



**HAL**  
open science

# Hypertextualité, construction du sens et représentation des connaissances.

Marc Augier

► **To cite this version:**

Marc Augier. Hypertextualité, construction du sens et représentation des connaissances.. domain\_stic.gest. Université Nice Sophia Antipolis, 2005. Français. NNT: . tel-00080166

**HAL Id: tel-00080166**

**<https://theses.hal.science/tel-00080166v1>**

Submitted on 14 Jun 2006

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**Université de Nice - Sophia Antipolis**

*Faculté de Lettres, Arts et Sciences Humaines*

ÉCOLE DOCTORALE LETTRES, SCIENCES HUMAINES ET SOCIALES

**I3M**

INFORMATION, MILIEUX, MÉDIAS, MÉDIATIONS

## **Hypertextualité**

# **Construction du sens et représentation des connaissances**

Tome I

## **THÈSE**

présentée et soutenue publiquement le 2 décembre 2005

*Pour l'obtention du grade de:*

**Docteur**

*Discipline :* **Sciences de l'information et de la communication**

par

Marc Augier

Directeur de thèse :

Monsieur le Professeur Norbert Hillaire

Jury:

Monsieur le Professeur Éric Bruillard (IUFM Caen)

Monsieur David Piotrowski (Chargé de recherche CREA-CNRS)

Monsieur le Professeur Pascal Vidal (CERAM Sophia Antipolis)

Monsieur Georges Vignaux (Directeur de recherche CNRS)

Monsieur le Professeur Norbert Hillaire (UNSA)

Mis en page avec la classe thloria.

## Remerciements

Les remerciements....

Exercice de style pendant les remises de prix de chaque cérémonie officielle, c'est en les rédigeant que l'on comprend mieux pourquoi les gens qui les lisent ont les yeux mouillés.

Ce travail, *a priori* et forcément personnel, n'a toutefois été possible que grâce au soutien, direct ou indirect, de nombreuses personnes. Certains m'ont aidé directement, d'autres indirectement, mais tous ont joué un rôle déterminant même s'ils n'étaient pas présents à mes côtés. Je pense en premier lieu à mes grand-parents, Félix et Emma ; c'est grâce à eux que j'ai développé confiance en moi et curiosité, quand je n'étais encore qu'un enfant. Ces qualités m'ont suivi bien après qu'ils nous aient quittés et il se passe peu de journées sans que j'ai une pensée pour eux. Je remercie aussi mes parents, qui m'ont laissé le libre arbitre de mes choix, ce qui m'a permis de me construire de manière autonome. Je remercie mon épouse et mes enfants, qui m'ont accompagnés sur le chemin de cette thèse ; leurs soirées et leurs vacances en ont souvent souffert, mais je n'ai jamais entendu de reproche et toujours des encouragements. Maria, tu as souvent su me redonner courage quand le doute m'envahissait. Matthieu, j'espère te voir bientôt sur les bancs de l'Université, maintenant c'est à toi ! Et *last but not least* Daphné, je compte sur toi pour tous nous rattraper et même nous doubler.

Il y a aussi des remerciement plus professionnels, qui n'en sont pas moins chaleureux. Je voudrais tout d'abord remercier tout particulièrement RMS, sans le savoir, c'est lui qui a rendu tout cela possible. Il ne me lira sans doute malheureusement jamais, mais les remerciements sont aussi le moment de rêver un peu... Je remercie Yannick Geffroy et Norbert Hillaire pour m'avoir fait confiance ; un ingénieur en construction mécanique et informaticien se

sent parfois comme un éléphant dans un magasin de porcelaine. En m'accueillant et en me permettant de participer à votre projet commun avec le CNRS, vous m'avez permis de découvrir que mes expériences professionnelles et personnelles précédentes avaient malgré les apparences un fil conducteur. Pour terminer, je remercie chaleureusement avec une attention toute particulière Georges Vignaux et Marc Silberstein pour leur accueil et la confiance qu'ils m'ont témoignée dans le cadre du projet *CoLiSciences*. Ils ont souvent fait preuve de beaucoup de patience pour comprendre, adopter mais quelquefois aussi rejeter avec beaucoup de tact les expériences technico-informatiques que je lançais sans crier gare. Je leur suis infiniment redevables pour leurs précieux conseils et le soutien qu'ils m'ont apporté tout au long de mon projet personnel de thèse qui voit son aboutissement dans ce document. L'aboutissement de ce projet n'est bien entendu pas tout, et j'espère que nous aurons encore de nombreuses occasions de collaborer, que ce soit pour *CoLiSciences* ou pour d'autres expériences, que je serai heureux et fier de pouvoir accompagner.

*Je dédie cette thèse  
à Maria, mon épouse.*



# TABLE DES MATIÈRES

---

---

---

<b>Introduction Générale</b>	<b>1</b>
------------------------------	----------

---

---

<b>Chapitre 1</b>
<b>Introduction</b>

1.1	Objectifs thématiques et finalités . . . . .	5
A.	Objectif Général . . . . .	8
1.	Analyser . . . . .	8
2.	S'interroger . . . . .	9
3.	Croiser . . . . .	9
4.	Discuter . . . . .	10
5.	Numériser . . . . .	11
6.	Éditer . . . . .	11
B.	Objectifs Particuliers . . . . .	12
1.	Hypertexte . . . . .	12

2.	Parcours . . . . .	18
3.	Sens et cognition . . . . .	18
1.2	Cadre de référence théorique . . . . .	20
A.	Première idée régulatrice : la simulation . . . . .	21
B.	Deuxième idée régulatrice : la rétroaction . . . . .	21
C.	Troisième idée régulatrice : l'information codée . . . . .	22
1.3	Éléments méthodologiques . . . . .	23
A.	Finalité de cette recherche . . . . .	23
1.	Action . . . . .	23
2.	Analyse de corpus . . . . .	24
3.	Études expérimentales . . . . .	24

<b>Chapitre 2</b>
-------------------

<b>Cadre épistémologique</b>
------------------------------

2.1	Les questions à résoudre . . . . .	27
A.	Problématique . . . . .	27
2.2	Organisation par thèmes de recherche . . . . .	27
A.	Intelligence artificielle . . . . .	28
1.	Définition . . . . .	28
2.	Système expert . . . . .	32
3.	Traitement du langage naturel . . . . .	33
4.	Pensée et apprentissage . . . . .	34
5.	Approche connexioniste . . . . .	35
B.	Cartographie . . . . .	38
C.	Orientation et repérage . . . . .	41
D.	La linguistique cognitive . . . . .	44
E.	La communication . . . . .	45
F.	L'informatique linguistique . . . . .	47
G.	La sémiotique . . . . .	47
2.3	Organisation de la thèse . . . . .	49

---

---

## Partie I Du livre au Web

### L'hypertexte nouveau support de pensée ?

---

---

<b>Chapitre 3</b>
-------------------

<b>L'écrit, la médiation, la numérisation</b>
---

3.1	Introduction . . . . .	55
3.2	Le document . . . . .	56
A.	Définition . . . . .	56
B.	Théorie du support numérique . . . . .	58
C.	La chaîne documentaire . . . . .	62
1.	Définition . . . . .	62
2.	L'application, notion technique imposée . . . . .	63
3.	Approche dynamique du document . . . . .	67
3.3	La notion de collection . . . . .	68
A.	La définition du dictionnaire . . . . .	69
B.	La collection pour valoriser . . . . .	71
3.4	Le livre, du papier à l'électronique . . . . .	73
A.	Le livre traditionnel . . . . .	75
1.	Poids économique . . . . .	79
B.	Qu'entend-on par livre électronique ? . . . . .	81
1.	Première définition du livre électronique . . . . .	83
2.	La définition du dictionnaire . . . . .	85
3.	Notre définition . . . . .	85
C.	Le point sur les technologies . . . . .	86
1.	Modèles ou contenus anglophones . . . . .	86
2.	Modèles ou contenus francophones . . . . .	88
3.	Cartables et manuels numériques . . . . .	90
4.	Normes, standards et formats . . . . .	93
5.	Papier et encre électroniques . . . . .	98
D.	Livre électronique : Rupture ou continuité ? . . . . .	101
1.	Texte imprimé ou électronique . . . . .	105

2.	Le futur du livre ? . . . . .	106
3.	Rupture . . . . .	108
4.	Les métadonnées . . . . .	118
5.	Lecture sur papier et lecture sur écran . . .	120
6.	Modélisations sémantiques et cognitives . .	128
3.5	Conservation numérique . . . . .	136
A.	Les formats . . . . .	136
B.	La numérisation comme changement de support . .	138
C.	La chaîne de numérisation . . . . .	138
1.	Soin continu . . . . .	139
2.	Rafraîchissement . . . . .	140
3.	Migration . . . . .	140
4.	Émulation . . . . .	141
5.	Conservation de la technologie . . . . .	141
3.6	Les corpus électroniques . . . . .	144
A.	Pourquoi un corpus en ligne ? . . . . .	145
1.	Pérennisation et patrimoine . . . . .	145
2.	Travail collaboratif . . . . .	147
B.	Les grands projets de numérisation de textes . . . . .	149
1.	Les projets Français . . . . .	150
2.	Les projets étrangers . . . . .	158

<b>Chapitre 4</b>
-------------------

<b>Réflexions sur l'hypertexte</b>
------------------------------------

4.1	Introduction à l'hypertexte . . . . .	165
A.	Historique . . . . .	166
1.	Un mot . . . . .	166
2.	Un concept . . . . .	166
3.	Les ancêtres . . . . .	169
B.	Définition . . . . .	170
4.2	Fondements théoriques de l'hypertexte . . . . .	176
A.	Modèles théoriques hypertextuels . . . . .	177
1.	Modèle de Marc Nanard . . . . .	177
2.	Modèle de Dexter . . . . .	178

---

3.	Modèle à trois modules . . . . .	179
4.3	Les notions autour de l’hypertexte . . . . .	180
A.	Interactivité . . . . .	180
B.	Hypertexte et connaissances . . . . .	183
C.	Hypertexte collaboratif . . . . .	187
1.	Définition de la collaboration . . . . .	187
2.	Écriture hypertextuelle : Wiki . . . . .	188
3.	Wikipédia . . . . .	190
D.	Les systèmes hypertextes . . . . .	199
1.	Web et hypertexte . . . . .	201
4.4	La Gestion électronique de Documents . . . . .	204
A.	Définition . . . . .	204
B.	Le traitement du document . . . . .	204
1.	L’indexation . . . . .	207
2.	Le catalogage . . . . .	212
3.	Le résumé . . . . .	212
C.	Les normes de balisage . . . . .	213
1.	DTD . . . . .	213
2.	SGML . . . . .	213
3.	XML . . . . .	213
D.	Les DTD de l’information–documentation . . . . .	216
1.	TEI Text Encoding Initiative . . . . .	217
2.	DocBook . . . . .	218
3.	BiblioML . . . . .	218
4.	Dublin Core . . . . .	219

---



---

## Partie II Construction et représentation du sens

---



---

**Chapitre 5**

**Construction du sens**

5.1	Introduction . . . . .	225
	A. Définition du sens . . . . .	226
	B. Définition de la quantité d'information . . . . .	230
5.2	Mémoire de lecture . . . . .	233
	A. Réseau et apprentissage . . . . .	234
	B. Restitution des parcours . . . . .	237
	C. Modèle de représentation . . . . .	238
5.3	Le sens . . . . .	239
	A. Cadre de référence théorique . . . . .	239
	B. Créer du sens, pour qui, par qui ? . . . . .	240
	C. Navigation et recherche d'information . . . . .	244
	D. Représentation de la connaissance . . . . .	245
	1. L'intelligence artificielle . . . . .	245
	2. Le Web Sémantique . . . . .	246
	3. Métadonnées et annotations . . . . .	251
5.4	Sémantique interprétative . . . . .	252

**Chapitre 6**

**Le dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences***

6.1	Introduction . . . . .	257
	A. Présentation et objectifs . . . . .	258
6.2	Description du système <i>C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences</i> . . . . .	261
	A. Une base de connaissances (multimédia) . . . . .	264
	B. Un espace de travail collaboratif . . . . .	266
	C. Une bibliothèque virtuelle . . . . .	267
6.3	Le dispositif technique . . . . .	268
	A. Modèle de publication . . . . .	273
	B. Le stockage de données XML . . . . .	275
	C. Délinéarisation et découpage du corpus . . . . .	281
	1. L'unité d'information . . . . .	281
	2. Délinéariser mais pas désorienter . . . . .	282
	3. Génération automatique des liens . . . . .	283

4.	Problématique du changement de support .	283
D.	L'influence des réseaux de Petri . . . . .	284
E.	Structuration . . . . .	289
F.	Appropriation et apprentissage . . . . .	290
1.	<i>C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup></i> : un outil pour l'étude des processus d'appropriation . . . . .	291
G.	Les médiations autour de <i>C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup></i> . . . . .	294
1.	Pourquoi publier sur la toile . . . . .	294
2.	Les dispositifs disponibles dans <i>C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup></i> . . . . .	297
3.	Les interfaces . . . . .	299
6.4	Notre architecture . . . . .	302
A.	Numérisation . . . . .	304
B.	Architecture serveur . . . . .	306
C.	Le corpus disponible . . . . .	307
D.	Construction de la DTD . . . . .	308

## Chapitre 7

### Représentation des parcours de lecture

7.1	Navigation à travers <i>C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup></i> . . . . .	317
7.2	Parcours . . . . .	326
A.	Qu'est-ce qu'un parcours ? . . . . .	326
B.	La co-construction des parcours de lecture . . . . .	326
C.	Reconstructions du sens et parcours . . . . .	327
D.	Lecture sur écran et parcours cognitifs . . . . .	328
1.	Typologie des parcours et Navigation . . . . .	331
2.	La compréhension de textes . . . . .	332
7.3	Topologie . . . . .	333
A.	Hypertextes et repères pour la navigation . . . . .	333
B.	Le cas de la bibliothèque numérique . . . . .	335
C.	Représentation . . . . .	337
1.	Les Schémas de représentation . . . . .	338
2.	Les attributs techniques . . . . .	341
7.4	Lectochromie . . . . .	342
A.	Description . . . . .	343
1.	Un système commun de référence . . . . .	343

	2.	Matrice de parcours . . . . .	343
B.		Cartographie ou statistique ? . . . . .	343
	1.	Statistiques de lecture . . . . .	343
	2.	La cartographie pour rendre visible . . . . .	344
	3.	La couleur pour donner du sens . . . . .	345
C.		Mise en œuvre . . . . .	347
D.		Modélisation . . . . .	360
E.		Analyse . . . . .	362
	1.	Construction de parcours . . . . .	362
	2.	Les lecteurs . . . . .	363
	3.	Les possibilités d'évolution du système . . . . .	364

---

---

## Partie III Genèse du dispositif *CoLiSciences*

### Retour d'expérience et mise en œuvre

---

---

<b>Chapitre 8</b>
-------------------

<b>Les Logiciels Libres</b>
-----------------------------

8.1		Introduction . . . . .	369
A.		Définitions . . . . .	376
	1.	Logiciel . . . . .	376
	2.	Logiciel Libre . . . . .	377
	3.	Open Source . . . . .	383
B.		Les licences libres . . . . .	386
	1.	Licences pour les logiciels . . . . .	387
	2.	Licences pour les documents et créations artistiques . . . . .	389
C.		Les textes fondateurs . . . . .	392
	1.	La cathédrale et le bazar . . . . .	392

---

2.	Le chaudron magique . . . . .	393
D.	La rencontre . . . . .	396
8.2	Nouveaux cadres pratiques, nouveaux modèles théoriques .	399
A.	Quel statut? . . . . .	399
B.	Programmeur de Logiciel Libre? . . . . .	401
1.	Motivation . . . . .	401
2.	Éthique du programmeur . . . . .	403
3.	Plus simplement . . . . .	403
C.	Logiciel libre et innovation . . . . .	405
1.	Méthode scientifique . . . . .	407
8.3	Logiciels Libres et recherche scientifique . . . . .	409
A.	Introduction . . . . .	409
B.	Une impasse scientifique . . . . .	410
C.	Le domaine public en danger . . . . .	411
1.	Propriété Intellectuelle contre domaine pu- blic . . . . .	413
2.	Copyright contre communauté . . . . .	414
3.	Ordinateur de confiance . . . . .	414
4.	Brevets Logiciels . . . . .	418
8.4	Pourquoi des Logiciels Libres dans un projet de numérisation	419
A.	Nos objectifs et contraintes . . . . .	420
1.	Pérennité . . . . .	420
2.	Fiabilité . . . . .	423
3.	Contraintes Techniques et Logicielles . . . . .	424
4.	Contraintes Budgétaires . . . . .	425
B.	Composants logiciels . . . . .	426
1.	Le logiciel système . . . . .	426
2.	Les logiciels applicatifs . . . . .	429
C.	Méthode de développement . . . . .	430
1.	Une méthode pour formaliser les interactions	431
2.	eXtreme Programming . . . . .	432

<b>Chapitre 9</b>
-------------------

<b>Le processus de concrétisation d'une bibliothèque numérique</b>
--

9.1	Introduction . . . . .	439
9.2	Bibliothèque et réseau...	
	... ou réseau de bibliothèque. . . . .	440
A.	Fédérer un réseau de bibliothèques . . . . .	442
	1. Faut-il le faire ou laisser Google le faire pour nous? . . . . .	442
	2. Grand projet institutionnel . . . . .	446
B.	Quel lien avec les bibliothèques traditionnelles? . . . . .	447
9.3	Bibliothèque électronique...	
	... ou système d'information. . . . .	448
A.	Les outils . . . . .	448
B.	La relation outil et utilisation . . . . .	451
C.	Quelle bibliothèque? . . . . .	452
	1. Les typologies . . . . .	453
D.	Le Web Sémantique . . . . .	453
9.4	Bibliothèque et société . . . . .	454
A.	Les aspects économiques . . . . .	455
B.	Les aspects juridiques . . . . .	458

---

---

## Conclusion et perspectives

---

---

<b>Chapitre 10</b>
<b>Conclusion</b>

10.1	Conclusion Générale . . . . .	465
	A. Commentaires critiques sur la réalisation . . . . .	465
	B. Notre modèle hypertextuel . . . . .	466
	C. Contextes de publication sur le Web . . . . .	468
10.2	Perspectives de recherche . . . . .	469
	A. Le futur du livre : Écouter un livre . . . . .	469

---

B.	Bibliothèques numériques . . . . .	471
C.	Le droit d’auteur numérique . . . . .	472
D.	Production de document en collaboration . . . . .	473

---



---



---

## Bibliographie, Index, Glossaire et Annexes

---



---

<b>Bibliographie</b>	<b>477</b>
<b>Les index</b>	<b>505</b>
<b>Glossaire</b>	<b>523</b>

<b>Les Annexes</b>
--------------------

<b>Annexe A La GED et la chaîne de numérisation</b>	<b>551</b>
<b>Annexe B Les projets de numérisation</b>	<b>553</b>
<b>Annexe C Le projet <i>C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences</i></b>	<b>555</b>
<b>Annexe D Les <i>referers C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences</i></b>	<b>559</b>
<b>Annexe E Message au Président du Conseil Européen</b>	<b>567</b>
<b>Annexe F Les <i>statistiques C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences</i></b>	<b>571</b>
<b>Annexe G Manuel de l’administrateur <i>C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences</i></b>	<b>573</b>
G.1 Manuel de l’administrateur <i>C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences</i> . . . . .	574
A. Introduction en forme d’avertissement . . . . .	574
B. Modification de la page d’accueil . . . . .	574
<b>Annexe H Références sur les fondateurs de l’hypertexte</b>	<b>585</b>

<b>Annexe I Licences Libres</b>	<b>589</b>
I.1 La licence GPL . . . . .	589
1. Preamble . . . . .	590
<b>Annexe J Brevetabilité des logiciels</b>	<b>601</b>
J.1 Les brevets logiciels . . . . .	602
A. Pourquoi faut-il être contre les brevets logiciels? . .	602
B. Conséquences pour <i>CoLiSciences</i> . . . . .	608
1. Pages dynamiques . . . . .	609
2. Aide contextuelle . . . . .	609
3. Systèmes d’onglets . . . . .	610
<b>Annexe K Achat d’ouvrages numériques en ligne</b>	<b>613</b>
K.1 Petit voyage à travers DRM, cas pratique . . . . .	614
<b>Annexe L Ouvrages numériques et domaine public</b>	<b>623</b>
L.1 Numérique contre domaine public . . . . .	624

# TABLE DES FIGURES

---

1.1	Le moteur <i>CoLiSciences</i> . . . . .	14
1.2	Schéma technique . . . . .	15
1.3	Les trois modules . . . . .	16
1.4	Le moteur hypertexte vu selon notre modèle . . . . .	17
1.5	Le triple modèle à trois modules . . . . .	19
2.1	Développement comparé des thésaurus . . . . .	40
3.1	Prototype Xerox de papier électronique . . . . .	100
4.1	L’ancrage dans le modèle à deux couches de Marc Nanard . . . . .	177
4.2	Le modèle à trois couches de Dexter . . . . .	179
4.3	Le modèle à trois modules . . . . .	180
4.4	Le modèle à trois modules en détail . . . . .	181
4.5	La page RèglesDeFormatageDesTextes sur un Wiki . . . . .	189
4.6	Courbe de progression du Wikipédia francophone . . . . .	192
4.7	Historique de modification d’un article . . . . .	195
4.8	Problème de pertinence d’un article . . . . .	197

Table des figures

---

4.9	Les tâches du documentaliste . . . . .	207
4.10	Le principe de la classification de Dewey . . . . .	211
6.1	Le plan du site <i>C<sup>o</sup>L<sup>i</sup>S<sup>ci</sup>ences</i> . . . . .	259
6.2	Généalogie du dispositif <i>C<sup>o</sup>L<sup>i</sup>S<sup>ci</sup>ences</i> . . . . .	270
6.3	Le dispositif <i>C<sup>o</sup>L<sup>i</sup>S<sup>ci</sup>ences</i> avance comme une chenille . . . . .	272
6.4	Modèle de publication traditionnel . . . . .	274
6.5	Modèle de publication mixte . . . . .	276
6.6	Modèle de publication XML . . . . .	277
6.7	Page d'accueil <i>C<sup>o</sup>L<sup>i</sup>S<sup>ci</sup>ences</i> . . . . .	284
6.8	Graphe d'un parcours géré par le moteur hypertextuel . . . . .	285
6.9	Les nœuds d'un réseau de Petri . . . . .	286
6.10	Modélisation des quatre saisons (Petri) . . . . .	287
6.11	Modélisation du moteur <i>C<sup>o</sup>L<sup>i</sup>S<sup>ci</sup>ences</i> (Petri) . . . . .	288
6.12	DTD préparée pour le prototype Claude Bernard . . . . .	308
6.13	DTD préparée pour la mise en ligne <i>C<sup>o</sup>L<sup>i</sup>S<sup>ci</sup>ences</i> . . . . .	309
6.14	Le fac-similé de la page 1 de <i>IEME</i> . . . . .	311
6.15	Le paragraphe 1 de <i>IEME</i> sous forme de fichier XML . . . . .	312
6.16	Le paragraphe 1 de <i>IEME</i> tel qu'il apparaît en ligne (parcours notionnel) . . . . .	312
6.17	DTD préparée pour la numérisation par Jouve . . . . .	313
7.1	Page d'accueil <i>C<sup>o</sup>L<sup>i</sup>S<sup>ci</sup>ences</i> . . . . .	318
7.2	S'identifier sur <i>C<sup>o</sup>L<sup>i</sup>S<sup>ci</sup>ences</i> . . . . .	318
7.3	Les informations personnelles . . . . .	319
7.4	Page d'accueil <i>C<sup>o</sup>L<sup>i</sup>S<sup>ci</sup>ences</i> après identification . . . . .	320
7.5	L'accès au corpus par la liste des auteurs . . . . .	321
7.6	L'accès au corpus par disciplines . . . . .	322

---

7.7	L'accès au corpus par domaines . . . . .	323
7.8	L'accès au corpus par parcours notionnel . . . . .	324
7.9	Le liste des signets disponibles sur le corpus . . . . .	325
7.10	Arborescence sans organisation hiérarchique . . . . .	334
7.11	Arborescence en organigramme . . . . .	335
7.12	Public@tions Électroniques de Philosophi@ Scient@e . . . . .	339
7.13	Bibliothèque et Portail . . . . .	340
7.14	Bibliothèque virtuelle KeeBook . . . . .	341
7.15	Exemple d'échelle de couleur pour la lectochromie . . . . .	347
7.16	Un paragraphe du corpus <i>C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup></i> . . . . .	348
7.17	Le menu d'accueil de la lectochromie . . . . .	350
7.18	Les parcours de l'internaute dans l'ouvrage en cours . . . . .	351
7.19	les parcours de tous dans l'ouvrage en cours . . . . .	352
7.20	Les parcours de l'internaute dans le corpus . . . . .	355
7.21	Les parcours de tous dans le corpus . . . . .	356
7.22	les parcours de tous dans le corpus (1/5) . . . . .	357
7.23	les parcours de tous dans le corpus (2/5) . . . . .	358
7.24	les parcours de tous dans le corpus (3/5) . . . . .	358
7.25	les parcours de tous dans le corpus (4/5) . . . . .	359
7.26	les parcours de tous dans le corpus (5/5) . . . . .	359
8.1	Les parts de marché Apache <i>vs</i> Microsoft . . . . .	425
8.2	Tout est parti d'un message de Linus . . . . .	427
8.3	Le projet eXtreme Programming . . . . .	435
8.4	Les itérations du projet eXtreme Programming . . . . .	435
D.1	La page des référants dans l'administration <i>C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup></i> . . . . .	560
F.1	Les statistiques de connexion . . . . .	572

## Table des figures

---

G.1	La page d'accueil <i>CoLis</i> ciences de la version 1 . . . . .	575
G.2	La page d'accueil actuelle . . . . .	576
G.3	La page de connexion de l'administrateur . . . . .	576
G.4	Les icônes du menu d'administration . . . . .	577
G.5	Les icônes du menu d'administration . . . . .	578
G.6	La gestion du contenu . . . . .	580
G.7	La gestion du contenu . . . . .	581
G.8	L'ajout d'une catégorie . . . . .	582
G.9	La liste des catégories . . . . .	583
G.10	L'ajout d'une page . . . . .	584
J.1	Les parts de marché Apache <i>vs</i> Microsoft . . . . .	603
K.1	Confirmation commande NumiLog . . . . .	615
K.2	Le mail de confirmation . . . . .	617
K.3	Version d'Adobe Reader non reconnue . . . . .	618
K.4	DRM et Adobe . . . . .	619
K.5	DRM avec Adobe Reader sans ouvrir de compte Passport . . . . .	619
K.6	id Adobe . . . . .	620
K.7	Création d'un compte Adobe (1 <sup>ère</sup> partie) . . . . .	621
K.8	Création d'un compte Adobe (2 <sup>ème</sup> partie) . . . . .	621

# LISTE DES TABLEAUX

---

3.1	La production d'ouvrages papier . . . . .	81
4.1	Tableau comparatif entre encyclopédies . . . . .	193



# **Introduction Générale**



## CHAPITRE 1

# INTRODUCTION

---

*« Chacun doit savoir ce qu'il cherche  
pour éviter de ne pas voir ce qu'il trouve. »*

Claude Bernard

*« Une expérience scientifique est [...] une expérience  
qui contredit l'expérience commune. »*

Gaston Bachelard

*La Formation de l'esprit scientifique*

## Sommaire

---

<b>1.1 Objectifs thématiques et finalités . . . . .</b>	<b>5</b>
A. Objectif Général . . . . .	8
B. Objectifs Particuliers . . . . .	12
<b>1.2 Cadre de référence théorique . . . . .</b>	<b>20</b>
A. Première idée régulatrice : la simulation . . . . .	21
B. Deuxième idée régulatrice : la rétroaction . . . . .	21
C. Troisième idée régulatrice : l'information codée . . . . .	22
<b>1.3 Éléments méthodologiques . . . . .</b>	<b>23</b>
A. Finalité de cette recherche . . . . .	23

---

## 1.1 Objectifs thématiques et finalités

Cette thèse se présente dans le champ interdisciplinaire des sciences de l'information et de la communication (SIC), et a pour but d'explorer la question de la construction du sens dans un corpus de textes sous format électronique en relation avec la question de la représentation de cette construction et des connaissances apparentes ou sous-jacentes du corpus. Les éléments fondamentaux de problématique qui guident ce travail sont centrés sur l'hypertexte, notion qui s'est imposée dans le champ de l'usage des technologies de l'information et de la communication depuis une quinzaine d'années. De nombreuses activités ont été proposées et testées, des produits ont été diffusés.

*« Aujourd'hui, le formidable succès du Web a consacré la notion d'hypertexte, mais en la réduisant, dans l'esprit de nombre d'utilisateurs, à une simple technologie d'accès à des informations ou des documents. En effet, si l'hypertexte peut se définir comme un dispositif informatisé permettant l'interconnexion de documents de divers types, non sur la base d'un modèle hiérarchique ou relationnel, mais par des mécanismes associatifs sous contrôle de l'utilisateur (la navigation), ce n'est pas ainsi qu'il est en général véritablement compris. La vision dominante n'est pas tant celle d'un dispositif que d'une action locale (ponctuelle) : des pages sont reliées entre elles et on se « déplace » de l'une à l'autre en cliquant sur des mots ou expressions mises en exergue ou des zones de l'écran « réactives ». Dans ce contexte, on peut rechercher des informations. On est loin d'avoir l'idée que l'hypertexte puisse aider à en développer de nouvelles. C'est une technologie d'accès, pas de structuration. (Bruillard, 2002) »*

Plus largement, nous entendons mener une analyse de l'hypertexte qui va bien au-delà de l'objet technique. Notre finalité se trouve plutôt dans l'étude des relations et interactions de l'hypertexte avec les autres composantes des sciences de l'information et de la communication. Elle se retrouve en partie dans la citation suivante :

« Comment l'analyse des usages d'un objet technique lié à un objet de savoir et aux concepts qui caractérisent les connaissances associées à un domaine spécifique, permet-il de penser l'intégration et l'utilisation qui est faite des TICE. (Agostinelli, 1999, page 20) »

Le traitement numérique et la mise en réseau transforment en profondeur les relations aux documents pour les individus, les communautés et les sociétés. Parmi les enjeux socio-économiques de ces transformations, on trouve notamment l'accès partagé à l'information, la gestion des connaissances pour les organisations, une valeur ajoutée pour les services (commerce électronique, transports, tourisme, santé, l'éducation, etc.), la transformation des industries culturelles.

Cette recherche s'appuie sur les actions menées pour le programme de recherches *CoLiSciences*<sup>1</sup> de l'équipe « *Hypertextes et textualité électronique* » dirigée par Georges Vignaux à la Maison des Sciences de l'Homme Paris Nord. Il s'agit de la réalisation et du développement d'un grand hypertexte sur les biologistes et naturalistes du XIX<sup>e</sup> siècle en langue française, mis en ligne sur Internet à l'adresse : <http://www.colisciences.net>. Nous y reviendrons plus loin de manière plus approfondie, en particulier dans le chapitre 6 *Le dispositif CoLiSciences*. Pour présenter succinctement ce programme, il s'agit de la création d'un grand corpus électronique d'histoire des idées

---

1. *CoLiSciences* est la poursuite d'un premier programme de recherche qui avait conduit au lancement du projet COLIS, acronyme de Corpus de Littérature Scientifique.

centré sur la période fin XVIII – courant XIX, et la naissance d’une science moderne du vivant. Dans ce corpus, où prédominent les grand naturalistes et les biologistes français, nous pouvons voir émerger les concepts fondamentaux de la biologie moderne en même temps que nous y trouvons formulées nombre de questions toujours actuelles (théorie de l’évolution, inné/acquis, hérédité, morphogenèse, classification...). Les ouvrages scientifiques qui ont été rassemblés pour constituer le corpus *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* appartenaient à la Bibliothèque André Desguine, et ils sont maintenant conservés aux Archives départementales des Hauts-de-Seine. Ils sont au nombre de 14 et représentent environ 5500 pages, dont le détail avait été donné en annexe de notre mémoire de DEA (Augier, 2002). Pour terminer, un article de présentation du projet *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* paru dans le journal « *Le Monde* » est disponible en annexe C.

*C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* répond à trois objectifs :

1. Se donner les moyens de l’édition en ligne de collections d’auteurs biologistes et naturalistes du XIX<sup>e</sup> siècle, en langue française, et par là valoriser un patrimoine historique et un moment de grande importance dans l’histoire des sciences et de la pensée ;
2. Prendre ce corpus inédit comme base de réflexion sur la question scientifique de la nature de l’hypertextualité : processus, parcours, liens, lectures, navigations ;
3. Contribuer, ce faisant, aux travaux sur l’histoire des idées dans ce contexte précis du développement et de la consolidation d’une science « moderne » du vivant.

De ces trois objectifs résultent plusieurs axes de travail :

- la constitution d’un corpus raisonné et annoté, permettant la mise en valeur scientifique et patrimoniale de certains états de la pensée (au sens d’« histoire des idées ») au sein d’un domaine – la biologie – constam-

- ment traversé par des controverses et des innovations méthodologiques et conceptuelles ;
- la prise en compte de la spécificité de l’hypertextualisation quand il s’agit de tester sa pertinence comme outil pour aborder les questions de la « navigation » dans une masse documentaire particulièrement étendue, profuse et diverse ;
  - le questionnement sur les processus d’acquisition de connaissances via ce dispositif particulier.

## A. Objectif Général

Les modalités de mise en œuvre du dispositif *CoLis* sciences constituent l’objectif général de notre recherche. La question dont nous cherchons la réponse pourrait se résumer de la manière suivante : *Pourquoi constituer un corpus et comment ?* Toutefois notre objectif général peut se décliner sous la forme d’une suite de questions plus précises, auxquelles nous tenterons de répondre. Celles-ci sont souvent très imbriquées entre elles bien que nous les présentions ici une par une. Par exemple *interrogation* et *analyse* sont liées par la détermination des conditions préliminaires à la collecte, ce qui nous ramène à notre première question.

Notre recherche va s’articuler autour de la liste d’interrogations suivante que nous allons chercher à éclaircir :

### 1. Analyser

Analyser les documents (structure, évolutivité, collection, etc...), les médiations (techniques et sociales) et leur relation avec l’activité humaine et ses limites (navigation, recherche documentaire et utilité, construction sociale des normes, etc.).

## 2. S'interroger

S'interroger sur les outils à construire, sur les modalités de leur construction et sur les conséquences de leur utilisation. Parler de développement d'outils informatiques dans un travail de recherche sur l'émergence du sens dans un corpus hypertextuel pourrait paraître paradoxal, il n'en est pourtant rien tant les enjeux socioculturels sont liés aux enjeux techniques.

*« Bien que nous vivions en démocratie, les processus sociotechniques font rarement l'objet de délibérations collectives explicites, encore moins de décisions prises par l'ensemble des citoyens. (Lévy, 1990, page 8) »*

Ces réflexions seront développées plus largement dans le chapitre 8 : « Logiciels Libres », dans lequel nous détaillerons comment se sont faits nos choix technologiques. Ce sera l'occasion de poser la question de la relation entre outils technologiques et travail de recherche. Pour résumer brièvement notre propos, il s'agit d'établir l'indépendance du travail de recherche par rapport aux outils qu'il utilise ou qu'il analyse. Ensuite, il s'agit de proposer une solution à ce problème de liberté et d'ouverture sous la forme prise par des outils développés suivant les règles des logiciels libres.

## 3. Croiser

Croiser l'informatique (conception des systèmes et outils), la linguistique (analyse de textes, construction de terminologies, analyse sémantique), les sciences cognitives (construction du sens, modèles de compréhension et de lecture), et les « sciences de l'information » (modèles d'indexation, sémiotique, pragmatique, construction des modèles sociaux).

#### 4. Discuter

Discuter la notion de « document », reconfigurée par le développement numérique. Les recherches sur la notion de document sont assez riches et il est possible d'en trouver plusieurs définitions, plus ou moins générales.

**Vu de l'utilisateur :** un document est un support contenant une information lisible par l'homme.

*« Une information lisible par l'homme peut être :*

- un produit de l'écriture, au sens d'un système de notation ou de code du discours humain, autrement dit un texte ; ces documents-là remontent à l'invention de l'écriture, à 5000 ans d'ici ;*
- une pensée humaine, exprimée sans le secours de l'écrit codé et donc directement « lisible » par tout humain : dessin, croquis, image ;*
- un élément de matière, étranger à l'homme mais que celui-ci, grâce à ses connaissances et à facultés créatives, peut analyser, commenter, interpréter.*

*Les trois cas de figures ne sont pas étanches et on les trouve fréquemment mêlés. (Chabin, 2004, page 2) »*

**Vu de l'auteur :** le document est une trace ou un signe.

Avec le numérique nous allons nous trouver dans des situations où les documents n'ont plus un auteur humain, voire même un lecteur humain. Par exemple, dans la chaîne de production d'un ouvrage papier sur le Web, nous allons trouver des étapes de numérisation, de traduction d'un format à un autre, jusqu'à ce que le *document* puisse être lu en ligne. Le livre, ses pages,

vont apparaître successivement en ayant pris corps sous plusieurs instances qui sont normalement considérées comme des documents (intermédiaires).

Quel est le statut de ces documents ? La plupart ont été écrits par un logiciel et seront lu par un autre logiciel qui les transforme. L'opérateur du logiciel peut-il être considéré comme un auteur ou un lecteur ?

Ces documents intermédiaires ont-ils un statut à part entière ou héritent-ils du statut du document source, en l'occurrence le livre ?

## 5. Numériser

Nous employons le terme dans son sens le plus large qui englobe toutes les phases d'une chaîne de numérisation, depuis la collecte des textes jusqu'à leur mise sous un format électronique exploitable de manière informatique. La numérisation, les modalités de présentation d'un texte en ligne, les apports et les contraintes induites par le nouveau support, comment l'hypertextualité est mise en œuvre, dans quel but. Livre papier ou livre électronique ? De nombreux travaux de recherche portent souvent sur l'un ou l'autre, en comparant les aspects, s'interrogent sur les nouvelles médiations que le numérique introduit par rapport à son ancêtre papier. Nous commencerons par un état de l'art en la matière, que nous essayerons de poursuivre par l'étude de la numérisation. En effet, plutôt que d'opposer les deux médias, plutôt que chercher des points communs ou de divergence dans leur utilisation comme *lecteur*, nous verrons à travers la numérisation un nouveau travail d'*auteur* et d'*éditeur*.

## 6. Éditer

Pour prolonger notre réflexion sur le travail d'un éditeur numérique, nous poursuivrons le travail entrepris avec la numérisation pour détailler

les nouvelles pratiques de l'édition électronique. Comment mettre en forme par l'utilisation de feuilles de style, de traduction à base de technologies XML et autres procédés informatiques. Les nouveaux outils de l'édition ont migré depuis des technologies plus anciennes vers celles du numérique. Cette évolution est quelquefois une rupture, il y a eu et il y a encore une phase d'adaptation pendant laquelle à la fois les outils et leurs usages sont remis en cause, redéfinis et améliorés. Pour amoindrir le sentiment de rupture, il faudrait que le même savoir-faire puisse être mis en œuvre avec les nouveaux outils avec la même aisance que précédemment, voire mieux.

## **B. Objectifs Particuliers**

Les objectifs particuliers de notre recherche sont au nombre de trois, ils nous ont conduit à la réalisation de plusieurs dispositifs techniques. Ceux-ci conduisant à des expérimentations qui permettent de valider leur bon fonctionnement technologique et surtout la vérification de certaines de nos hypothèses.

### **1. Hypertexte**

Le dispositif essentiel est l'architecture du site *C<sup>o</sup>L<sup>i</sup>Sciences*, en rupture avec les technologies « Web » traditionnelles et qui permet de naviguer à l'intérieur du corpus en suivant les notions évoquées par le texte plutôt que les mots du texte. Ce n'est pas parce que l'interface naturelle du Web est hypertextuelle que nous nous sommes penchés sur ce domaine.

*« Le domaine de l'hypertexte concerne les diverses manières d'accroître les possibilités humaines de structurer et d'accéder à de vastes ensembles d'informations et de connaissances. (Bruillard, 1997, page 197) »*

C'est donc pour pouvoir naviguer à travers les pages du corpus *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, que nous avons adopté une interface hypertextuelle. Pour cela, nous sommes revenus aux sources de l'hypertexte et aux idées que Ted Nelson développe pour décrire Xanadu. Le dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* s'intéresse aux ouvrages et aux idées qu'ils transmettent. Il permet des parcours érudits à l'intérieur du corpus, rendant ainsi possible de suivre de paragraphe en paragraphe, par des liens hypertextuels, non pas des mots comme cela est généralement le cas, mais des notions. Les études qui s'intéressent à la navigation comme moyen d'accès à l'information sont essentiellement consacrées au développement de prototypes pour l'interrogation. La validation expérimentale de ces prototypes, l'appropriation qu'en réalisent les usagers, les démarches d'interrogation qu'ils induisent ne sont que très rarement abordées. Ainsi encore des recherches sur les types de navigations (« butinage » ou élaboration de requêtes), basées sur les connaissances dans des domaines, réseaux ou cartes sémantiques de documents. Les systèmes traditionnels de représentation de la connaissance sont centralisés et hiérarchisés. Ils nécessitent que chacun partage exactement une définition identique des concepts communs. Ce contrôle central est étouffant. De plus, l'augmentation de la taille et de la portée de tels systèmes les rend rapidement très difficiles à gérer. Pour essayer de résoudre ce problème de taille et de complexité, ces systèmes limitent les questions pouvant être posées (pour que l'ordinateur puisse répondre de manière fiable et à toutes les questions). L'exemple intéressant est celui des chercheurs du Web sémantique, qui au contraire, acceptent les paradoxes. Les questions sans réponse sont le prix à payer pour acquérir de la souplesse. On construit un langage de règles aussi significatif que nécessaire pour permettre au web de raisonner autant qu'on le veut. Le défi du Web sémantique est de fournir un langage qui exprime à la fois des données et des règles pour

raisonner sur les données et pour que les règles de n'importe quel système de représentation de la connaissance puissent être exportées sur le Web.

La recherche proposée est d'une part de revisiter le concept d'hypertexte pour lui intégrer la logique permettant de décrire les propriétés complexes des objets, sans empêcher les agents de résoudre un paradoxe. Et d'autre part de mettre au point des outils d'évaluation de la recherche par navigation.

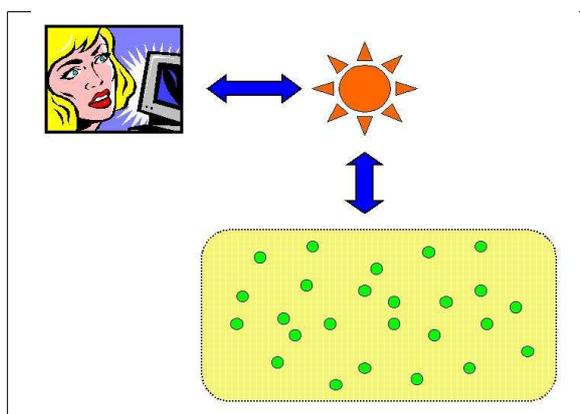


FIG. 1.1: Le moteur  $C_o^L_i^S_ciences$

Une première réponse avait déjà été apportée dans notre mémoire de DEA. Dans ce travail il s'agissait de revenir sur la théorie hypertextuelle et d'en comparer les modèles au nouveau modèle hypertextuel introduit par  $C_o^L_i^S_ciences$ . Sur la figure 1.1 nous avons représenté une des façons de « voir »  $C_o^L_i^S_ciences$  : un automate ou un agent intelligent<sup>2</sup>, capable de repérer dans le réseau hypertexte dont il a le contrôle le nœud qui répond le mieux aux attentes du lecteur.

---

2. Un agent intelligent est une sorte de robot virtuel. C'est à dire un logiciel qui accomplit des tâches à la manière d'un automate et en fonction de ce que lui a demandé son propriétaire. Dans le contexte d'Internet les agents intelligents sont liés au Web sémantique, dans lequel ils sont utilisés pour faire à notre place les recherches et les corrélations entre les résultats de ces recherches en fonctions de règles pré-définies. Ils ont capables d'une certaine autonomie, en particulier de dialoguer entre-eux. Par exemple l'agent intelligent d'une personne qui souhaite faire un achat sera capable de dialoguer avec les agents des vendeurs pour comparer prix, qualité et prestations. Le site <http://www.agentland.fr/> présente plusieurs applications déjà disponibles.

Plutôt que de présenter uniquement la structure logique de cette architecture hypertextuelle, il peut être intéressant de lui confronter la représentation de sa structure technique. Les deux schémas n'ont pas vocation à être comparés, en revanche cela permet déjà de constater que la même architecture peut être symbolisée de manière radicalement différente suivant comment on l'aborde. Le moteur hypertextuel représenté par sa structure technique interne pourrait être décrit comme cela est visible sur la figure 1.2, qui reprend les différents modules composant *C<sup>o</sup>L<sup>i</sup>s<sup>c</sup>iences* tout en schématisant leurs interactions, ceci au niveau technologique. Nous pourrions continuer de la sorte à représenter *C<sup>o</sup>L<sup>i</sup>s<sup>c</sup>iences* de différents points de vue, extérieurs cette fois, avec celui du lecteur ou de l'administrateur.

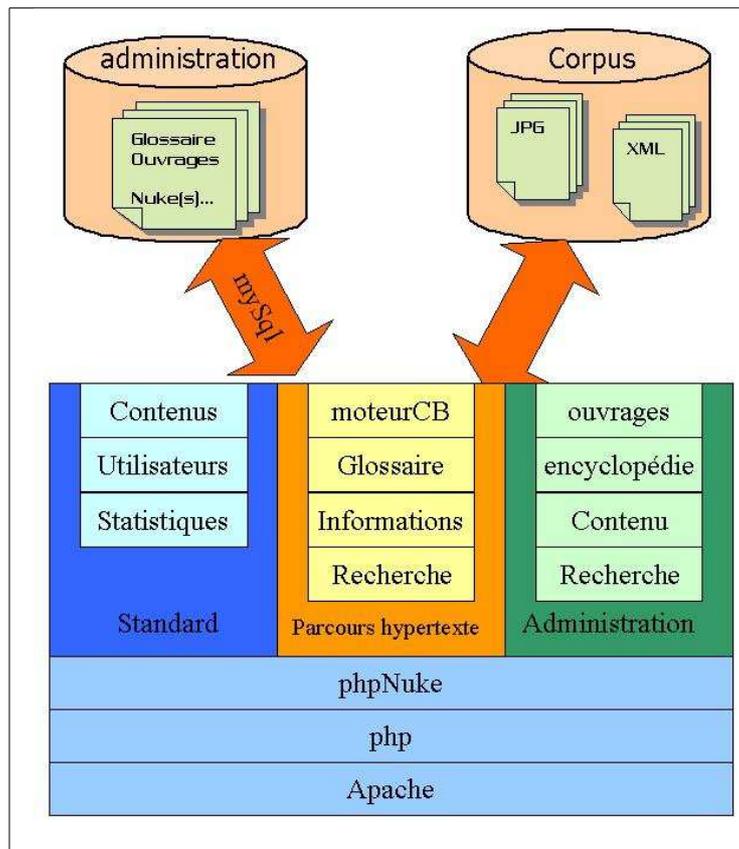


FIG. 1.2: Schéma technique

Dans notre première représentation logique, le réseau hypertexte n'est pas complètement construit. La trame générale est donnée puisque les nœuds sont figés, ce sont les paragraphes de texte codés en format XML, mais les parcours à travers le réseau sont construits par l'agent intelligent. C'est celui-ci qui décide d'établir un lien ou pas, en fonction bien entendu des règles qui lui ont été indiquées, mais ces règles peuvent être plus ou moins flexibles et lui donner plus ou moins de latitude dans ses actions.

Cette construction pourrait laisser croire que le réseau hypertextuel n'est qu'un amas de nœuds connectés le plus souvent aléatoirement et qu'à vouloir une structure souple, nous avons construit un système sans aucune structure. Pour ne pas tomber dans cet écueil nous avons pris garde à construire l'agent intelligent de manière à ce qu'il respecte un modèle hypertextuel théorique précis. Nous présenterons les éléments théoriques majeurs qui ont conduit notre réflexion dans le chapitre 4 consacré à l'hypertexte, pour l'instant nous nous contenterons de présenter le résultat de cette recherche.

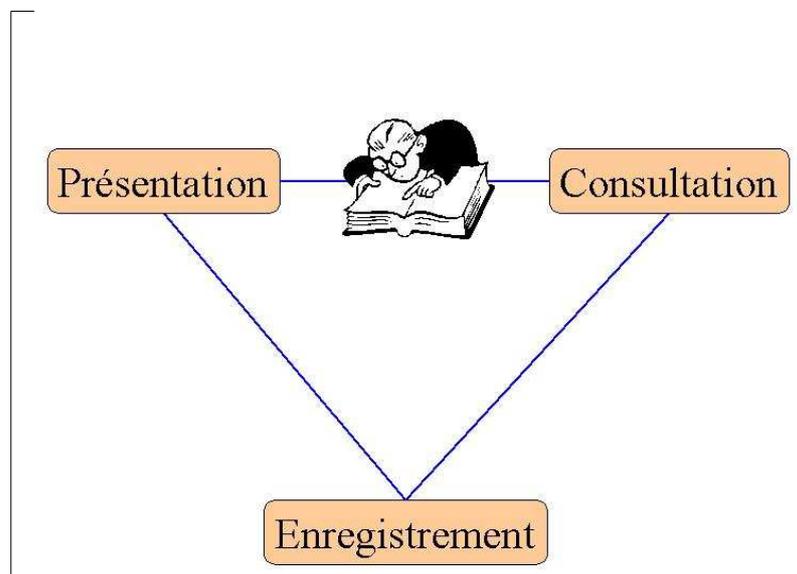


FIG. 1.3: Les trois modules

En effet, l'équipe « *Hypertextes et textualité électronique* » a développé son propre modèle à 3 modules, basé sur les recherches de David Piotrowski (Piotrowski, 1996). C'est ce modèle qui est représenté sur la figure 1.3.

Nous avons appliqué ce modèle à notre architecture technique pour obtenir ce qui est représenté par la figure 1.4. L'agent intelligent, appelé ici plus modestement « moteur hypertextuel », est responsable de la présentation et de la consultation. L'enregistrement étant sous le contrôle de la structure apportée par XML.

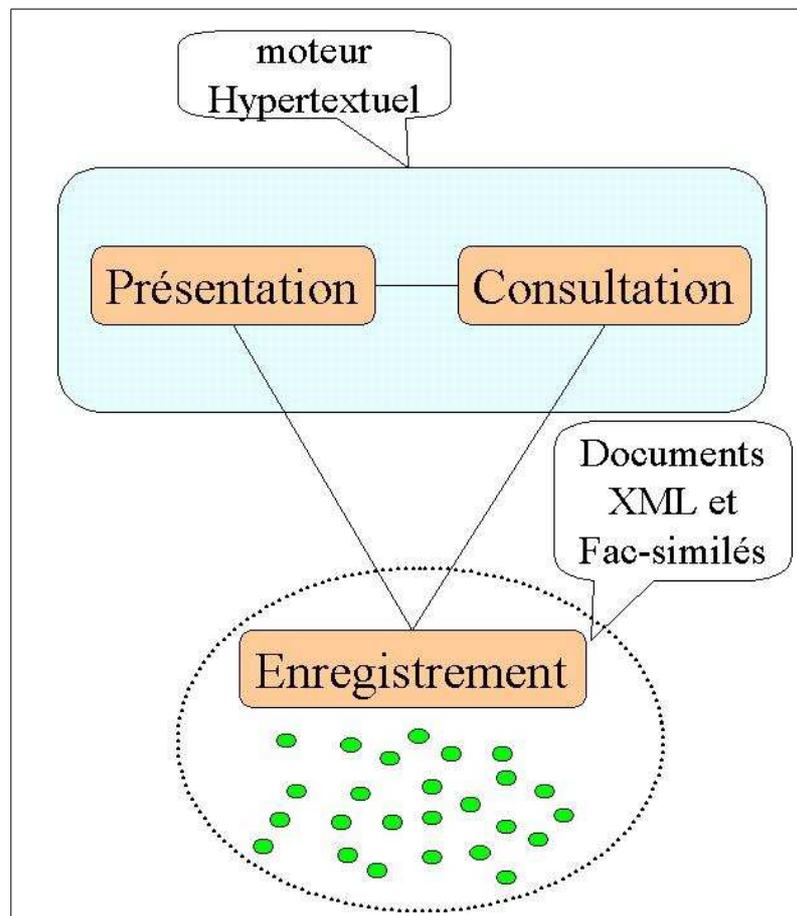


FIG. 1.4: Le moteur hypertexte vu selon notre modèle

## 2. Parcours

Intimement lié au dispositif de navigation, un deuxième dispositif s'intéresse plus particulièrement à l'analyse des usages qui sont faits du corpus et à la manière dont les lectures s'effectuent dès lors qu'elles sont libérées du support matériel d'une part et que l'esprit du lecteur peut réellement suivre son propre chemin d'autre part, grâce au dispositif hypertextuel décrit précédemment.

Nous l'avons baptisé « lectochromie » car il n'existe pas à notre connaissance de dispositif équivalent ni encore de mot pour le qualifier. La lectochromie doit être comprise comme un outil de synthèse de parcours de lectures à travers un corpus hypertextuel, pour la mise en évidence des représentations que les lecteurs ont cherché à construire, ou plus simplement ont construit au cours de leur « butinage » à travers le corpus.

Nous pouvons reprendre la présentation de l'architecture hypertextuelle de *C<sup>o</sup>L<sup>i</sup>Sciences* faite dans la section précédente, pour lui adjoindre la vision des parcours. Sur le schéma précédent, nous pouvons combiner la triple possibilité de parcours : fac-similé, texte ou notionnels. C'est donc finalement ce modèle que nous avons présenté de manière détaillée, en ajoutant la triple vision au travers des triples parcours, sur le schéma de la figure 1.5

## 3. Sens et cognition

Le troisième dispositif sera la mise en relation des deux premiers, de manière à donner au réseau hypertexte certaines des qualités d'un réseau neuronal. Si nous reprenons la figure 1.4 présentée précédemment, nous pouvons voir de manière synthétique le dispositif hypertextuel qui anime *C<sup>o</sup>L<sup>i</sup>Sciences*. Une approche purement computationnelle pourrait conduire à implémenter un moteur hypertextuel disposant de technologies issues de

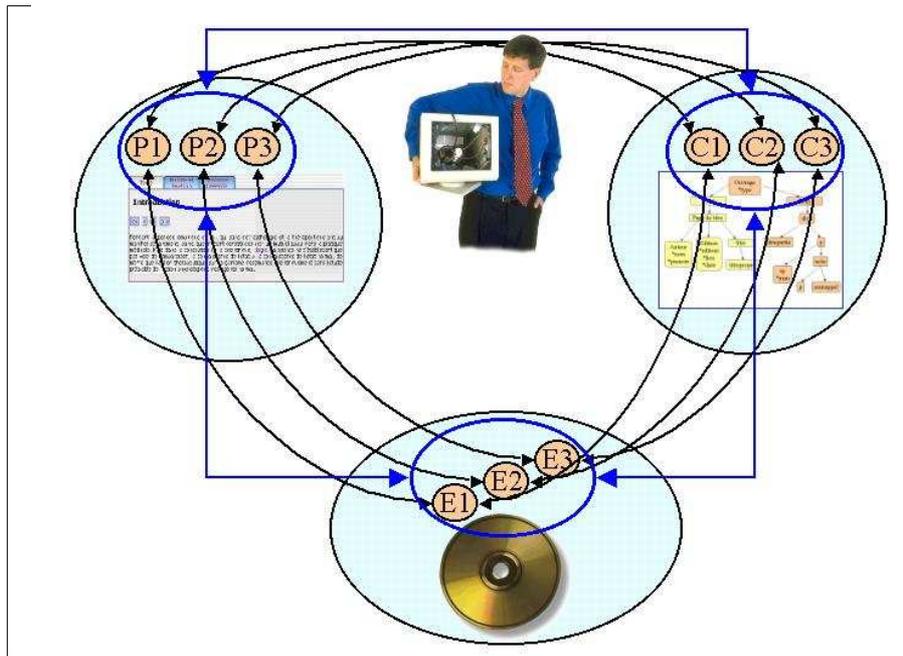


FIG. 1.5: Le triple modèle à trois modules

l'Intelligence Artificielle, ce qui aurait l'avantage de séparer les fonctionnalités du dispositif et les éléments du corpus. De cette manière, il serait possible de reproduire sur d'autres corpus ce qui aurait été mis en place une première fois dans *C<sup>o</sup>L<sup>i</sup>sciences*.

Mais le moteur hypertextuel n'est *a priori* pas capable de tirer parti des parcours réalisés par les internautes pour évoluer et s'enrichir. Bien entendu, il est toujours techniquement possible d'envisager d'ajouter cette possibilité au moteur, mais elle sera artificielle, parce qu'elle demande que le moteur hypertextuel soit capable d'enregistrer des informations sur le réseau hypertextuel, en dehors de ce réseau. Ce qui conduirait à avoir un moteur hypertextuel spécifique pour chaque corpus.

De cette réflexion est née tout d'abord l'idée de doter le réseau hypertextuel de capacités à enregistrer les parcours des lecteurs, ce qui s'est poursuivi par l'idée que le réseau hypertextuel pourrait aussi être considéré comme un

réseau neuronal doté de fonctions d'apprentissage. Ce n'est plus le moteur qui s'enrichi des lectures, c'est directement le corpus.

Les parcours de lecture, par un phénomène assimilable à la rétro-action, enrichissent le réseau hypertextuel, qui lui-même en retour ouvre de nouveaux parcours sur la base de ces nouvelles connaissances. Lecteur et hypertexte interagissent, ouvrant ainsi de nouvelles voies, symbolisées par ces nouveaux parcours.

Le dispositif *CoLiSciences* pourrait être considéré comme un catalyseur qui conduit un lecteur à se poser des questions, pour lesquelles il cherche une solution dans les liens qui lui sont présentés. Le problème du dispositif est d'identifier les liens qui proposent une bonne réponse, problème d'autant plus difficile à résoudre que la question n'est pas connue et que la même page peut susciter plusieurs questions. Si cette identification pouvait être rendue possible, il faudrait ensuite que les « bons » liens soient mis en évidence pour que le prochain lecteur puisse en bénéficier. Nous pouvons aussi envisager de faire apparaître ou disparaître certains liens, suivant leur pertinence, évaluée par les parcours des lecteurs précédents.

## 1.2 Cadre de référence théorique

Nous nous appuyons sur un certain nombre de travaux, les uns à visée théorique, les autres d'inspiration plus nettement appliquée, tous traduisant le caractère interdisciplinaire de cette recherche.

Du côté théorique, il s'agira notamment de faire le point en ce qui concerne les modèles sémantiques (Georges Vignaux), mais aussi ceux du Web sémantique menés par le W3C. En effet ces derniers prennent une position nettement

en contraste avec l'hypertexte tel qu'il est actuellement mis en œuvre sur la toile et tendent vers la mise en œuvre de systèmes locaux d'inférences.

Du côté pratique et opératoire, on s'inspirera des recherches sur l'hypertextualité menées en collaboration avec l'équipe « *Hypertextes et textualité électronique* » (David Piotrowski, Marc Silberstein, Georges Vignaux) de la Maison des Sciences de l'Homme - Paris Nord (CNRS-Paris), partenaire de I3M dans le cadre du Programme *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*. Rappelons les trois « idées régulatrices » qui sous-tendent les sciences cognitives, telles qu'elles sont présentées dans le recueil de textes fondateurs d'Aline Péliissier et Alain Tête (Péliissier et Tête, 1995). Pour chaque idée, nous pouvons faire correspondre des fonctions ou des usages de *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*.

### **A. Première idée régulatrice : la simulation**

*C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* est un site Web, mais il comporte une particularité très intéressante : le système hypertextuel utilisé pour générer les parcours à la demande de l'internaute se comporte comme un automate programmable. Les parcours enregistrés sont en quelque sorte le « code source », le programme de l'automate. Quand un utilisateur navigue page par page, il code pas à pas l'automate, quand un parcours est affiché on exécute programme automatiquement, simulant une navigation.

C'est la partie à développer pour permettre l'identification notionnelle.

### **B. Deuxième idée régulatrice : la rétroaction**

L'enregistrement des parcours, au fur et à mesure de la navigation, permet de générer un système de rétroaction qui enrichit chaque nouvelle lecture des enseignements des lectures précédentes.

Comment développer cette idée de *feedback* pour participer à l'enrichissement du site, quelles modalités d'enrichissement sont envisageable ? En particulier il serait déjà intéressant de faire le distingo entre les enrichissements apportés volontairement par un lecteur et ceux apportés, non pas à son insu, mais sans sa participation active au processus d'enrichissement. C'est par exemple la différence entre un système où il est proposé d'ajouter ses propres commentaires et un autre qui analyserait la fréquentation de certaines pages pour les mettre en avant<sup>3</sup>.

Pour aller encore plus loin sur cette idée de rétroaction entre le système *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* et ses lecteurs, nous pourrions imaginer de ne plus simplement afficher les parcours tels qu'ils ont été réalisés par chaque lecture individuelle, mais plutôt les synthétiser pour mettre en évidence les « grandes routes » qui traversent *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, les chemins les plus souvent empruntés.

### C. Troisième idée régulatrice : l'information codée

Le codage de l'information donne lieu à deux approches, la première est liée à l'annotation notionnelle et le deuxième en découle presque naturellement puisqu'il s'agit de l'utilisation du langage de balisage XML pour coder les textes et surtout, pour coder les notions. De cette manière, nous pourrions considérer XML comme une passerelle entre le texte, les mots et le sens. XML structure l'écrit.

---

3. Nous pouvons faire un parallèle avec des sites commerciaux. D'un côté avec le site Amazon de vente en ligne de livres, vidéos, CD, etc., à l'adresse <http://www.amazon.fr> pour le site français, qui propose à ses acheteurs de commenter les produits proposés. À l'opposé nous trouvons le modèle du moteur de recherche Google, qui analyse les liens cliqués par les internautes, en fonction des mots clefs spécifiés dans leur recherche. Les sites les plus souvent choisis « remontent » ainsi automatiquement dans les classements.

## 1.3 Éléments méthodologiques

### A. Finalité de cette recherche

Comme souvent dans le cadre de la recherche action, ce travail de recherche s'est construit au fil du temps, allant du général au particulier. Le sujet général correspondait plus à un ensemble d'interrogations personnelles qu'à une véritable construction, qui est venue plus tard, quand toutes les pièces ont été précisées et isolées, pour être finalement assemblées. Sans finalité pratique, les recherches sont validées par la communauté universitaire et non par le terrain. La finalité pratique de cette recherche, qui en quelque sorte permettrait de la valider par le terrain, se traduit par l'aboutissement du projet *C<sup>o</sup>l<sup>i</sup>sciences*, mais la finalité n'est pas uniquement pratique. Nous nous sommes situés dans une articulation étroite des deux lieux : académique et terrain. Pour parvenir à articuler la volonté de répondre aux problèmes qui se posent sur le terrain et l'exigence universitaire de construction d'un savoir objectif, nous pouvons distinguer trois étapes pour caractériser notre démarche :

- action
- analyse de corpus
- études expérimentales

#### 1. Action

Il ne faut pas entendre « observation » dans un sens naturaliste, comme le dit Gérard Vergnaud,

*« On n'a de chance de connaître vraiment quelque chose que si on essaye de le transformer. (Vergnaud, 1994) »*

En recherche action le chercheur n'est pas spectateur mais acteur.

La recherche action connaît deux types de prolongements. Sur le plan de la recherche, elle se transforme en exploration systématique au cours de la seconde phase de la démarche. Sur le plan professionnel, elle conduit à la construction d'outils pédagogiques qui, élaborés avec des maîtres participant aux recherches-actions, sont ensuite proposés aux maîtres en général.

## **2. Analyse de corpus**

Les idées nées au cours de la phase précédente vont prendre corps ici, conduisant à la constitution d'un corpus. Une analyse de corpus, aussi objective et précise soit-elle, demeure insatisfaisante : les propositions produites dans ce cadre relèvent plus de l'interprétation que de la preuve. Conclusions pour la deuxième phase, elles deviennent hypothèses dans la troisième. Pour atteindre le statut de preuve, nous devons passer à une autre phase de la recherche, de nature expérimentale.

## **3. Études expérimentales**

La mise en œuvre des conclusions de la deuxième phase à travers le projet *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* donne lieu à différents types d'expérimentations. Le projet lui-même est une sorte de macro-expérimentation, notre équipe multidisciplinaire, les outils techniques d'une part : plate-forme de publication, numérisation, outils d'édition comme les outils cognitifs d'autre part : lectochromie, construction de parcours, participent tous de l'étude expérimentale de construction d'un système hypertextuel savant. Plusieurs micro-expérimentations ont aussi été mises en œuvre autour de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, qui visaient à valider des choix particuliers de publication, ou d'analyse de textes, sémantique ou statistiques et qui ont donné lieu à de nouvelles constructions d'outils, techniques comme théoriques.

## CHAPITRE 2

# CADRE ÉPISTÉMOLOGIQUE

---

*« Dans la pensée scientifique, la médiation de l'objet par le sujet  
prend toujours la forme du projet. »*

Gaston Bachelard

*Le nouvel esprit scientifique*

## Sommaire

---

<b>2.1</b>	<b>Les questions à résoudre . . . . .</b>	<b>27</b>
A.	Problématique . . . . .	27
<b>2.2</b>	<b>Organisation par thèmes de recherche . . . . .</b>	<b>27</b>
A.	Intelligence artificielle . . . . .	28
B.	Cartographie . . . . .	38
C.	Orientation et repérage . . . . .	41
D.	La linguistique cognitive . . . . .	44
E.	La communication . . . . .	45
F.	L'informatique linguistique . . . . .	47
G.	La sémiotique . . . . .	47
<b>2.3</b>	<b>Organisation de la thèse . . . . .</b>	<b>49</b>

---

## 2.1 Les questions à résoudre

### A. Problématique

Cette recherche doit nous amener à aborder un ensemble de questions relevant de :

**L'Intelligence artificielle** L'intelligence artificielle (I.A.) sera étudiée à travers son rapport à l'élaboration de modèles de la cognition et de la mémoire, de mécanismes d'interactions entre niveaux de connaissances dans la compréhension de textes, de représentation de situations cognitives et de mécanismes d'apprentissage.

**La linguistique cognitive** La linguistique cognitive sera abordée à travers des recherches sur les opérations fondamentales de construction des représentations et de traitement du sens.

**La communication** La communication nous amènera à procéder à l'étude de modèles de la communication langagière et textuelle.

**La sémiotique** L'informatique linguistique et ses techniques seront utilisées pour la mise au point d'analyseurs syntaxico-sémantiques et morphologiques, de procédures de validation. Ceci donnera lieu à un certain nombre d'expérimentations sur corpus.

## 2.2 Organisation par thèmes de recherche

Les thèmes de recherche que nous visons à développer découlent de la mise en place d'un dispositif de communication mettant en œuvre des technologies numériques. Nous aurions pu mentionner la cybernétique plutôt que la terminologie *a priori* plus vague de « technologies numériques », mais cela aurait en fait réduit notre champ de recherche qui tentera d'intégrer plus

largement les technologies de communication et de numérisation, dans leur sens le plus large possible. En cela, nous nous associons à la citation suivante :

« *La technologie doit envisager l'universalité des objets techniques. En ce sens, la Cybernétique est insuffisante. (Simondon, 1969, page 48)* »

## A. Intelligence artificielle

### 1. Définition

L'intelligence artificielle (IA) est définie par l'un de ses créateurs, Marvin Lee Minsky, comme

« *La construction de programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisantes par des êtres humains car elles demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique. (Minsky, 1987)* »

Cette discipline est liée à la question de savoir si les machines peuvent penser, sur laquelle il faut relever les recherches de Alan Turing<sup>4</sup> qui ont abouti à la modélisation d'un test qui permettrait d'affirmer ou non qu'une machine est capable de penser.

« *Je propose d'étudier la question, "Est-ce que les machines peuvent penser ?". Ceci devrait commencer par la définition des termes "machine" et "penser". Les définitions devraient être formulées de manière à refléter autant que possible l'utilisation habituelle de ces termes, mais cette attitude est dangereuse. Si la signification des mots "machine" et "penser" doit être trouvée en examinant comment ils sont habituellement utilisés il*

---

4. Il pourrait être intéressant de remarquer que Alan Turing a été l'élève de Wittgenstein dont nous avons mentionné le *Tractatus Logico-Philosophicus* dans la section 4.1 du chapitre 4 consacré à l'hypertexte qui commence page 67.

*est difficile d'échapper à la conclusion que la signification et la réponse à la question "Est-ce que les machines peuvent penser?" devrait être trouvée dans une enquête statistique telle qu'un sondage de l'institut Gallup. Mais ceci est absurde. Au lieu d'essayer une définition de ce type, je devrais plutôt remplacer la question par une autre, qui lui serait liée de près et n'utiliserait pas de mots ambigus. La nouvelle forme du problème peut être exprimé suivant les termes d'un jeu que nous appellerons "le jeu de l'imitation". Il se joue avec trois personnes, un homme (A), une femme (B), et un enquêteur (C) qui peut être de n'importe quel sexe. L'enquêteur est dans une pièce différente des deux autres. Le but du jeu pour l'enquêteur est de déterminer lequel des deux autres est l'homme et lequel est la femme. Il les connaît par des étiquettes X et Y, à la fin du jeu il dit soit "X est A et Y est B" ou "X est B et Y est A". L'enquêteur peut poser des questions à A et B. Nous nous posons maintenant la question "Qu'est-ce qui se passerait si une machine prenait la place de A dans ce jeu?" Est-ce que l'enquêteur se tromperait aussi souvent quand le jeu est fait sous cette forme ou quand le jeu se fait avec un homme et une femme? Ces questions remplacent celle originale "Est-ce que les machines pensent?"<sup>5</sup>(Turing, 1950) »*

---

5. I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous, If the meaning of the words "machine" and "think" are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, "Can machines think?" is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words. The new form of the problem can be described in terms of a game which we call the 'imitation game.' It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A." The interrogator is allowed to put questions to A and B. We now ask the question, "What will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often when the game is played like this

L'Intelligence Artificielle est un domaine qui a suscité de nombreux débats. Après les promesses des débuts, la désillusion de certains a donné lieu à un certain nombre de controverses qu'il est possible de retrouver jusque dans la définition qu'en donne *Le dictionnaire des sciences*. C'est d'une part :

« Un "artifice", c'est-à-dire un art consommé qui fait illusion en produisant des leurre fabriqués tout exprès pour nous tromper en laissant accroire que les machines seraient effectivement intelligentes, sans toutefois nous duper vraiment. (Serres et Farouki, 1997, page 479) »

Cette définition fait référence aux logiciels dont le comportement mime l'intelligence humaine « de l'extérieur », mais qui ne sont en fait basés que sur les techniques traditionnelles de l'informatique, c'est-à-dire faisant appel à la logique, aux mathématiques et à l'algorithmique.

« Cependant l'intelligence artificielle se distingue des autres branches de l'informatique en ce qu'elle aborde des problèmes mal formulés, pour lesquels il n'existe pas de solution universellement admise. (Serres et Farouki, 1997, page 480) »

Les progrès faits en matière d'algorithmique font que certains problèmes résolus par l'IA hier sont aujourd'hui résolus par des techniques paradoxalement plus traditionnelles, tout en étant plus modernes.

C'est peut-être quand il doit être fait appel à l'imagination que nous entrons tout à fait dans le champ de l'IA, en particulier pour ce qui se rapproche de la création artistique.

« Certains programmes d'intelligence artificielle font reposer le travail de l'imaginaire à l'œuvre chez le compositeur sur une modélisation de la mémoire où se conjugueraient deux mécanismes heuristiques :

---

as he does when the game is played between a man and a woman? These questions replace our original, "Can machines think?"

*un mécanisme d'évocation, simulant les processus de réminiscence par lesquels un contexte [...] remet en mémoire des épisodes anciennement vécus, et un mécanisme de recombinaison, qui mettrait bout à bout et ajusterait les fragments sortis de la mémoire[...] L'élaboration, la simulation et l'évaluation de ces critères heuristiques relève précisément de l'intelligence artificielle. (Serres et Farouki, 1997, page 480) »*

La liste ci-dessous, non exhaustive, des travaux de l'IA montre comment nous partageons nombre d'objectifs :

- Les systèmes experts et la résolution de problèmes
- La représentation des connaissances
- Les modèles linguistiques développés pour la compréhension du langage naturel (par les machines) et la traduction automatique.

Bien que dans l'architecture *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* nous ne faisons pas à proprement parler appel à l'IA, nous faisons appel à de nombreux procédés largement inspirés de celle-ci.

En premier lieu, le moteur hypertextuel de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* est construit sur le modèle d'un automate fini<sup>6</sup> qui réagit en suivant des règles similaires à celles implémentées par des réseaux de Petri<sup>7</sup>, comme nous le verrons dans la section D., page 118 du chapitre 6. Cet automate peut-être échangé avec un autre, ou agrémenté d'autres fonctions qui pourraient aller jusqu'à l'utilisation d'un moteur d'inférences comme on en trouve dans les systèmes experts. En second lieu, cet automate est capable, avec plus ou moins de précision de

---

6. Un automate fini est constitué d'une entrée, d'une sortie, d'un ensemble d'états et de transitions étiquetées entre ces états

7. Un réseau de Petri permet d'étudier des systèmes dynamiques complexes. Ils ont été proposés dans les années 60 par Carl Adam Petri, puis développés au MIT, et sont maintenant utilisés pour spécifier, modéliser et comprendre les systèmes (dans lesquels plusieurs processus sont interdépendants). Le formalisme des réseaux de Petri est un outil permettant l'étude de systèmes dynamiques et discrets.

mettre en œuvre des caractéristiques que l'on retrouve généralement dans des systèmes relevant de l'IA : représentation des connaissances en particulier.

## 2. Système expert

Les systèmes experts interviennent dans le champ de l'I.A. de plusieurs façons, pour l'usage qu'il peut en être fait, mais aussi pour les outils qu'ils nécessitent de mettre en place. Nous pourrions de la même manière baser notre étude plutôt par rapport à l'utilisation d'un système expert pour contrôler les parcours érudits d'un lecteur à travers un corpus numérique, ce qui irait sans doute dans la direction d'une assistance à l'identification des notions traitées par le texte, mais pourrait aussi s'étendre à la propositions de voies de découvertes partant du parcours initié par ce même lecteur. Une autre voie à explorer est celle des outils nécessaires à la construction d'un système expert. Pour cela il faut généralement décrire les connaissances mobilisées par un expert humain, pour pouvoir les reformuler de façon à ce qu'elles soient accessibles au système expert et cela demande de commencer par l'énumération des mots du domaine.

Les ensembles de mots relatifs aux différents champs de compétences sont regroupés dans ce qui est appelé, en termes techniques, des ontologies de domaine. Il ne faut pas confondre cette notion avec celle d'ontologie prise au sens de système des concepts, car ces ontologies de domaines ne forment pas à proprement parler des systèmes, mais plutôt des inventaires. Une fois adoptée une ontologie de domaine, il faut transcrire des fragments de raisonnements d'experts sous forme de règles utilisant les termes de l'ontologie, c'est-à-dire du vocabulaire des experts.

Il y a dans notre dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* l'utilisation d'une expertise humaine qui pourrait être l'objet de la création d'une ontologie de domaine, ce qui

pourrait donner lieu ensuite à l'implémentation d'un système expert. Ceci pourrait se faire en effet à travers les résultats de l'analyse notionnelle qui a été faite manuellement et implémentée dans le réseau hypertexte.

### 3. Traitement du langage naturel

**Introduction** Les recherches portant sur le langage naturel visent principalement à faciliter les interactions entre hommes et machines, au travers d'interfaces (IHM - Interface Homme Machine). Nous « parlons », nous « utilisons » des langages différents. Celui des hommes est complexe et supporte de nombreuses nuances, de nombreuses interprétations, dont les machines sont bien incapables. Les interfaces graphiques sont un exemple d'apport des recherches en IHM, il n'est plus indispensable d'apprendre les commandes de l'ordinateur, une représentation graphique des ressources utilisables est proposée à l'utilisateur, les manipulations avec la souris sont censées ou prétendues intuitives et donnent accès aux commandes disponibles sur les ressources pointées. Il n'y a plus de commande cachée, tout est accessible « naturellement » en quelques clics de souris. Pourtant tout n'est pas si simple, le « bureau » symbolisé sur un écran d'ordinateur est vertical, un bureau réel est horizontal. Sur un « bureau » symbolisé il est possible de trouver, pêle-mêle, des documents, des applications (donc des objets virtuels, des verbes en quelque sorte), le disque dur de l'ordinateur et même quelquefois l'ordinateur lui-même. L'autre partie de l'interface se trouve donc dans le cerveau de l'utilisateur, c'est lui qui doit réinterpréter ses besoins pour les traduire en actions que l'ordinateur sera capable d'effectuer. Dans l'imaginaire populaire, le développement ultime de l'IHM c'est HAL<sup>8</sup>

---

8. Prenons quelques lignes pour faire une digression à propos de ce nom et de la « légende urbaine » qui s'y attache. Si on avance d'un rang dans l'alphabet chaque lettre de « HAL », on obtient « IBM ». Ce petit jeu de lettres fait dire à certains, que HAL est l'ordinateur IBM du futur, une place devant. Arthur C. Clarke, co-scénariste et auteur de la nouvelle à l'origine

9000, l'ordinateur du film de Stanley Kubrick « 2001 : Odyssée de l'Espace<sup>9</sup> ». Avec HAL, il suffit de s'exprimer à haute voix et il comprend, interprète, propose des solutions et même des jeux, en particulier les échecs, à son interlocuteur humain.

**Les enjeux** Le traitement du langage humain est donc avant tout une problématique d'interface : faciliter l'accès de tous aux ordinateurs en rendant leur interface capable de comprendre notre langage. Mais il y existe d'autres enjeux tout aussi importants. Le traitement du langage naturel donne aussi accès à la traduction automatique par un ordinateur, à l'écriture de résumés, voire de commentaires, toujours de manière automatique.

#### 4. Pensée et apprentissage

Ce qui nous amènerait ensuite à essayer de répondre à des questions sur les machines auto-apprenantes, par l'étude du comportement de notre dispositif face à un texte nouveau et non balisé.

*« Pourquoi les ordinateurs d'aujourd'hui n'apprennent-ils pas comme les gens le font ?*

**la Mémoire** *n'est pas un endroit pour conserver les descriptions comportementales faites par un observateur au cours du temps (schémas, plans, grammaires), mais une compétence pour rétablir et combiner les manières précédentes de se comporter.*

**Apprendre** *n'est pas un processus isolé, mais une partie intégrale de chaque perception ou mouvement.*

---

du film, affirme que c'est pure coïncidence s'il est possible de faire une connexion entre ces deux noms et qu'il était trop tard pour revenir en arrière quand cela avait été mis en évidence. « HAL » signifie « Heuristically programmed ALgorithmic computer ». Toutefois, IBM n'est pas absent du film : les panneaux de l'ordinateur de la navette qui vient s'arrimer à la station, l'écran sur le bras de la combinaison spatiale de Dave, les écrans portables en portent le logo.

9. 2001 : A Space Odyssey, film américain de 1968

**les Représentations** *ne sont pas simplement des schémas syntaxiques, mais relu et réinterprétés à chaque utilisation par un commentaire qui fournit un contexte pour les interpréter.*

*Parler c'est concevoir et comprendre.*<sup>10</sup>(Clancey, 1991)

Les premiers programmes de jeu d'échecs ont essayé de reproduire la manière de jouer des humains, en utilisant des modélisations à base de systèmes experts et de règles du type de celles qu'un grand maître met en œuvre quand il joue. Mais finalement les progrès de l'algorithmique, conjugués à la puissance de traitement croissante des machines permettent aujourd'hui d'obtenir des résultats bien meilleurs. L'ordinateur Deep Blue de la compagnie IBM, qui a battu le champion du monde humain Gary Kasparov en 1997 n'a pas appris à jouer aux échecs, il ne fait pas des combinaisons stratégiques comme un humain, il utilise au mieux ce qu'il sait faire le mieux : le traitement heuristique de données, ces données étant en particulier des bibliothèques de parties déjà jouées, le but étant de reconnaître dans la partie qu'il est en train de jouer une combinaison qui est déjà apparue dans une partie déjà jouée.

## 5. Approche connexioniste

Le modèle connexioniste est né de réflexions sur le cerveau humain, qui est capable d'interpréter à partir de données non exhaustives et mouvantes une

---

10. Why Don't Today's Computers Learn the Way People Do ?

**Memory** is not a storage place for descriptions of how the behavior appears to an observer over time (schemas, plans, grammars), but a capacity to directly reenact and compose previous ways of behaving.

**Learning** is not a separate process, but an integral part of every perception and movement.

**Representations** are not just syntactic patterns, but re-perceived and given new meaning in every use by commentary that supplies a context for interpreting them.

Speaking is conceiving.

solution, à laquelle il peut aussi donner un niveau de fiabilité. Ce modèle se trouve en quelque sorte à l'opposé de la machine de Von Neuman, qui sépare mémoire et programme, et qui s'interdit toute capacité de généralisation. La machine de Von Neuman accepte un certain nombre de données en entrée, procède à des calculs arithmétiques à partir de celles-ci et fournit des données de sorties, nous ne retrouvons dans la mémoire que les données entrées ou qui en est une conséquence directement calculable. Les connexionistes veulent trouver des modèles informatiques capables de généralisation.

Les premières idées furent développées par les cybernéticiens McCulloch et Pitts en 1943. Ils montrèrent alors que des fonctions logiques pouvaient être calculées par des automates simples constitués en réseau. Par analogie avec le cerveau ces unités simples furent désignés comme « *neurones formels* ». Les neurones sont capables de réaliser des « échanges synaptiques ». Un « neurone formel » se comporte comme un additionneur à seuil de toutes les valeurs qui lui sont apportées par les autres neurones, chaque participation d'un neurone étant modulée par les « poids synaptiques ».

Les problèmes fortement combinatoires, se modélisant par un ensemble de contraintes sur des objets, gagnent à être distribués. Dans notre cas, le corpus des ouvrages de *CoLiSciences* étant modélisé sous la forme d'un réseau hypertextuel de paragraphes, chacun étant doté d'attributs et de contraintes, l'approche connexioniste s'impose de façon presque naturelle. Le corpus ne se résume plus à une collection de textes étiquetés, il devient une base de connaissances qui s'enrichi au cours du temps. Pour cela, il faut le doter de certaines capacités, inspirées des sciences cognitives et de l'approche connexioniste.

Tout d'abord, l'approche computationnelle nous a poussé à mettre en œuvre des parcours au travers d'un dispositif qui pourrait être assimilé à

un automate fini. En fonction des étapes précédentes et de déclencheurs prédéterminés, l'automate décide d'afficher tel ou tel paragraphe de telle ou telle manière, fac-similé, texte ou texte analysé. Ce premier degré d'autonomie de notre système permet de ne pas avoir un texte balisé par des liens hypertextuels, c'est l'automate qui décide de placer les liens. Si on se réfère aux théories des systèmes hypertextuels, que nous verrons dans la section 4.2, page 71 du chapitre 4, cela veut dire que nous respectons un modèle à trois modules et que le module d'enregistrement est bien distinct et indépendant du module de présentation. En pratique cela veut dire que nous pouvons complètement modifier l'interface, voire l'ergonomie du site en intervenant sur la couche responsable de la présentation ou celle de la consultation sans toucher à son contenu qui dépend de la couche enregistrement.

Dans un premier temps, nous voyons donc que la théorie computationnelle nous a permis de mettre en place un système hypertextuel conforme à la théorie. Mais les apports des sciences cognitives ne s'arrêtent pas là. En se référant à la théorie connexioniste, nous pourrions voir dans notre réseau de nœuds hypertextes autre chose qu'un ensemble de paragraphes :

*« Le cerveau allait devenir une fois encore la source de métaphores et d'idées pour d'autres domaines des sciences cognitives. Ici, le point de départ n'est pas une description symbolique abstraite, mais une armée de constituants simples et non intelligents qui, comme des neurones, expriment des propriétés globales intéressantes lorsqu'ils sont reliés. (Varela, 1996, page 57) »*

Notre point de départ est le même : un réseau de paragraphes de texte, simples et non intelligents. Ces paragraphes ne disposent d'aucune autonomie, ils prennent sens uniquement quand ils sont assemblés entre eux et alors, comme les neurones, ils expriment des propriétés globales (dans notre

cas, un texte porteur de sens) au travers de la relation qui s'est instituée entre eux (la connexion ayant lieu *a priori* par les notions issues de l'analyse manuelle, et *a posteriori* par l'activité de lecture des internautes et la création de parcours de lecture plus ou moins redondants). La mise en place d'une approche connexionniste reviendrait à reproduire avec les paragraphes issus des ouvrages du corpus ce qui est décrit un peu plus loin sur la même page :

« *Si deux neurones essaient de s'activer au même moment, leur lien est renforcé ; autrement il est diminué. (Varela, 1996, page 57) »*

Il faudrait donc que le dispositif soit à même de renforcer les liens entre les paragraphes d'un même parcours, dans l'espoir de voir émerger un « méta-parcours » après un certain nombre d'expérimentations. Mais pour cela, encore faut-il définir ce qui fait la force d'un lien et comment elle est manifestée au lecteur.

## **B. Cartographie**

La cartographie est utilisée traditionnellement en géographie, mais elle n'est pas réservée à ce seul domaine. La carte est une représentation spatiale du concept de territoire. Partant de là, il est possible d'utiliser la cartographie pour permettre des représentations spatiales de concepts quels qu'ils soient, ce qui est fait depuis longtemps comme par exemple avec les planches anatomiques. Toutefois, avec le territoire comme avec l'anatomie, la cartographie permet de représenter des concepts physiques, c'est son utilisation pour représenter des concepts qui n'ont pas de forme physique, qui permet donc de les concrétiser et d'en permettre la visualisation qui va trouver tout son intérêt dans notre cas. Une analyse bibliographique (Roger *et coll.*, 1998) réalisée pour le *Programme d'Aide à la Recherche en INFOrmation (PARINFO)*,

dans le cadre d'une recherche menée à l'ENSSIB donne un large panorama de l'utilisation de ce concept. Cette analyse couvre tour à tour plusieurs domaines de recherche, il y est fait tout d'abord mention des sciences de l'éducation et du regain d'intérêt suscité par l'utilisation de la cartographie dans les années 1980. Après que Novak propose en 1984 les « concept maps »<sup>11</sup>. Par ailleurs, le même concept de représentation graphique a été repris en 1992 sous la forme d'arbres de connaissances par Authier et Lévy (1998), mais dans ce cas ce n'est plus une connaissance qui est cartographiée pour être plus accessible, ce sont les connaissances d'un individu ou d'un groupe qui sont représentées. Ce n'est plus un savoir à transmettre, mais un état des lieux. Nous pouvons aussi relever que de nombreuses représentations spatiales de connaissances scientifiques sont apparues à la suite des cartes conceptuelles : trames conceptuelles, réseaux conceptuels, Vee-diagrams, Vee maps, knowledge mapping ou K-maps, diagrammes conceptuels, etc.

Pour continuer par les sciences cognitives, nous pouvons relever qu'elles font également usage de représentations spatiales de concepts : réseaux sémantiques (fin des années 60) utilisés pour la représentation des connaissances expertes et de la mémoire sémantique, graphes conceptuels (années 70) pour la compréhension du langage naturel. La représentation graphique des connaissances pose des problèmes de formalismes, des types de liens à prendre en compte et de la manière à suivre pour représenter les méta-connaissances. Les sciences de l'information font, depuis longtemps, appel à des représentations spatiales de concepts avec les terminogrammes, les schémas fléchés des thésaurus graphiques. Le développement des thésaurus graphiques, se révèle plus lent et limité ainsi que le montre le diagramme pré-

---

11. (Jacobi et Boquillon, 1994) font observer que « map » peut se traduire de deux façons : carte conceptuelle ou carte de concepts ; on trouve également l'expression « mapping » qui, elle, est intraduisible. Selon ce même auteur, mapping comme projet en acte est probablement ce qu'il faudrait comprendre et interroger pour analyser les cartes. » (ibid. p.13).

senté sur la figure 2.1 tiré de Bertrand-Gastaldy et Davidson (1986, page 228).

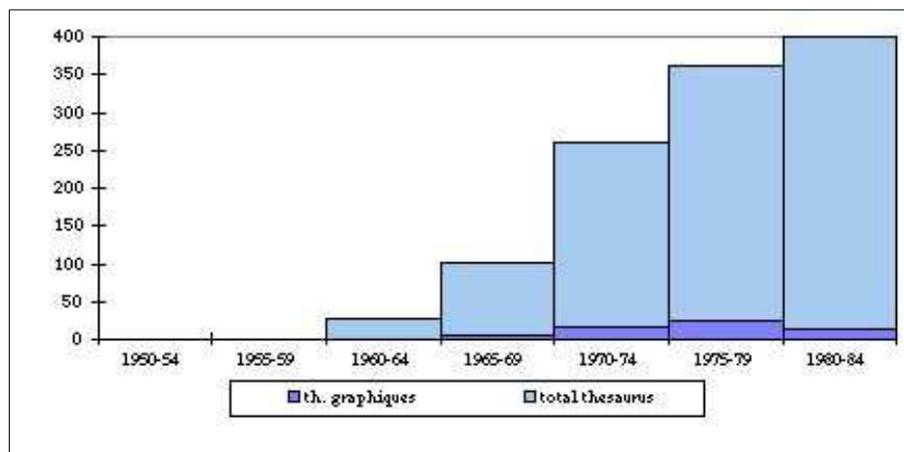


FIG. 2.1: Développement comparé des thésaurus

D'autre part, les nombreux travaux suscités par les thésaurus ont porté sur leurs méthodes de construction mais n'ont pas pris en compte leurs usages, qu'il s'agisse de ceux des professionnels de l'information ou de ceux des utilisateurs. Selon certains auteurs, la consultation du thésaurus est l'une des stratégies de recherche des professionnels qu'il faut implémenter sur les systèmes destinés au grand public, ceci sans que ne soient précisées les stratégies d'exploration et de consultation de ces thésaurus. Les premiers thésaurus grands publics ont été des CD-Roms transposés directement de l'outil papier, sans nouvelle fonctionnalité ni même d'adaptation.

*« Thésaurus et cartes conceptuelles présentent des points communs mais aussi des différences. Tous deux visent à représenter un domaine de connaissances et tous deux privilégient les relations hiérarchiques. Cependant, les thésaurus possèdent un caractère normatif que les cartes conceptuelles ne présentent pas. Loin de se vouloir « la » représentation de référence, celles-ci ont surtout valeur pour celui/ceux qui les créent. La question de leur lisibilité pour autrui ne se pose donc pas. Les tailles des*

*domaines pris en compte, dans l'un et l'autre cas, ne sont pas non plus les mêmes : Novak fait figurer une vingtaine de concepts sur une carte alors qu'un champ sémantique comporte beaucoup plus de descripteurs. Par contre, les thésaurus font appel à un très petit nombre de types de liens, standardisés, entre concepts alors que les cartes conceptuelles ne les normalisent pas. (Roger et coll., 1998) »*

Nous avons plusieurs questions à résoudre pour faire avancer notre recherche sur le domaine de la cartographie active telle que nous l'imaginons. En particulier comment formaliser les concepts et les traduire visuellement de manière à ce qu'ils soient immédiatement accessibles au lecteur. Par exemple, comment devons nous dessiner les parcours, et à l'intérieur de ceux-ci, comment représenter les temps de lecture, les sauts à travers le corpus. Il existe aussi des questions qui sont liées aux autres thèmes, hypertexte et désorientation par exemple : comment devons-nous articuler cartographie et lecture, en particulier comment chaque dispositif peut-il faire référence à l'autre, et faire voyager le lecteur entre les deux, ce qui va faire basculer le lecteur d'un plan de compréhension du texte à un plan de navigation à travers le corpus.

### **C. Orientation et repérage**

La question de la navigation suit logiquement celle de la représentation spatiale. Les problèmes d'orientation et de repérage sont incontournables dans tout ce qui touche à l'hypertexte, la cartographie est à la fois une manière de résoudre ce problème et une manière de le complexifier parce qu'elle peut en fait apporter une surcharge cognitive supérieure à la simplification apportée par la synthétisation de la navigation. Commençons par énoncer

le cycle à trois phases qui représente l'activité de l'utilisateur selon Rouet et Tricot (Rouet, 1992) (Tricot et Bastien, 9-11 Mai 1996) :

**Sélection de l'information par sélection d'un nœud** La phase de sélection met en jeu plusieurs processus, il s'agit pour le lecteur de gérer à la fois où il se trouve et où il voudrait aller, son activité de lecture et processus de traitement des informations relationnelles.

**Traitement de cette information** La phase de traitement s'effectue au niveau local du nœud (compréhension du contenu du passage lu), mais aussi à un niveau plus global, c'est ce qui conduira le lecteur, dans l'étape suivante, à évaluer si le passage lu contribue au but poursuivi et, dans ce cas, intégrer cette information à celles acquises aux cycles précédents.

**Évaluation de la pertinence du nœud sélectionné** Ce qui conduira le lecteur à se poser de nouvelles questions pour sélectionner le nœud suivant.

En essayant de faire une compilation des nombreux travaux qui cherchent à préciser quelles fonctionnalités il faut proposer pour atténuer le phénomène de désorientation, nous pouvons constater que ces travaux sont souvent parcellaires et de divers types. Certains parlent d'ergonomie de manière générale, certains sont plus adaptés à la navigation sur le Web. Il peut être intéressant de mettre au point telle fonctionnalité de manière isolée pour en mesurer l'impact précisément, mais il est sans doute préférable de chercher à comparer l'efficacité de plusieurs moyens d'orientation combinés, le cumul de l'impact de chacun étant sans doute inférieur à l'impact global quand les fonctionnalités sont présentes ensembles. Là encore, les différents travaux sont difficilement comparables et donnent des résultats parfois contrastés.

Nous avons relevé, toujours dans le même rapport, au travers des études comparatives et des recommandations de conception une liste de fonctionnalités qui sont sujettes à limiter la désorientation des utilisateurs :

- les retours arrière
- les visites guidées
- les signets (bookmarks)
- les cartes conceptuelles
- les vues panoramiques (fisheye views)
- la mise en place de repères dans le sommaire
- la sélection, par l'utilisateur, des passages qu'il veut approfondir.

Nous ne pouvons pas terminer sans noter que la cartographie ne fait tout de même par l'unanimité, Brown (1989) récuse tout intérêt aux cartes conceptuelles et autres navigateurs graphiques. A la recherche de fonctionnalités plus performantes et qui permettraient de s'en passer il propose principalement la sélection, par l'utilisateur, des passages qu'il veut approfondir. Ces passages viennent s'insérer dans le texte courant affiché à l'écran et se donnent à lire quasi séquentiellement à l'utilisateur. Il propose également la personnalisation de l'affichage pour des utilisateurs novices, en effet il ne faut pas oublier l'importance des différences individuelles, ce qui oriente les uns alourdit la navigation des autres. Il est aussi proposé d'utiliser une organisation hiérarchique reprise dans un système de menus et sous-menus.

Un autre aspect encore est le typage de liens testé par Baron, Tague-Sutcliff et Kinnucan (1996). Pour cela, ils définissent trois types de liens :

- sémantiques : tels que ressemblance, partie de, est une sorte de
- rhétoriques : définition, explication, illustration, continuation et récapitulation ;
- pragmatiques : mise en garde, prérequis, usages, exemples.

Une autre étude des cartes conceptuelles est présente chez Gaines et Show (1995), mais dans leur cas ils se sont plus attachés à en étudier la mise en œuvre et les usages possibles plutôt que de valider leur impact par des

expérimentations. Il faut toutefois noter que comme pour Rouet (1992), Rouet et Tricot (1995) ils relèvent l'importance des différences individuelles.

Parmi les principes de cohérence, trois visent à améliorer la cohérence locale, les deux autres la cohérence globale. Il convient :

- d'étiqueter les liens, d'indiquer les équivalences (lorsqu'un même nœud est présent dans plusieurs fenêtres),
- de préserver le contexte en montrant les nœuds voisins d'un nœud donné et leurs relations. de créer des unités de plus haut niveau d'information, les nœuds composites,
- de visualiser la structure en donnant une vue générale du domaine avec ses principaux sujets et leurs relations, par exemple au moyen de cartes conceptuelles.

Les trois derniers principes visent la réduction de la charge cognitive :

- fournir à l'apprenant des indicateurs de sa position, du chemin parcouru et des options possibles,
- fournir des indicateurs de direction et de distance,
- offrir une interface stable.

## **D. La linguistique cognitive**

La linguistique cognitive doit nous aider à comprendre comment fonctionne le langage, afin de comprendre les modes de pensée et les diverses façons de communiquer. Nous ne pourrons pas entrer dans le débat entre les principaux courants contemporains, qui se réclament de deux grands paradigmes théoriques :

- le paradigme classique du cognitivisme (appelé paradigme computationnel symbolique), qui a été adopté par les « grammaires formelles », et notamment par la grammaire chomskienne

- un paradigme alternatif (parfois appelé paradigme constructiviste), représenté principalement par les « grammaires cognitives » dont une figure centrale est Ronald (Langacker, 1987), mais auquel se rattachent également certains tenants de la linguistique fonctionnaliste, ainsi que des approches typologiques et diachroniques des langues.

Il pourra être fait appel à la linguistique pour offrir un autre éclairage à notre problématique à travers ses interactions avec d'autres disciplines, abordant le langage non plus à partir de l'étude spécifique de la structure et du fonctionnement des langues, mais en tant que faculté supérieure de l'espèce humaine mise en œuvre par des sujets, donnant lieu à l'activation de certaines zones du cerveau, et susceptible d'être simulée sur ordinateur.

## **E. La communication**

Les sciences de la communication ont pour objet l'étude de la communication, mais il n'y a pas une science de la communication puisque la communication fait appel à plusieurs disciplines. La communication est plutôt un objet de connaissance interdisciplinaire, au carrefour des disciplines traditionnelles et des savoirs récents liés à une formidable expansion.

*« La communication est l'action complexe résultant de l'interaction de trois actions archétypales : l'acte technique, par lequel un signal devient transmissible (le niveau A, chez Weaver), l'acte sémantique, par lequel le signal émis devient signification (niveau B), et l'acte pragmatique, par lequel le signal reçu induit un changement de comportement du récepteur (niveau C, chez Weaver, qui parle du problème de l'effectivité : l'effectivité du signal se définissant par le changement de comportement observable du récepteur). (Le Moigne, 1994, page 196) »*

La communication est au centre de beaucoup de nos préoccupations modernes, même si ses racines remontent à l'antiquité. A quelques exceptions près, les hommes ont toujours cherché à communiquer, à se rapprocher et à apprendre des autres cultures et la communication était au centre du développement démocratique dès le forum Romain. Interdite ou manipulée, elle a aussi été un outil de pouvoir et de soumission des totalitarismes. Dans le monde moderne, l'explosion des outils dédiés à la communication, la multiplication des médias a entraîné un début de chute des frontières et des barrières. Même si la manipulation existe toujours (peut-être même plus encore qu'auparavant parce que la manipulation bénéficie des mêmes outils), Internet semble être le lieu où tout citoyen peut être en prise directe avec des événements et des idées surgis à l'autre bout de la planète. Le phénomène des « Blogs » succède à celui des sites personnels d'il y a quelques années, l'idée est la même mais l'information se structure.

Nous nous intéresserons aux sciences de la communication alternativement sous trois de ses facettes :

- Nous aurons besoin d'étudier la communication dans ses rapports avec le cerveau : perception, mémoire, traitement de l'image et du langage ;
- Nous devons aussi, de manière plus précise, nous concentrer sur l'analyse des problèmes de communication entre l'homme et les machines.
- Enfin, et c'est sans doute là l'aspect le plus lié à notre problématique, nous étudierons la communication entre individus et organisations, ainsi que l'impact des techniques de communication sur le fonctionnement de la société.

## **F. L'informatique linguistique**

L'informatique linguistique propose de nombreux outils pour assister les chercheurs dans l'analyse de grand corpus de manière plus ou moins automatisée. C'est bien entendu un aspect qui pourrait prendre une grande part dans le projet *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, même si nous n'avons pas toutefois vocation à développer de tels outils, ni d'en faire spécifiquement un sujet de recherche. Nous essaierons donc d'assortir notre dispositif d'outils issus de cette recherche et qui seront utilisés principalement pour construire des liens avec les autres thèmes.

## **G. La sémiotique**

Nous l'avons dit, le contexte général de notre recherche nous amène à réfléchir sur la publication en ligne sur le Web de corpus textuels numérisés, sous la forme commune d'un « site web » que nous appellerons plus simplement « bibliothèque numérique ». Il est bien entendu que le site web est à la bibliothèque numérique ce qu'un immeuble est à une bibliothèque traditionnelle, rien de plus qu'un espace pour héberger et abriter les ouvrages. Pourtant un architecte qui doit dessiner les plans d'une bibliothèque se posera des questions différentes de celui qui doit bâtir un stade ou un hôpital. Il devrait donc en être de même pour le webmaster qui doit concevoir une bibliothèque numérique. En tant que formes d'organisation et de gestion d'information, les questions qui se posent à nous concernant ces bibliothèques numériques peuvent être posées de la façon suivante :

*« L'évolution des modes de communication implique la nécessité d'une maîtrise permanente du sens produit par les différents supports utilisés et diffusés. Une connaissance approfondie et évolutive des mo-*

*tivations et attentes du public-cible est donc devenue particulièrement importante dans l'optique d'un usage raisonné de produits multimédias inscrits dans une politique communicative cohérente et réfléchie. La conception et le suivi d'utilisation de systèmes d'information multimédia (logiciels, CD-ROMs, sites Internet et autres supports) font appel depuis plusieurs années à l'analyse ergonomique afin d'optimiser notamment le rapport homme-machine instauré. Or, l'application de cette seule méthode n'est pas suffisante pour appréhender pleinement des rapports tels que celui qui unit le texte à l'image, les interactions qui gouvernent les variables visuelles ou encore les jeux qui s'installent entre les différentes isotopies. Il s'agira ici d'utiliser ce que la sémiologie, en complément des éléments vus ci-dessus, peut apporter à la conception et à l'analyse de systèmes d'information multimédia. (Masselot, 2004) »*

Pour préciser notre propos, nous pouvons ajouter suivant (Stockinger, 1999), que la sémiotique s'intéresse à ces formes d'organisation pour les trois raisons suivantes :

1. en tant que *producteur et fournisseur d'information* et de produits et services d'information ;
2. en tant que *témoin d'une culture d'information*, d'une histoire ou d'une tradition d'information ;
3. en tant qu'*utilisateur et consommateur d'informations* en vue du développement de (nouveaux) produits et services d'information et de son positionnement dans un environnement socio-culturel et économique.

Dans un premier temps, la sémiotique appliquée au cadre restreint des bibliothèques numériques va nous amener à nous pencher sur la description comparative de leur organisation suivant différentes approches :

- les scénarii et scènes composant un scénario organisant les espaces thématiques ;
- le scénario d'un site (ou d'une de ses parties) sous forme de régions propres à une page ;
- la structure narrative, les chemins ou parcours selon lesquels un site et ses prestations peuvent être explorés et exploités par un utilisateur ;

Cette étude devra permettre dans un second temps d'être mise en application sur le dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* de manière à assurer la mise en place d'une méthodologie pour la conception et la spécification du site, ainsi que la production de scénarios ou modèles.

## 2.3 Organisation de la thèse

Le compte rendu de notre travail de recherche s'organise suivant deux grandes directions, qui donneront lieu chacune à une partie. Dans la première partie nous étudierons l'hypertexte et à travers ce dispositif, l'évolution de l'écrit introduite par les supports numériques (page 53). Nous commencerons donc par analyser la mutation de l'écrit imprimé vers le numérique. Pour cela, nous reviendrons d'une part sur les dispositifs techniques qui mettent à disposition des lecteurs non seulement le contenu d'un livre, mais aussi jusqu'au contenu d'une bibliothèque. Cette constatation restant tout à fait quantitative, dans un deuxième temps nous aborderons pour ce qui est du qualitatif, la « qualité » de la lecture, médiatisée à travers de tels dispositifs. Nous étudierons d'autre part l'hypertexte, en dehors de tout contexte spécifique, puis dans le cadre de sa mise en œuvre sur la « toile », le Web. Cette modalité d'écriture et de lecture semblant proposer une manière d'articuler

un texte selon comment le lecteur et sa manière de penser l'attendent. Dans cette étude des hypertextes nous aborderons :

- le sens
- la lecture
- les parcours

que nous ne manquerons pas de mettre en rapport avec les expérimentations menées dans le cadre du projet *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*.

Dans la deuxième partie, nous reviendrons plus précisément sur le dispositif mis en place par le projet *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*. Ceci par rapport à deux axes d'étude bien différenciés :

- le dispositif dans son rapport avec les Logiciels Libres
- le dispositif présenté comme une forme originale d'édition de textes.

À ce compte rendu s'ajoute une troisième partie faisant l'analyse et la synthèse des différentes expérimentations, dans lequel nous proposerons un cadre de développement pour les bibliothèques numériques en particulier et plus généralement des modalités d'usages des NTIC dans le cadre d'actions de recherche.

# **Première partie**

## **Du livre au Web**

**L'hypertexte nouveau support de pensée ?**



L'ÉCRIT, LA MÉDIATION, LA NUMÉRISATION

---

« *Les droits imprescriptibles du lecteur*

1. *Le droit de ne pas lire.*
2. *Le droit de sauter des pages.*
3. *Le droit de ne pas finir un livre.*
4. *Le droit de relire.*
5. *Le droit de lire n'importe quoi.*
6. *Le droit au bovarisme*  
*(maladie textuellement transmissible).*
7. *Le droit de lire n'importe où.*
8. *Le droit de grappiller.*
9. *Le droit de lire à voix haute.*
10. *Le droit de nous taire. »*

Daniel Pennac

*Comme un roman*

## Sommaire

---

<b>3.1</b>	<b>Introduction</b>	<b>55</b>
<b>3.2</b>	<b>Le document</b>	<b>56</b>
A.	Définition	56
B.	Théorie du support numérique	58
C.	La chaîne documentaire	62
<b>3.3</b>	<b>La notion de collection</b>	<b>68</b>
A.	La définition du dictionnaire	69
B.	La collection pour valoriser	71
<b>3.4</b>	<b>Le livre, du papier à l'électronique</b>	<b>73</b>
A.	Le livre traditionnel	75
B.	Qu'entend-on par livre électronique?	81
C.	Le point sur les technologies	86
D.	Livre électronique : Rupture ou continuité?	101
<b>3.5</b>	<b>Conservation numérique</b>	<b>136</b>
A.	Les formats	136
B.	La numérisation comme changement de support	138
C.	La chaîne de numérisation	138
<b>3.6</b>	<b>Les corpus électroniques</b>	<b>144</b>
A.	Pourquoi un corpus en ligne?	145
B.	Les grands projets de numérisation de textes	149

---

## 3.1 Introduction

Avant d'arriver aux corpus électroniques en ligne, il nous faut présenter les différents médias, ou plutôt les différentes générations de médias qui ont participé à la construction de cette catégorie de corpus.

Nous partirons du document et en passant par la définition du livre, nous arriverons ensuite à une clarification de la notion de collection et de corpus. Du corpus nous pourrions appréhender ses différentes médiations électroniques, en particulier celles faisant appel à une interface hypertextuelle. Du corpus à l'hypertexte nous verrons que le corpus, médiatisé à travers une interface électronique sur Internet, s'organise et prend naturellement corps sous la forme d'un système hypertextuel.

L'hypertexte nous permettra de continuer par la construction d'un parcours raisonné dans le dernier chapitre ; pour l'instant ce sera simplement l'occasion de rappeler les différents problèmes inhérents à ce type de système et les manières de les aborder.

Ce chapitre est en quelque sorte une réflexion sur les « savoir faire » du livre numérique, que nous compléterons par l'étude des processus d'appropriation dans le dernier chapitre.

Par son mode de fonctionnement, l'hypertexte peut être perçu comme un élément structurant, qui va permettre de lier entre elles des idées, ou une source de désorientation, par la création d'un trop grand nombre de connexions entre un trop grand nombre de concepts. La désorientation qui affecte le lecteur perdu au milieu d'un trop grand nombre de chemins

possibles est plutôt un problème d'implémentation du système hypertextuel qu'un problème directement lié à l'hypertexte.

En revanche, dans le cas du corpus, la quantité d'information donne toute sa pertinence à l'utilisation d'un système hypertextuel, qui permet de mettre en place des parcours raisonnés et ainsi servir de guide, ou d'assistant à la lecture. L'hypertexte n'est plus seulement l'ouvrage, ni même la bibliothèque physique mais l'organisation bibliothèque c'est à dire les murs, les ouvrages les notices et les personnes qui conseillent habituellement les visiteurs.

Avant d'analyser parcours de lecture et hypertexte dans la suite, il nous faut situer le contexte du livre numérique. Pour cela nous allons revenir sur les notions de document, de corpus et de collection dans cette première partie. Cela nous permettra ensuite d'évoluer vers les notions de lecture, d'appropriation du texte et de ce idées quand elles sont portées par le média numérique.

## 3.2 Le document

### A. Définition

Une première définition du document pourrait être représentée par l'équation : *Document traditionnel = support + inscription*. C'est à quelques variantes près la plus courante, mais il manque dans cette description la finalité du document, c'est-à-dire l'information qu'il recèle, son contenu. Dans un premier temps en effet, l'accent est mis sur un support manipulable (au sens premier) sur lequel est fixé une trace interprétable selon sa forme par un œil, une oreille humaine, ou par le toucher pour la lecture en braille et, pourquoi pas demain, d'autres sens, avec ou sans prothèse. Cette trace représente le contenu, matérialisé par une inscription. Le support traditionnel est le papier,

la trace l'écriture, manuscrite ou imprimée. C'est la notion de document telle qu'elle est déclinée dans les réflexions du groupe « Roger T. Pédaque » sous la forme de trois axes : la forme, le medium et le signe.

*« La notion de document peut se décliner selon différentes facettes ou figures :*

- 1. Selon le statut physique ou matériel (papier, image, photo, vidéo, film),*
- 2. Selon le statut social ou communicatif : il s'agit de la place de ce document dans un domaine, un contexte, une société,*
- 3. Selon le statut cognitif et informatif : qu'est-ce que le document apporte comme connaissance, comme information et sur quoi ?*

*On peut dégager trois grands cas de figures :*

- Le document est une pièce unique, isolée : il est une preuve d'existence : c'est la carte d'identité, qui témoigne de l'inscription du porteur dans une société, dans un pays. Sa définition, c'est sa dénomination : un extrait d'acte de naissance, un passeport, etc. Il peut aussi avoir une valeur indicielle : c'est un témoignage, une preuve dans un dossier (criminel, judiciaire, fiscal). Dans ce dernier cas, c'est un élément d'un regroupement opportun en vue d'un but, d'une fin. Le document, dans cette situation générale est ancré comme pièce, il est auto suffisant, encore que certains supermarchés aujourd'hui réclament deux pièces d'identité pour accepter un chèque.*
- Le document est une partie de, un élément d'un dossier ou d'une collection : ce qui compte alors, c'est sa qualification, son indexation. Tel timbre est un document d'un tirage postal*

*historique, telle photo est un élément d'un reportage sur un voyage.*

- *Le document est un matériau d'une collection ; il est composable avec d'autres éléments. Il est porteur d'une interprétation qui lui confère une certaine détermination, c'est-à-dire un mode d'être dans un domaine ou un univers de connaissance et de culture. L'important alors, ce sont les critères qui vont lui être appliqués pour légitimer sa présence dans une classification. C'est sa valorisation.*

*(R.T.Pedauque, 2003) »*

Nous reviendrons un peu plus loin sur la notion de collection, qui est l'objet de la section 3.3, page 68. Dans un premier temps nous allons terminer notre étude du document par ses aspect techniques : théorie du support et chaîne documentaire.

## **B. Théorie du support numérique**

Le numérique a ouvert la voie à une évolution de l'utilisation des supports, la mutation essentielle étant le passage de l'inscription d'un signal analogique à un signal numérique, avec toutes les facilités de traitement informatique qui l'accompagnent. Ceci nous amène à faire trois constatations :

1. Le support papier autorise différentes instrumentations, dépendantes des usages. Le support numérique devrait permettre la même chose, mais nous verrons les restrictions apportées par la notion d'application par la suite.
2. Le support numérique permet de mimer les possibilités du papier, mais dans ce cas il n'apporte aucun bénéfice, parce que son implémentation

technique est plus lourde et il est de toute manière limité dans ses possibilités de reproduction du papier.

3. Ce sont les possibilités computationnelles qui marquent véritablement la différence, le document numérique est alors non pas un document inscrit mais un document calculé dynamiquement à la demande et en fonction d'un contexte particulier.

En revanche le support peut-être source de questionnement, nous pouvons en effet identifier plusieurs types de supports : logiques comme par exemple le fichier ou physiques : disque dur, disquettes et autres « supports » matériels. Ceux-ci sont des supports « permanents » au sens que le document a été placé, écrit par son auteur sur le support fichier ou disque, mais il existe aussi des supports « transitoires » qui sont utilisés par les systèmes informatiques dans le cadre de leur fonctionnement interne, par exemple quand un fichier est copié à travers un réseau, il est généralement découpé en fragments plus petits, ces fragments parcourent le réseau en ordre dispersé et sont rassemblés en un seul fichier, identique à celui de départ, sur la machine d'arrivée. Dans ce cas, nous pouvons considérer le réseau informatique comme un support. Il faut alors noter que le fichier est inaccessible pour un utilisateur humain et que son existence sur ce support ne dure que pendant le transfert.

Concernant ces dispositifs techniques, nous pouvons signaler sans plus insister le fort appareillage, domestique ou professionnel, indispensable au lecteur aujourd'hui pour accéder au document dans sa forme numérique et le parcourir.

Nous pouvons aussi ajouter un certain nombre d'interrogations, sur l'effacement et l'obsolescence du support.

Les mêmes « supports » d'ailleurs peuvent accueillir indifféremment n'importe quel type de représentation pourvue qu'elle soit numérique et il

sera même possible de « lire », étroitement imbriqués, du texte, de l'image fixe, de l'audio ou de l'image animée, le seul problème étant l'existence d'une application capable de lire un tel document. Nous reviendrons sur le rôle des applications dans les architectures informatiques telles qu'elles sont conçues actuellement, c'est bien souvent l'application qui prend le pas sur le document et nos systèmes sont très rigidelement centrés sur ce concept plutôt que sur les documents que nous désirons produire. De fait, la publication électronique a rendu possible la production à la demande d'un document (indifféremment sur écran ou sur papier). Le support a perdu sa position privilégiée au profit de l'affichage électronique. Une des évolutions importantes de la publication assistée par ordinateur a été l'avènement du *wysiwyg* (*What you see is what you get*, ce qui est en raccourci une formule qui veut dire « Ce que vous voyez est ce que vous obtiendrez », sous-entendu lorsque vous imprimerez le document). C'était un des principaux slogans commerciaux des premiers micro-ordinateurs grand public dotés d'une interface graphique (Apple Macintosh, 1984), qui ont rendu possible la visualisation à l'écran du document en préparation comme il sera finalement produit et imprimé sur le papier. *Wysiwyg* souligne le lien fort qui existe entre la production électronique et l'imprimé. Implicitement, ce que ce slogan exprime est que tout ce qui est produit sur un ordinateur est destiné à être imprimé, ce qui pouvait être encore vrai dans les années 90 mais qui est devenu très contestable aujourd'hui. En effet, à l'heure d'Internet, le fait que de nombreux documents numériques ne soient faits que pour être visualisés sur écran n'est plus discutable. Le *wysiwyg* a par ailleurs lui aussi évolué, par exemple dans le cadre du développement de sites Web, de nombreux logiciels proposent des outils de composition qui permettent de dessiner les pages d'un site telles qu'elles seront affichées sur l'écran de l'internaute. Ce qui semblerait donner

à l'écran le statut qu'avait auparavant le support papier (ou l'imprimante, comme passerelle entre le support électronique et le support papier).

Le *wysiwyg* est la parfaite illustration du lien qui unit document numérique et document papier : frères ennemis, unis par leur même destination de support de la pensée et des idées, mais intrinsèquement différents, voire opposés par leur mode de présentation et surtout leurs modalités d'usages. Le document numérique est plus fluide, mouvant. Il est présenté souvent comme en rupture avec le document traditionnel sur papier, mais n'est-il pas plus proche de notre mode de pensée, de l'oralité? Le *wysiwyg* s'inscrit lui aussi dans cette discussion sur la rupture et la continuité, c'est un avatar technologique qui permettait de rassurer à une époque où l'écrit électronique était systématiquement imprimé. Ce n'est plus le cas aujourd'hui, 20 ans après son invention ce mot est déjà en train de disparaître de notre vocabulaire et ceci pour plusieurs raisons. D'une part les documents sont maintenant la plupart du temps échangés de manière électronique plutôt qu'imprimés et postés, d'autre part la taille moyenne des disques durs de nos machines personnelles permet maintenant de stocker des centaines, voire des milliers d'équivalents d'ouvrages papiers qu'il serait très difficile de stocker, tout simplement pour des raisons physiques de place et d'encombrement.

L'inscription sur le papier assurait la stabilité du document, elle est maintenant remplacée par un signal électronique qui se déplace d'un support à un autre et rend le document éphémère et volatile comme pourrait l'être un discours oral. C'est donc plus en termes d'organisation du signal et de structure que les problèmes sont posés dans le cadre du document numérique, tout comme la rhétorique et l'argumentation ont été développés en leur temps quand les mêmes questionnements faisaient jour pour le discours. Mais cette fois l'orateur est remplacé par le lecteur, ce n'est plus l'émetteur

mais le récepteur qui a besoin de structure. Il s'agit à nouveau d'aider le lecteur à structurer du sens à travers son parcours de lecture plutôt que de déclencher la confusion par l'amoncellement d'informations, thème que nous développerons plus loin.

## C. La chaîne documentaire

### 1. Définition

La chaîne documentaire fait référence aux diverses opérations de traitement d'un document et elle a été conçue bien avant la création de documents numériques. Pourtant, c'est sans doute la preuve de sa profonde validité et de sa cohérence, elle s'applique toujours parfaitement au traitement de documents numériques. Il faut sans doute voir là le fait qu'à travers les outils, ce sont toujours les mêmes activités humaines qui prennent place. Les nouveaux outils viennent là pour faciliter certaines étapes, mais ils ne les remettent que rarement en question.

- sélection, acquisition, enregistrement, équipement. Avec un document numérique nous devons simplement ajouter numérisation, si le document primaire est papier.
- description, catalogage.
- analyse du contenu (résumé, indexation).
- mise en mémoire (catalogues), classement. Les notices sont déjà numériques depuis plusieurs années, c'est le même processus qui a lieu, mais cette fois notice et documents sont sur le même support.
- diffusion, communication qui sont facilitées par la mise en ligne, que ce soit à travers un réseau privé, à travers Internet ou un Extranet.
- contrôle (prêt de documents, évaluation des demandes et des réponses).

## 2. **L'application, notion imposée par le modèle technique**

Le « document » est une notion qui est devenue intuitive du fait de son omniprésence dans notre quotidien, et ceci est encore plus vrai depuis que le document est devenu numérique. C'est un terme fourre-tout qui désigne aussi bien un texte structuré comme un mémoire, qu'un tableau de chiffres ou, à la limite, qu'un morceau de musique. Avec le numérique les documents sont aussi devenus interactifs, ce qui les rend dépendants des interfaces techniques, des applications, qui permettent de les afficher et de les animer.

Dans le modèle technique utilisé pour mettre en œuvre les différents logiciels qui peuvent être mis en œuvre sur un ordinateur, c'est une approche par application qui a été retenue plutôt qu'une approche par contenu. Les raisons de ce choix sont essentiellement techniques et à bien y réfléchir c'est une approche qui pourrait paraître très anti-naturelle. Nous pouvons trouver là un début de piste parmi les questionnements autour de la difficulté de certains de s'adapter à une interface numérique. En effet, le support papier peut avoir des usages très différents et être destiné à des documents de types très différents. Même réunies dans un même cahier, comme c'est par exemple le cas avec un cahier de brouillon, chaque page peut être destinée à recevoir des informations très différentes de ses voisines : nous pouvons mélanger des notes, des textes, des adresses, des dessins, etc dans un même cahier. Bien entendu la spécialisation existe aussi avec le papier : nous notons généralement les adresses dans un répertoire et pour dessiner nous utiliserons un papier différent suivant qu'il s'agit de croquis au crayon ou d'aquarelles, mais en règle générale cela ne prend pas l'ampleur de la sur-spécialisation qui a cours avec le numérique, d'autant que qu'il devrait plutôt nous permettre de nous affranchir de toutes limitations et permettre la réutilisation d'un même contenu dans des contextes très divers.

C'est en fait une approche technique centrée sur l'application plutôt que sur le document qui conduit à cette situation. Nous partageons dans ce domaine la vision qu'a Ted Nelson de cet environnement où la technologie impose trop souvent une solution avant même que le problème n'ait été posé. De ce fait la solution proposée est généralement parcellaire et nécessite de nombreux efforts pour être utilisée conjointement avec d'autres. Prenons pour exemple une utilisation très simple, si ce n'est la plus simple qui soit d'un ordinateur personnel, c'est-à-dire l'écriture de courriers, électroniques ou traditionnels, papier. La finalité de cette activité est la réalisation d'un document comportant notre adresse si c'est un courrier plutôt officiel, une partie de message qui est l'objet du courrier et l'adresse à laquelle le courrier doit être expédié. Ces informations ont été codées dans des champs de notre logiciel de messagerie électronique et puisque nous avons déjà des adresses enregistrées dans la partie du logiciel destinée à la gestion de notre agenda, nous n'avons pas à taper l'adresse du destinataire, nous avons aussi sans doute enregistré une signature automatique quand nous envoyons un courrier électronique. Pourquoi alors devons nous entrer à nouveau toutes ces informations quand nous tapons un courrier qui sera mis sous enveloppe dans un logiciel de traitement de texte ? Pourquoi utiliser un logiciel différent pour taper un courrier papier d'un courrier électronique ? Et pourquoi faut-il encore changer de logiciel si nous voulons ajouter des croquis à notre texte ? Le document « courrier » est toujours le même, mais les applications à utiliser sont différentes et de ce fait les « documents », ou plutôt les fichiers produits sont différents.

Une explication pourrait être donnée par le fait que les informaticiens à l'origine de ces architectures ont sans doute confondu documents et fichiers, ce qui nous amène à aborder un autre phénomène que tout le monde accepte

sans le remettre en cause : la structure hiérarchique et arborescente des fichiers stockés dans un ordinateur.

Bien sûr cela reprend le même modèle que celui du monde physique : une feuille dans un dossier, un dossier dans un tiroir, un tiroir dans une armoire. Mais pourquoi se limiter au monde physique quand le numérique nous permettrait de stocker les informations sur un modèle plus riche et plus souple, un modèle plus proche du fonctionnement de notre cerveau. Nous verrons dans le chapitre dédié à l'étude des systèmes hypertextes que les chercheurs qui les ont mis en œuvre cherchaient à reproduire un modèle proche de notre mode de pensée associatif. Une notion dans notre esprit est liée à un grand nombre d'autres notions, qui sont chacune liées à un grand nombre d'autres notions et ainsi de suite. Ce modèle n'est pas hiérarchique, il n'y a pas de notion « racine » de laquelle découlent toutes les autres en arborescence. De la même manière, quand nous enregistrons sur le disque dur de notre ordinateur un article sur les documents numériques, il devrait pouvoir être retrouvé dans le dossier contenant les articles sur la notion de document, ainsi que dans le dossier contenant les articles portant sur la notion de numérique. Ce n'est pour l'instant pas possible sur les systèmes de fichiers actuels, sans l'ajout d'une couche applicative qui ajouterait cette souplesse, sans doute à travers l'utilisation d'une base de données. Dans ce cas se pose la question de l'indexation, c'est-à-dire qu'il faut à la fois définir comment nous voulons que les documents soient indexés, mais aussi comment nous voulons pouvoir y accéder, quel type de recherche nous voulons être capables de faire pour retrouver un document particulier. Le problème n'est en fait ni nouveau, ni particulier aux documents numériques puisque c'est déjà celui-ci que Paul Otlet cherchait à résoudre en construisant le Mundaneum dont nous reparlerons dans l'introduction à l'hypertexte du chapitre 4.

Le numérique bouscule profondément la notion de document. Avec le papier, le document n'était au départ que du texte, puis les enluminures sont venues s'ajouter, les gravures et ensuite les photographies. Avec l'arrivée de nouveaux supports, le document a pu s'animer en devenant cinématographique. Mais nous avons alors des documents différents utilisant des supports différents. Le numérique a apporté un support universel qui permet de contenir texte, images, sons et vidéo. C'est tout simplement ce qui correspond au vocable « multimédia ». Du support traditionnel au numérique les transformations se repèrent facilement. Le support numérique reste le même, quelque soit le contenu qu'il renferme, en revanche ce contenu n'est plus directement accessible et le document multimédia va imposer d'utiliser des dispositifs différents pour être décodé. L'aspect matériel, le traitement cognitif, la perception ou encore l'usage ont subi une mutation profonde.

Pour les entreprises de ce secteur, les activités de recherche et de développement sont intenses, activées par le sentiment de devoir être le premier pour pouvoir développer un nouveau marché, voire déposer un brevet qui marquera définitivement le territoire, avant que les concurrents ne maîtrisent la nouvelle technologie. Les découvertes sont vite traduites en outils professionnels et domestiques par des industriels intéressés et des pionniers enthousiastes sans être vraiment pensées socialement, et encore moins économiquement maîtrisées. C'est ce qui explique sans doute le nombre de nouveaux matériels proposés au public au début des années 2000, lorsque le développement d'Internet et les perspectives de gains en bourse avaient dopé tous les acteurs de l'économie numérique. Nous avons vu apparaître lecteurs de « e-books », de CD multimédia ou autres, la plupart ayant disparu quelques années plus tard. La pression technologique et commerciale rend ces matériels obsolètes rapidement, ce qui renforce encore le sentiment de doute

et d'insécurité parmi le grand public. Aucun dispositif n'a su se faire accepter, hormis ceux les plus triviaux tels le micro-ordinateur, capable d'interpréter tous les formats, et le lecteur de DVD « de salon » pour sa simplicité d'utilisation.

### 3. Approche dynamique du document

Dans ces conditions, le document réinventé par le médium électronique est devenu un concept plus éphémère, remis en question régulièrement par l'arrivée de nouveaux dispositifs, de nouvelles pratiques et la disparition des anciens.

*« Puisque le mouvement prime sur la réflexion, c'est lui qu'il faut observer, et tout particulièrement le mouvement initié par les chercheurs très actifs dans le domaine. Il peut être éclairant de regrouper les recherches en cours selon trois approches qui construisent progressivement un nouveau compromis entre une multiplicité d'acteurs pour, de fait, réinventer des documents au travers du numérique : le signe (ou la forme), le texte (ou le contenu) et le médium (ou la relation). Ainsi un document ne serait finalement, de notre point de vue, qu'un contrat entre des hommes dont les modalités anthropologiques (lisibilité-perception, signe), intellectuelle (intelligibilité-assimilation, texte) et sociales (sociabilité-intégration, médium) fonderaient une part de leur humanité, de leur capacité à vivre ensemble. Autrement dit nous voudrions mesurer comment le paradigme technique « numérique » transforme le contrat à trois facettes qui fonde la notion de document. (Salaiin, 2003) »*

Pour continuer dans cette approche, il nous faut aussi constater que la frontière entre document et base de données devient elle aussi plus floue. Quand un texte est structuré par exemple sous la forme d'un document

XML, il devient accessible à travers des requêtes comme le seraient des enregistrements d'une base de données. Quand il participe d'un hypertexte, il devient le nœud d'un réseau distribué et il peut perdre son statut de document fini, parce qu'il sera en relation avec d'autres nœuds en constructions. D'autre part le document affiché dans la fenêtre de l'explorateur Web peut très bien n'exister nulle part ailleurs que dans cette fenêtre, il aura été construit dynamiquement en fonction de requêtes et de résultats provenant de fichiers XML ou d'enregistrements de base de données.

*« Selon cette hypothèse dont la validité est sans doute variable selon la spécificité des contextes, un document n'aurait alors d'existence à proprement parler qu'à deux moments : éventuellement celui de sa conception par son auteur qui devra le visualiser ou l'entendre, pour s'assurer qu'il correspond à ses choix et surtout celui de sa re-construction par un lecteur. (Salain, 2003) »*

Toujours par rapport à cette construction dynamique nous pouvons même aller plus loin et avancer l'hypothèse d'un document numérique qui n'existe que quand un lecteur le réclame, et pendant la lecture. Il n'existe ni avant, ni après cet acte de lecture, puisque son contenu pourrait même avoir évolué quand un nouveau lecteur fait la même demande, pour laquelle il obtiendrait donc une réponse différente.

### **3.3 La notion de collection**

A suivre la définition de l'historien Krzysztof Pomian, le terme « collection » désignerait

*« Tout ensemble d'objets naturels ou artificiels maintenus temporairement hors du circuit d'activité économiques, soumis à une protection*

*spéciale dans un lieu clos aménagé à cet effet, et exposés au regard.*

*(Pomian, 1987) »*

On pourrait corriger cette définition d'historien en précisant qu'une collection n'est pas nécessairement un regroupement d'objets, mais plutôt un ensemble d'entités, d'éléments tels que l'acte de les regrouper implique un certain regard sur eux ou une certaine représentation construite sur eux. Ces entités sont alors considérées comme dotées de d'attributs qui les relient entre elles. Autrement dit, « ce qui fait collection », c'est aussi l'œil de l'observateur.

## **A. La définition du dictionnaire**

Concernant le mot *collection*, nous trouvons dans le dictionnaire *Le Trésor de la Langue Française Informatisé*<sup>12</sup> la définition suivante :

« 1. - Vx. [En parlant d'inanimés] Action de réunir, recueillir, rassembler. A. - [Ce qui est réuni est considéré comme un tout] Les collections d'eau stagnante (E. BRUMPT, *Précis de parasitologie*, 1910, p. 73). 1. [En parlant d'animés ou d'inanimés] Etat résultant d'une réunion d'éléments B. - [Ce qui est réuni est constitué d'éléments juxtaposés, conservant leur individualité] Ensemble d'éléments groupés en raison de certains points communs. 1. [La collection résulte de choix successifs faits systématiquement par un individu ou un groupe d'individus dans une intention particulière]

– Ensemble non fini (le plus souvent classé) d'objets réunis par un amateur, en raison de leur valeur scientifique, artistique, esthétique, documentaire, affective ou vénale. Collection d'insectes, de tableaux, de timbres ; collection particulière, publique ; pièce de

---

12. disponible sur le Web à l'adresse <http://atilf.inalf.fr/tlfv3.htm>

*collection ; vente de la collection (de) X ; faire (la) collection de ; dépareiller une collection :*

- Série de volumes contenant les œuvres d'un auteur ou de volumes publiés sous un titre commun, et édités le plus souvent de façon uniforme.*
- Ensemble complet de fascicules d'un périodique parus pendant une certaine période.*
- Ensemble des modèles (vêtements ou accessoires) créés par un couturier pour une saison. Collection de prêt-à-porter ; présentation de (la) collection ; la, les collection(s) de printemps-été, d'hiver : - Fam. Toute une collection de. Un grand nombre, une grande quantité de choses ou de personnes mises ensemble sans beaucoup de discernement. »*

À partir de cette définition, nous pouvons résumer la notion de collection autour de trois grandes « idées » :

- celle de réunir, de mettre ensemble des objets,*
- celle d'éléments juxtaposés ainsi réunis, mais qui conservent leur individualité,*
- celle d'un principe de réunion au nom d'éléments communs ou encore fondé sur des caractéristiques communes.*

La collection est ainsi un rassemblement d'éléments guidé par une volonté, impulsée par un individu ou un groupe. Cette idée directrice pourrait nous amener à un deuxième niveau de description. C'est souvent cette idée qui constitue le ciment de la collection, parce qu'elle résulte de la volonté affective, scientifique ou artistique qui a abouti à sa constitution. Dans tous les cas, elle est d'abord le résultat d'une volonté de pérennisation ou de valorisation d'un patrimoine.

## B. La collection pour valoriser

Dans le cas de notre projet *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, le travail de collection s'est effectué en deux étapes, la première a été la constitution d'une bibliothèque physique par André Deguine. Après sa disparition, le fonds de cette bibliothèque André Deguine a été versé aux archives départementales des Hauts de Seine. Si les archives étaient à même d'en assurer la pérennité, la valorisation était un objectif plus difficile à assumer. C'est ce qui a conduit à la seconde étape, qui a été le transfert de la bibliothèque papier sous format électronique pour en rendre les ouvrages disponibles à tous à travers le Web. Dans les deux cas, c'est la valorisation patrimoniale qui a guidé la structuration de la collection.

Cette courte description des étapes de la vie de la collection qui constitue le corpus du projet *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* nous conduit à réfléchir aux types d'opérations et de processus qui sont mobilisés à l'occasion de l'émergence d'une collection :

- le choix de quelques objets initiaux (livres, tableaux, bibelots, archives),
- la mise en évidence de propriétés communes ou proches (aux plans des formes ou des fonctions) entre ces objets,
- la confirmation d'une certaine permanence de ces propriétés : la collection doit se prévaloir d'une stabilité mesurable dans l'espace et dans le temps (physique, fonctionnelle, objectale : collectionner les compteurs à gaz peut se comprendre...),
- la légitimation de cette mise en commun au nom de valeurs qui peuvent être, selon les circonstances, symboliques (les fèves des galettes des rois), matérielles (les flacons de parfum), monétaires (les autos de luxe), patrimoniales (les objets de Picasso), historiques (les ivoires de Dieppe) voire sentimentales (mes « peluches »).

En conséquence, une collection traduira la mise en œuvre d'une ou plusieurs valeurs :

- *cognitives* : on l'a conçue et construite ; elle est donc porteuse d'idées, de représentations, de connaissances ;
- *pratiques* : elle est mesurable, on peut la référencer à un domaine, elle a des limites, des clôtures ;
- *fonctionnelles* : on peut la traiter, l'analyser, la décomposer, la recomposer, la vendre comme tout ou partie ;
- *symboliques* : historiques, originales (objets insolites), emblématiques (les objets retrouvés du Titanic).

De ces processus dépend la valorisation des collections. Cette valorisation peut suivre plusieurs types de stratégies concrètes et symboliques, centrées sur les objets-exemplaires de la collection ou centrées sur les extensions de ces objets :

- référentielles : définition et illustration d'un domaine sous forme d'ancrage dans ce domaine soit en extension (les timbres poste français de 1960 à nos jours) soit en rareté (les timbres des îles Comores de 1970 à 1980) ;
- testimoniales : la collection témoigne d'une époque ou d'un art (les épées du XVe siècle espagnol ; les œufs du sculpteur Fabergé) ;
- promotionnelles : il s'agit alors de la mise en valeur de créations (la mode) ou d'époques ou de lieux (les châteaux de la Loire) ;
- fiduciaires : fondées sur la rareté (telle ou telle collection de bijoux ou de voitures anciennes).

Le lien qui unit documents textuels et connaissances est symétrique. Dans un sens, on constate que les documents et collections de documents sont envisagés souvent comme de possibles sources de connaissances d'un domaine.

La notion de collection, quand elle touche aux documents, se présente généralement sous la forme d'une collection de textes organisés dans le but d'une analyse particulière, comme c'est le cas pour *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*. Dans ce cas nous désignerons la collection plus précisément comme un corpus, notion sur laquelle nous reviendrons pour traiter des corpus électroniques dans la section 3.6, page 144.

### **3.4 Le livre, du papier à l'électronique**

Comparer livre électronique et livre papier classique n'est pas chose évidente. Jusqu'à présent les textes électroniques sont la plupart du temps une simple retranscription au format électronique d'un texte imprimé, ce qui ne permet pas d'exploiter facilement toutes les opportunités que pourrait offrir un livre électronique. Au final, le livre électronique ne pourra s'imposer dans la panoplie des objets numériques qu'avec l'aide de plusieurs acteurs. Les constructeurs de matériel informatique, s'ils parviennent à inventer un support au moins aussi pratique et robuste qu'un ouvrage papier. Les éditeurs s'ils adaptent leur production à ce nouveau format et proposent de nouvelles présentations.

Tous les acteurs économiques peuvent être sollicités et il est difficile de prévoir comment se fera cette évolution. Parfois des leviers inattendus sont actionnés, comme par exemple l'introduction du « premier cahier intelligent<sup>13</sup> » par la société Clairefontaine<sup>14</sup>. Il s'agit d'un cahier papier, presque normal, il est simplement muni de capteurs particuliers qui interagissent avec un stylo numérique spécifique. Le stylo est muni d'une électronique qui lui permet d'analyser l'écriture, de la transmettre à un ordinateur portable

---

13. <http://www.clairefontaine-paperpc.com/>

14. <http://www.clairefontaine.com>

ou directement à un espace privatif sur le Web au travers d'un téléphone portable. Les premiers articles de presse sur ce produit sont apparus en août 2004 et le peu de recul ne nous permet pas de juger s'il a reçu un bon accueil ou non de la part des utilisateurs. Le dispositif reste éminemment technique, même s'il annonce une extrême facilité d'utilisation pour une puissance fonctionnelle sans équivalent : la possibilité de diffuser ses écrits manuels presque en temps réel à travers une interface numérique.

Mais il ne faudrait pas perdre de vue que le livre électronique, même s'il sera le *futur du livre*, n'est finalement qu'un outil dans une longue liste des systèmes d'écriture et plus largement encore des systèmes de communication.

*« La chronologie des progrès de l'écriture ne coïncide pas avec la chronologie historique générale. C'est une chronologie particulière, avec des déphasages, ce qui me semble mériter réflexion. Il faudrait en particulier se demander en quelles périodes l'usage courant de l'écrit s'impose pour débloquer une situation, pour imposer le démarrage d'une économie <sup>15</sup>. Voilà comment je vois le livre. Il est aujourd'hui pour moi une étape dans mon effort pour me dégager de l'histoire du livre imprimé avec ce qu'elle comporte de partiel, pour élargir les horizons. C'est ainsi que j'ai compris que le sort de l'écriture ne pouvait être envisagé que*

---

15. Le passé de Novgorod, ville du nord de la Russie est à ce titre exemplaire. Du X<sup>e</sup> au XV<sup>e</sup> siècle, la ville fut le centre d'une florissante république qui développa comme nulle part ailleurs le commerce et les arts. Une république avec son assemblée, ses représentants, ses élections... et ses écoles ! Alors que partout en Europe, l'écriture était le privilège du clergé, à Novgorod, nul n'était exclu de l'alphabétisation. Femmes, enfants, paysans : tous savaient lire et écrire. Outre les lettres, les archéologues ont également trouvé un dessin d'enfant qui accompagnait ses exercices d'écriture. Sa similitude avec les dessins des enfants d'aujourd'hui est saisissante. La ville de Novgorod doit sans doute son essor au fait que toute la population savait écrire, à moins que ce ne soit l'inverse. Grâce aux messages qu'ils pouvaient s'échanger, les seigneurs propriétaires, les *boyards* et les paysans qui travaillaient pour eux restaient en contact malgré la distance. Grâce à cela, les boyards pouvaient habiter la ville pour s'occuper de leurs affaires politiques ou commerciales tout en continuant à garder un œil sur leur exploitation.

*dans une étude générale des systèmes de communication et des rapports qu'entretiennent l'écriture et la parole. (Martin, 2004, page 227) »*

## **A. Le livre traditionnel**

Parce qu'il fait partie de notre paysage culturel depuis des siècles, le texte imprimé nous paraît commun et unique, nous y faisons référence sous l'unique vocable de « livre », mais en fait l'imprimé est exploité sous de très nombreuses formes : le livre est en quelque sorte une technologie, mais le livre peut être roman, encyclopédie, dictionnaire, guide ou manuel de référence. Le livre est souvent présenté de manière solitaire, mais il peut aussi faire partie d'un corpus plus large, ce peut être le tome d'une série plus large, voire un ouvrage dans une bibliothèque. Toujours uniquement identifié comme livre, les modalités de lecture peuvent en être très différentes, et nous pouvons encore complexifier ce rapide tour d'horizon de l'imprimé en ajoutant des supports plus éphémères comme le journal ou le magazine. Pour terminer d'essayer de montrer que l'imprimé peut faire appel à des formes vraiment très différentes et surtout très dynamiques, qui offrent énormément de souplesse dans la consultation et même dans l'enrichissement, qualités généralement attribuées aux documents numériques, nous pouvons aussi citer le dossier. Celui-ci sera composé de documents plus ou moins homogènes entre eux, qui vont pouvoir être ajoutés ou supprimés indépendamment les uns des autres et consultés de manière totalement aléatoire, sans que rien ne présage du « parcours » qu'un éventuel lecteur serait amené à y faire. L'encyclopédie ou le dictionnaire sont aussi des exemples assez traditionnels dans les recherches sur l'hypertexte et son rapport à l'imprimé pour montrer des tentatives de rendre le livre plus « interactif » comme nous avons l'habitude de le dire aujourd'hui.

Le mot français « Livre » nous vient du latin liber. Ce mot est défini par Littré comme une « réunion de plusieurs feuilles servant de support à un texte manuscrit ou imprimé ». Et dans son *Nouveau Dictionnaire universel* (édition de 1870), Maurice La Châtre le définit comme un

« Assemblage de plusieurs feuilles de papier, de vélin, de parchemin, imprimées ou écrites à la main cousues ensemble et formant un volume recouvert d'une feuille de papier, de carton, de parchemin, de basane, de veau, de maroquin, etc. »

Ce mot peut être aussi compris comme subdivision d'une œuvre littéraire, par exemple les « Fables » de La Fontaine se composent de douze livres. Ce sera alors un synonyme de partie.

Si nous nous en tenons à la description matérielle du livre, nous pouvons retenir les trois fonctions suivantes : support de l'écriture, diffusion et conservation d'un savoir, transportabilité.

Nous reviendrons sur le livre électronique dans la section suivante, mais nous pouvons d'ores et déjà constater que ces 3 fonctions qualifient identiquement livre papier et livre électronique sans qu'il soit possible de les différencier.

Il faudrait aussi noter que livre est devenu maintenant synonyme de livre imprimé, pourtant les premiers livres ont été des manuscrits et là aussi il y a eu évolution, peut-être même quelques révolutions. Ayant pour objectif premier de reproduire fidèlement les manuscrits, les premiers imprimeurs n'ont pas apporté de modification à leur présentation. L'étude critique des textes bibliques menée pendant la Renaissance semble être à l'origine de la segmentation moderne des textes.

Patrice Guerpillon nous résume les évolutions qui ont abouti à la création du livre imprimé tel que nous le connaissons aujourd'hui :

« Les premiers manuscrits ne comportent pas d'espace entre les mots ; la ponctuation y est exceptionnelle (*scriptio continua*). Ils ne montrent aucune marque de division sémantique. La segmentation du texte a longtemps obéi à une seule contrainte : la largeur des feuillets amenant le copiste à effectuer des retours à la ligne et à marquer des césures. Le texte est encore la transcription d'un récit oral dont la linéarité est artificiellement interrompue par les dimensions du support d'écriture.

*La stichométrie.* La segmentation des textes en *stichos* est observable sur certains codices. Le *stichos* correspondait à une ligne type égale à la longueur d'un alexandrin. La segmentation d'un texte en *stichos* aurait servi d'une part à calculer le salaire des copistes, et d'autre part à contrôler l'exactitude des copies réalisées. Il existait des relevés stichométriques indiquant le nombre de *stichoi* de chaque livre biblique. Ainsi, en comptant les *stichoi* d'une copie on pouvait déterminer si celle-ci comportait ou non des erreurs. *La colométrie.* Avec la colométrie apparaît une disposition du texte obéissant à un critère sémantique. Selon B.BOTTE, repris par FREY<sup>16</sup>, « la disposition colométrique – *per cola et commata* – est au contraire une disposition qui regroupe en courtes lignes les mots qui doivent être unis dans la lecture. Le codex de Bezae est un des plus anciens manuscrits en colométrie ». Le texte est alors divisé en strophes marquées par des alinéas dont l'initiale déborde dans la marge.

Certains manuscrits (*Alexandrinus* et *Vaticanus*) disposent le texte en regroupant plusieurs stophes à l'intérieur d'un chapitre (*capitula* en latin, *kephalaion* en grec). Les chapitres comportent des titres *titloi* ou *tituli* étant soit placés au début de chaque section, soit groupés en tête

---

16. Louis Frey. *Analyse ordinale des évangiles synoptiques*, Mouton/Gauthier-Villars, 1972

*du document pour en former le sommaire. Quand une numérotation apparaît, elle porte sur les césures plutôt que sur les chapitres.*

*L'établissement de correspondances entre les livres composant la Bible est à l'origine de différentes segmentations : sections ammoniennes, canons d'EUSEBE, système littéral de Hugues de SAINT-CHER, subdivisions saint chrétiennes.*

*Robert ESTIENNE est l'auteur de la segmentation moderne de la Bible en chapitres et versets numérotés. C'est avec l'édition de 1551 de son Nouveau Testament grec (le Textus Receptus) que s'imposera ce système. La numérotation est encore portée dans la marge du texte. La segmentation stéphaniennne, adoptée d'abord par les communautés protestantes, est approuvée par les catholiques en 1592. Elle est depuis lors la référence des ouvrages bibliques.*

*L'insertion de la numérotation dans le texte est due à Théodore de BEZE, en 1565. C'est à partir de cette date que le livre présente l'aspect logique et visuel que nous lui connaissons : utilisation de la typographie pour mettre en valeur une segmentation devenue sémantique, hiérarchique et inter subjective. Robert ESTIENNE et Théodore de BEZE ont inventé l'architecture moderne du texte, donné forme aux nouveaux lieux de mémoire. (Guerpillon, 2002) »*

Il est un autre aspect que nous n'aborderons pas immédiatement : le droit d'auteur et le droit du propriétaire de l'ouvrage. Kant y faisait déjà allusion et nous reviendrons sur cet aspect dans le chapitre 8 (page 367) de la partie III.

*« L'auteur et le propriétaire de l'exemplaire peuvent dire chacun avec le même droit du même livre : c'est mon livre ! mais en des sens différents. Le premier prend le livre en tant qu'écrit ou discours ; le second simplement en tant que l'instrument muet de la diffusion du discours*

*jusqu'à lui ou jusqu'au public, c'est-à-dire en tant qu'exemplaire. Mais ce droit de l'auteur n'est pas un droit sur la chose, à savoir l'exemplaire (car le propriétaire peut le brûler devant les yeux de l'auteur), mais un droit inné sur sa propre personne, à savoir celui d'empêcher qu'un autre le fasse discourir sans son consentement à l'adresse du public, consentement qui ne peut assurément pas être présumé, car il l'a déjà accordé exclusivement à un autre. (Kant, 1995 (édition originale 1796) »*

## 1. Poids économique

Une autre façon de peser l'importance du livre et de l'imprimé est de mesurer les données industrielles et son poids économique. Nous trouvons dans (Van Cuyck, 2003) une analyse du rapport « E-commerce and development report » de 2002 des nations unies et citant des sources de l'Unesco.

Parmi les chiffres de cette étude nous pouvons relever avec Alain Van Cuyck qu'au delà de la notion de fracture numérique, c'est d'une fracture de l'imprimé et donc de l'écrit dont il faut parler.

- Dans plus de 50 % des pays il y a moins de 10 titres de journaux, et seulement 8 % des pays en ont plus de 100.
- 50 % des pays produisent moins d'un livre par habitant, 30 % produisent entre un et trois livres par habitants et 20 % en produisent plus de quatre par habitants.
- 60 % des pays ont moins de 50 exemplaires de livres scolaires pour 1000 habitants alors que 20 % des pays en ont plus d'un pour chaque habitant
- 70 % des pays ont moins de 200 ouvrages disponibles en bibliothèque par million d'habitant, 16 % entre 200 et 500 et 15 % plus de 500.
- Dans à peu près 50 % des pays du monde entier, toutes librairies publiques confondues, la totalité des titres disponibles contient moins

d'un livre par habitant, 20 % disposent de un à trois livres par habitant et 30 % quatre livres ou plus.

Toujours selon les statistiques 2002 de l'UNESCO, La consommation moyenne par habitant dans le monde de papier journal est de 6,1 kg. Cette moyenne cache d'importantes disparités : en 1997, la consommation moyenne de papier impression et écritures était de 1,5 kg par habitant pour l'Afrique (1,3 en 1970), alors qu'elle était de 42 kg/hab pour l'Amérique du Nord en 97 (contre 22,1 en 1970).

La consommation mondiale en papier en ce qui concerne la presse quotidienne est relativement stable, elle a même plutôt tendance à régresser si l'on raisonne en terme de population touchée dans les pays du nord. Pourtant la production d'imprimés et de papier ne cesse d'augmenter, soutenue par le livre et les publications.

Si l'on se réfère à l'*International Standard Book Number* (ISBN) dont le siège est à Berlin, et qui regroupe 161 pays membres, l'augmentation des titres publiés de 1995 à 1999 est de plus de 50 % et il existerait près d'un demi-million de maisons d'édition répertoriées dans le monde.

Pour la France nous pouvons trouver sur le site du Centre National du Livre<sup>17</sup> un certain nombre d'informations qui permettent de confirmer cette tendance. Les chiffres disponibles sur la période 1999-2004 montrent une progression de 38.657 à 52.231 nouveautés (le tableau de la figure 3.1 présente une synthèse des chiffres disponibles sur le site pour cette période).

Toujours selon le même organisme, au cours de l'année 2000 il y a eu en France 44.618 nouveautés et nouvelles éditions, ce qui représente une augmentation de plus de 13 % par rapport à l'année précédente.

---

17. Ce site est visible à l'adresse <http://www.culture.gouv.fr/culture/sedocum/dll-cd.htm>. Les chiffres sont eux disponibles à l'adresse <http://www.culture.fr/culture/guides/dll/fiche10chiffres2000-2001.html> et <http://www.centrenationaldulivre.fr/doc/chiffres-cles.htm>

	1999	2000	2001	2002	2003	2004
Production				53.155	55.302	60.972
				-5%	+4%	+10%
Nouveautés	38.657		44.618	45.787	44.145	52.231
	2%		13%	+3%	-3,6%	+18,3%
Tirage Moyen	8.262	8.158		8.190	7.792	
	-1,7%	-1,3%		-1,9%		

TAB. 3.1: La production d'ouvrages papier

Les ventes en volume ont quant à elles atteint le chiffre record de 354 millions d'exemplaires. Toutefois le tirage moyen a tendance à stagner ou régresser, puisque le tirage moyen par livre tourne toujours autour de 8000 exemplaires durant toute la période étudiée.

## B. Qu'entend-on par livre électronique ?

Le livre électronique est déjà depuis quelques années un objet d'étude (Bruillard *et coll.*, 1998 ; Gimont, 1998 ; Vanderdorpe, 1999 ; Dupoirier, 1999). Le livre électronique n'est donc plus une nouveauté, mais il n'a pas encore trouvé une vraie place dans l'ensemble des objets culturels. C'est une expression qui met face à face deux termes, chacun porteur d'un sens qui semble s'opposer à l'autre : traditionnel contre nouvelle technologie, analogique contre digital. Il serait intéressant de s'interroger plus longuement sur le choix de ces termes, comme l'a fait Jean-Gabriel Ganascia dans le cadre d'un rapport de synthèse d'une réflexion menée sur le livre électronique par le GIS « Sciences de la Cognition » :

*« Comme point de départ, partons du terme « livre électronique » qui s'avère à la fois restrictif et inopportun, tout au moins au premier abord. Le terme est restrictif car le livre désigne un support particulier de l'écrit qui est advenu à un moment donné dans l'histoire ; il est*

*restrictif de parler de livre, là où tous les supports de l'écrit, du son et de l'image sont convoqués. (...) - Le terme est inopportun car la juxtaposition des deux mots, « livre » et « électronique » apparaît, à première vue, comme antithétique : le livre désigne d'abord le support physique de l'écrit (...) - à cette signification restrictive et antithétique du terme « livre électronique » s'ajoute un statut hypothétique : le « livre électronique » n'est pas encore réalisé, et quand bien même il le serait, son statut demeurerait encore largement indéterminé ; personne ne sait au juste quelle forme exacte il revêtira, non plus que la fonction sociale exacte qui lui sera assignée. (Ganascia, 2000) »*

C'est encore entre l'antagonisme et la dualité de ces termes que nous retrouvons cette autre définition provenant de la Commission sur le livre numérique du Ministère de la culture et de la communication :

*« L'expression « livre numérique » est à elle seule porteuse de deux concepts technologiques de nature bien différente : - le livre, désignant un objet physique bien identifié, en trois dimensions, résultat d'une chaîne d'opérations de fabrication, dont le contenu est définitivement figé, - le numérique, désignant un ensemble de technologies basées sur une transformation (la numérisation) d'un signal en nombres (suite de 0 et de 1), impliquant en général l'usage d'un ordinateur. (Cordier, 1999) »*

Pour essayer d'aborder de manière exhaustive ces nouveaux types de supports électroniques de l'écriture il est nécessaire de faire un distinguo entre « livre électronique » et « texte électronique », mais avant cela nous devons revenir sur le choix du terme « livre électronique ».

Un texte électronique est un fichier électronique contenant des caractères codés de manière numérique et formant un texte. Ce fichier nécessite un dispositif pour être consulté, pour l'instant c'est plus généralement un ordi-

nateur, muni d'un logiciel adéquat pour (i) traduire le codage numérique des caractères, (ii) traduire la présentation et (iii) permettre la navigation à travers le texte (feuilletage, recherche, ...). L'existence de ces textes électroniques, leur utilisation de plus en plus fréquente à travers les ordinateurs personnels, de bureau puis portables, voire « de poche » a poussé les fabricants de matériel électronique à développer des produits portables et spécifiques, ces dispositifs répondant à la dénomination de livre électronique.

Dans le chapitre 4 *Réflexions sur l'hypertexte* suivant, nous continuerons notre étude du texte numérique par son implémentation sous forme hypertextuelle. Il peut être déjà intéressant de noter que bien que le texte numérique autorise toutes les manipulations, et bien qu'il soit techniquement très facile d'implémenter des liens hypertextuels à travers un texte numérique, l'écrasante majorité des livres électroniques ne proposent pas de liens hypertextes. Mis à part les sommaires ou les index, et même pour ces éléments ce n'est pas une pratique systématique, le texte lui-même ne permet pas d'autre navigation que celle séquentielle du livre imprimé.

Les raisons de cela peuvent être multiples, le travail d'édition pour retravailler un ouvrage et lui ajouter des liens hypertextuels est complexe, et s'il est fait manuellement, cela peut être long et demander de nombreuses ressources. Le retour sur investissement reste problématique dans un marché qui a du mal à accepter le livre numérique. En contre-partie, pourquoi utiliser un livre numérique s'il n'apporte rien de plus que son homologue papier ?

### **1. Première définition du livre électronique**

Un livre électronique est un dispositif technique qui permet de restituer un texte sous format électronique. C'est effectivement ce qui pourrait se rapprocher le plus du livre traditionnel : autonome, portable, léger, il permet

uniquement la lecture plus quelques fonctions annexes autorisées par la miniaturisation des composants électroniques, sans que cela ne pénalise autonomie et taille. D'ailleurs, plus qu'un livre ce dispositif est plus souvent une bibliothèque, ou tout du moins une étagère de bibliothèque, puisqu'il permet de stocker plusieurs ouvrages.

Sous cette même terminologie nous allons pouvoir trouver plusieurs objets, réel ou « virtuels » tous différents mais tous ayant en commun d'être un support de communication écrite.

- une œuvre ou plus simplement un texte numérisé.
- le support électronique de ce texte numérisé, ce qui peut-être soit le support physique : CD, Carte électronique ou soit le support logique, virtuel c'est-à-dire le fichier contenant le texte.
- l'appareillage qui sert à présenter des textes numérisés. Ce sera le plus souvent un ordinateur portable dont la taille rappelle celle d'un ouvrage papier. Compte tenu des ressources limitées par la taille et de la spécialisation extrême de cette machine il est généralement uniquement possible de télécharger et lire des livres numériques. En fait, c'est plus une bibliothèque nomade qu'un simple livre numérisé. Il faudrait ajouter que l'interface de lecture est pour l'instant majoritairement un écran, ce qui veut tout et rien dire puisque cet écran peut avoir différents formats (tout comme le papier), utiliser des technologies différentes qui vont rendre l'affichage plus ou moins agréable, lisible (qualité du papier) et influencer la taille de l'appareil, par exemple les écrans révolutionnaires qui nous sont promis pour ce type de machine sont aussi simples qu'une feuille de papier dont l'encre s'écrit et s'efface de manière électronique. A la place d'un écran à lire avec les yeux, nous pouvons aussi imaginer une

interface braille, des haut-parleurs ou autres interfaces peut-être encore à inventer.

Dans un article sur le livre électronique paru dans les cahiers de médiologie, Georges Vignaux insiste sur la nécessité d'éviter deux discours souvent aussi peu informés l'un que l'autre, celui de *l'approbation lyrique* d'un côté, du scepticisme inquiet de l'autre, il articule sa réflexion sur la signification même du terme *livre électronique*, sur ses enjeux et les pratiques dont il est l'objet (Vignaux, 2000).

## 2. La définition du dictionnaire

Le Grand dictionnaire terminologique de l'Office de la langue française<sup>18</sup> donne « livre numérique » comme synonyme de « livre électronique », en français nous trouvons les termes : livre électronique, livrel et en anglais : e-book, eBook, electronic book. Le Grand dictionnaire terminologique de l'Office de la langue française, dictionnaire de référence définit ainsi le livre électronique : « Petit portable en forme de livre, muni d'un écran de visualisation, qui permet de stocker et de lire les publications en ligne disponibles par téléchargement dans Internet. » Le terme « livrel » a été formé sur le modèle de « courriel » (contraction des mots « LIVRe » et « électronique » selon la définition de l'OLF).

## 3. Notre définition

Nous pouvons résumer ce qui vient d'être dit en donnant pour définition du livre électronique : *Un ouvrage ou plus simplement un texte, stocké sur un support numérique et destiné à la lecture par un appareillage informatique.*

---

18. OLF. Grand dictionnaire terminologique <http://www.granddictionnaire.com>

## C. Le point sur les technologies, standards et fournisseurs

Le monde de l'édition électronique se partage entre fournisseurs de contenu et fabricants de matériel, des alliances se font quelquefois entre les deux, dans le but d'optimiser les produits pour faciliter leur utilisation, mais surtout dans le but de trouver un produit attractif et un modèle économique viable.

### 1. modèles nord-américains ou contenus anglophones

Une courte de liste des matériels qui ont existé, en particulier ceux qui ont été lancés au débuts des années 2000, profitant de l'essor de tout ce qui touchait à la « e-économie » et de ceux qui subsistent aujourd'hui, début 2005, montrera à quel point le livre électronique a du mal à se faire adopter du public.

- Gemstar eBook<sup>19</sup> Gemstar avait racheté en janvier 2000 deux spécialistes du e-book américains : le Rocket eBook de NuvoMedia et le Softbook Reader de SoftBook Press, à quoi s'est ajouté en 2001 l'un des pionniers de l'édition électronique en France, les éditions 00h00 SA<sup>20</sup>. (Lebert, mis en ligne le 26 octobre 2000) Cela a donné lieu à la commercialisation de deux modèles : le REB1100 (noir et blanc) qui succède au Rocket eBook et le REB1200 (couleur) qui succède au Softbook Reader. Ces deux modèles ont été construits et vendus sous le label RCA<sup>21</sup> (appartenant à Thomson Multimedia). Ces achats, partenariats et fusions n'ont pas suffi : en juillet 2003 Gemstar annonçait l'arrêt de la commercialisation des lecteurs. Le serveur Internet devrait rester opérationnel jusqu'en 2006 au moins, pour permettre aux utilisateurs des REB de stocker

---

19. <http://www.gemstarebook.com/>

20. <http://www.00h00.com/>

21. <http://www.rca.com/>

leurs ouvrages et un logiciel est mis à leur disposition pour en créer d'autres sur leur ordinateur personnel, qu'ils pourront ensuite lire sur leur REB. Comme entre temps les éditions 00h00 ont cessé elles aussi toutes activité commerciale, il semble que le seul moyen de pouvoir continuer à utiliser ces matériels sera de devenir son propre éditeur. Mais les difficultés restent grandes puisqu'il faut à la fois maîtriser un nouvel outil et trouver des ouvrages dont les droits d'auteur et la protection soient adaptés à cette manipulation.

– Everybook<sup>22</sup>

Il s'agissait là d'un livre électronique « double page », donc permettant de l'ouvrir comme un livre imprimé, ce qui semblait apporter un confort de lecture qui assurerait son succès. D'autre part ce système fonctionnait grâce au système d'exploitation GNU/Linux, adapté en l'occurrence pour ce type de dispositif en un système développé par la société Lineo. (Transfert.net, 2000) La société Everybook a disparu avant que le premier « EB » n'ait été commercialisé. La société Lineo continue d'exister à travers une alliance avec une autre, MetroWerks qui a fini par l'absorber.

– Audible.Com<sup>23</sup> Une alternative intéressante au texte électronique est la distribution d'une version sonore, ce qui permet d'écouter les livres électroniques. La technologie de compression permet de stocker une heure de voix dans un fichier d'environ 2 Mo. Pour l'instant, les livres ne sont qu'en anglais et leur prix varie de 10\$ à 20\$ suivant la durée d'enregistrement et très certainement la notoriété du lecteur (Robin Williams, Stephen King, etc...)... En 2005 Audible s'affiche même comme

---

22. L'ancienne adresse de cette société <http://www.everybook.net/>, conduit maintenant à une liste de liens *sponsorisés* menant vers des sites plus ou moins liés aux livres électroniques ou à la publication.

23. <http://www.audible.com>

- distributeur exclusif de livres sonores pour Amazon<sup>24</sup> et propose en plus du téléchargement de livres proprement dit des possibilités qui ressemblent plus à une autre forme de diffusion radiophonique : des reportages, des discours.
- Barnes & Nobles<sup>25</sup> avait ouvert une division e-book en février 1999, proposant à ses clients le Rocket eBook et des ouvrages pour celui-ci. En 2000 les formats *Microsoft Reader Software* et *Glassbook Reader Software* viennent s'ajouter au premier. Plus tard viendront s'ajouter les formats *Pocket PC* et *Adobe PDF* pour remplacer Rocket eBook et Glassbook défaillants. Cela n'empêchera toutefois pas Barnes & Nobles de fermer la division e-book en février 2004.
  - Rocket Editions<sup>26</sup> C'était le site original du Rocket eBook, racheté ensuite par Gemstar qui a finalement fermé sa division e-book en 2003.
  - Librius<sup>27</sup> était la société proposant le *Millénium eBook*, sorti en 1999 et qui a disparu en décembre 2003.

## 2. Modèles français ou contenus francophones

- BooKenStock<sup>28</sup> est un site créé en mars 2001. Il s'agit d'un portail Web dédié à la lecture, d'une part il permet de mettre à disposition de tous une littérature de qualité en langue française, accessible gratuitement à partir de n'importe quel ordinateur. D'autre part des idées de lecture issues directement des lecteurs sont proposées. Enfin, le site essaie de promouvoir l'utilisation des nouvelles technologies dans la découverte de la littérature pour amener au plaisir de la lecture (et au livre papier

---

24. <http://www.amazon.com>

25. <http://www.barnesandnoble.com/subjects/ebooks/ebooks.asp>

26. <http://www.powells.com/rocketeditions.html>

27. Ancienne adresse <http://www.librius.com/>, qui pointe désormais vers un revendeur de nom de domaines.

28. <http://www.bookenstock.com/>

dans un second temps!) les internautes qui ne sont pas spontanément « livrophiles ».

BookKenStock en quelques chiffres (janvier 2004) :

- plus de 14 000 membres dont 1200 auteurs
- 40 000 visites par mois
- 200 000 pages consultées par mois
- Cybook de la Société Cytale<sup>29</sup>

La société Cytale fondée en France en 1998 par Olivier Pujol a créé le premier modèle français du livre électronique. Une machine présentée comme étant en mesure de bouleverser les métiers de l'édition par Jean-Pierre Arbon, de 00h00.com, éditeur en ligne, racheté par la suite par la société Gemstar comme nous l'avons vu dans la section précédente. Cybook ressemblait à une ardoise électronique d'un format de 21 centimètres sur 16, pesant 900 g, muni d'un écran tactile et d'un stylet. Il disposait d'une mémoire pouvant contenir jusqu'à 30 livres de 500 pages, de la couleur, de six tailles de caractères différentes et d'un modem intégré.

- @folio<sup>30</sup> : Un projet de livre électronique ouvert

Le projet @folio est né à l'atelier Design de l'école d'architecture de Strasbourg en 1996. Depuis 2 ans, il est développé à l'Ensis (école Nationale Supérieure des Arts et des Industries de Strasbourg). Et contrairement aux livres électroniques qui peu à peu vont arriver sur le marché ce projet possède une particularité de taille : il est conçu comme un système complètement ouvert. Pierre Schweitzer, architecte designer à Strasbourg, est le concepteur de @folio (support de lecture nomade) et de Mot@mot (passerelle vers les bibliothèques numériques).

---

29. <http://www.cytale.com/>

30. <http://www.atfolio.net/>

- Keebook<sup>31</sup> : livre virtuel

Keebook n'est pas un livre électronique au sens strict : il ne se présente pas sous forme de « petit portable ». C'est un logiciel permettant de créer des livres virtuels multimédias commentés (à partir de documents variés comme des pages de sites Web, des fichiers de traitement de texte Word ou des images). Ces livres virtuels peuvent être envoyés sur Internet et lus sans que les applications d'origine soient nécessaires.

- LED *Livre Electronique Dynamique* de Numilog<sup>32</sup>

Livres créés par Music.text et diffusés en exclusivité en Europe par Numilog. Un LED est un texte auquel se superpose un contexte sonore (musique, onomatopées) et visuel (animation, dessins) qui vient soutenir et accentuer l'action décrite par l'auteur. Dans un LED, le/la jeune lecteur/trice peut cliquer sur les icônes qui parsèment le texte et découvrir des images en mouvement ou des dessins, ponctués de musique, dans le but de créer une atmosphère de surprise et d'incitation à la découverte. La lecture de ces livres requiert de posséder le lecteur Macromedia Flash. Ces livres comportent des animations sonores - et requièrent donc d'utiliser un ordinateur équipé de périphériques audio.

### 3. Cartables et manuels numériques

Le concept de cartable numérique ne recouvre pas une réalité unique. « Il s'agit en réalité d'utiliser une image familière, qui renvoie au lien entre l'école et la maison, pour suggérer l'intégration profonde des technologies de l'information et de la communication au sein et autour de l'école. ». Il fait partie d'une réflexion plus générale menée par le Ministère de l'éducation au niveau national visant à définir le schéma directeur des espaces numériques

---

31. <http://www.keeboo.com/>

32. <http://www.numilog.com/>

de travail. La Fing, Fondation Internet Nouvelle Génération<sup>33</sup> a publié à ce sujet un rapport en 2002. Plusieurs solutions sont en cours d'expérimentation dans des établissements scolaires ou des écoles françaises.

- Peu de temps après avoir été annoncé en grandes pompes, le cartable du futur de Havas Education (Bordas et Nathan)<sup>34</sup> fait déjà partie du passé. La branche éducation de Havas a présenté lors de l'université d'été à Hourtin en 2000, un prototype de cartable électronique construit en partenariat avec le ministère de l'éducation nationale par Fujitsu et les éditions Bordas et Nathan. La phase d'expérimentation du cartable numérique a été menée par les collégiens des académies d'Amiens (Somme), Strasbourg (Bas-Rhin), Boulogne-Billancourt (Hauts-de-Seine), Vivonne (Vienne) et Marseille (Bouches du Rhône) de la Toussaint 2000 à la fin de l'année scolaire 2001.

Il devait peser au départ seulement 1 kilo d'après le prototype présenté, mais le modèle final fait un peu plus de 2kg. Cela pourrait sembler n'être qu'un poids plume par rapport à la dizaine de kilos en moyenne du cartable des enfants en âge d'aller au collège, mais ce cartable électronique ne remplaçait pas complètement le cartable traditionnel. Il contenait uniquement les contenus des manuels « Sciences et vie de la terre » de Bordas, « Histoire et géographie » de Nathan et Le Petit Larousse Illustré (des ouvrages édités, évidemment, par Havas) qui étaient préchargés dans la machine, cela imposait aux élèves de transporter cartable classique plus numérique à la fois. Nous ne pouvons que relever l'apparente bonne volonté des ingénieurs, les deux manuels version papier pèsent 513g et 853g mais le dictionnaire papier pèse plus de 2 kilos, soit le poids du cartable électronique. D'après le cahier des

---

33. <http://www.fing.org/>

34. <http://www.havas.fr/>

charges, cela aurait pu passer pour une bonne idée mais en pratique, les élèves ne transportent pas Le Petit Larousse Illustré, qui reste chez eux et qu'ils utilisent plutôt dans leur travail personnel.

Le premier cartable numérique a été mis au point par Nathan, tandis que Bordas s'est chargé du second, les enjeux commerciaux sont immenses pour ces sociétés (à la taille du marché captif de l'éducation nationale). Par la suite Havas envisageait la mise à jour des programmes à distance, pour autant que l'infrastructure des collèges puisse pallier au manque de possibilités dans certaines familles et ainsi contribuer à réduire la fracture numérique. De cette façon, fini les livres jetés mais aussi les bourses au livres, puisqu'il n'est plus possible de revendre le livre de l'année écoulée. Malheureusement en 2003, Havas ne faisait déjà plus la promotion de cet outil. Nous voyons donc que sans standard ouverts et imposés aux différents fournisseurs, sans une volonté politique de comprendre et maîtriser ce marché, c'est finalement le « client final » qui risque de payer la facture, dispositif mal adapté, contraintes budgétaires mal maîtrisées et technologie instable. Les expériences se sont multipliées, dans plusieurs directions : cartable totalement « virtuel » c'est-à-dire en fait un portail Web, Pc portable, tablette électronique ou simple clef de stockage (type USB). Il ne semble pas qu'il y ait déjà une expérience réellement couronnée de succès dans ce domaine. L'accès haut débit est un des enjeux permettant de le résoudre, mais c'est plutôt l'accès libre et en tout lieu, qui éviterait de creuser le fossé entre ceux qui ont ADSL chez eux et les autres.

- Le i-m@nuel (Editronics Education)<sup>35</sup> Pas encore un cartable complet, mais une autre voie. Cet outil pédagogique réunit deux éléments com-

---

35. <http://www.editronics-edu.fr/>

plémentaires : un manuel papier et un site Internet. Le contenu a été rédigé en collaboration avec des enseignants, en conformité avec les programmes de l'éducation nationale.

- o Le manuel papier contient les connaissances de base à acquérir dans une matière donnée. Son volume est volontairement réduit : moins de 100 pages, soit le tiers environ d'un manuel classique.
- o Le site Internet [www.i-manuel.fr](http://www.i-manuel.fr) offre un environnement de travail complet et personnalisé : il reprend les textes du manuel papier et les enrichit de ressources pédagogiques. Ce site est accessible grâce à un mot de passe propre à chaque utilisateur, enseignant ou élève.

#### 4. Normes, standards et formats

Les livres électroniques ont connu une popularité avec des hauts et des bas, ce qui ne peut que contribuer à confirmer que leur succès n'est pas garanti en soi. Voici comment Hans Hansen, auteur et promoteur de livres rares en format électronique, et J. W. Olsen, auteur et observateur des technologies, voyaient l'avenir du livre électronique en 2000, en pleine période d'euphorie pour les technologies du numérique.

*« Hans Hansen pense que les éditeurs doivent respecter les trois critères principaux suivants pour s'assurer que les livres électroniques ne soient pas seulement un engouement passager.*

- *Premièrement, le contenu des livres électroniques devrait être préparé et révisé avec le même sérieux que s'il s'agissait d'un livre traditionnel. Les éditeurs de livres électroniques ne doivent pas sacrifier la préparation et l'édition à l'autel du téléchargement rapide.*

- Deuxièmement, les éditeurs de livres électroniques devraient intégrer à leurs publications des paramètres numériques (hyperliens, fonctions de recherche avancée, imagerie numérique), de façon à offrir un contenu complémentaire et à combler les besoins des lecteurs.
- Troisièmement, les livres électroniques doivent être agréables à l'oeil. Aussi les éditeurs devront-ils porter une attention particulière aux éléments typographiques : la longueur des lignes, l'interlignage, la coupure de mots, les alinéas et les en-têtes.

(Educause, 2000)<sup>36</sup> »

Dans cette liste, il n'a pas envisagé que les éditeurs puissent avoir quelque frilosité à diffuser des contenus numériques, sans avoir la possibilité de contrôler leur recopie à travers les canaux d'Internet. L'« affaire » Napster et le succès des réseaux de « peer-to-peer » sont passés par là et font voir les possibilités de développement grâce au numérique d'un tout autre regard aux diffuseurs de contenus.

D'ailleurs, pour sa part, Olsen est plus dans cette logique économique. Il pressent que les livres électroniques auront un avenir pour autant qu'ils permettront d'abaisser les coûts de production et de générer de nouvelles sources de revenu.

---

36. THE E-BOOK BOOM \* Despite growing popularity, the success of electronic books is by no means guaranteed. Hans Hansen, an author and developer of rare books in electronic format, and technology observer and author J. W. Olsen each present their views on the future of e-books. Hansen believes there are three key issues publishers need to address to ensure e-books are not simply a passing fad. First, the text of e-books should be prepared and reviewed as thoroughly as the text published in traditional paper formats—e-book publishers cannot sacrifice good editing and proofing for speedy upload. Second, e-book publishers should incorporate electronic features, such as hypertext links, advanced searching, and digital imagery, into e-publications in a manner that complements content and anticipates the needs of readers. Third, e-books must be pleasing to the eye, which means publishers should attend to typographical details such as line length, leading, hyphenation, indentation, and heading style. Olsen feels the future of e-books centers around their ability to generate cost savings and new revenue streams. (Publish, 3 May 2000)

Nous venons de voir les différents aspects du livre numérique et nous avons vu que cet objet est encore mal défini. Il est encore au cœur de plusieurs batailles industrielles et il tente toujours de trouver le standard qui lui apportera stabilité et surtout pérennité. Parmi les technologies qui nécessitent d'être mises au point, nous pouvons déjà identifier d'une part celles qui vont supporter le contenu et d'autre part celles qui vont permettre de diffuser ce contenu dans le respect des droits d'auteurs.

Les normes principales de codage d'information utilisées pour stocker et restituer un ouvrage sous forme numérique sont à la fois multiples si nous considérons toutes les possibilités de stocker du texte (depuis le texte brut jusqu'au HTML utilisé pour les pages Web, en passant par tous les formats de traitements de texte) et restreintes si nous ne considérons que celles qui permettent de coder un texte de manière à le restituer conformément à l'original papier. Voici donc cette liste restreinte :

**Le format PDF (*Portable Document Format*)** de la société Adobe<sup>37</sup> a une certaine longueur d'avance sur les autres. Le logiciel permettant la lecture de ce type de eBook est depuis longtemps disponible gratuitement et de nombreuses sociétés l'ont adopté, les industriels réunis sur le Forum Open ebook sont partagés entre l'intérêt de leur entreprise et l'intérêt général (ce qui n'est pas une nouveauté dans la rédaction de standards ouverts). Les poids lourds tentent de mettre en avant leurs propres solutions : la société Microsoft, omniprésente dans les applications de bureautique avec sa suite intégrée Microsoft Office, essaie d'introduire un nouveau format propriétaire. Adobe jusqu'à présent leader de ce type de technologie continue de le mettre en avant et travaille au développement d'un nouveau format baptisé SVG (*Scalable Vector Graphic*).

---

37. Leader de la production de logiciels de mise en page et de traitement d'images professionnels, visibles à l'adresse : <http://www.adobe.fr/products/main.html>

**Le format OEB (Open eBook)**<sup>38</sup> Ce consortium a défini un format standard pour livres électroniques, basé sur la norme HTML et la norme XML. Grâce à l'utilisation de ces formats ouverts, acheter et lire n'importe quel livre numérique sur n'importe quel support devrait être rendu possible. le Consortium Open eBook a défini un format de fichier standard pour l'e-book. On nous parle de supports électroniques nomades, de liberté d'utilisation mais pourra t-on demain compulser un même ouvrage sur tous les livres électroniques ? La réponse est oui. L'Open eBook (OEB), un format standard de lecture semble avoir été créé pour ça. Conçu par le National Institut and Standards Technology, il a vu le jour en septembre 1999.

**Le format SVG** SVG sont les initiales de *Scalable Vector Graphics*, c'est une spécification du W3C, ce qui veut dire que ce format est ouvert et accessible à n'importe quelle société qui voudrait l'utiliser pour l'implémenter dans ses propres produits. Il s'agit d'un langage de description de graphiques vectoriels bi-dimensionnels, basé sur XML. L'intérêt de ce type de graphiques est que pour décrire des images à base de formes géométriques, ils sont de taille assez modeste : plutôt que d'identifier tous les points d'une image, il suffit de décrire sont très peu gourmand en La société Adobe a décidé de l'utiliser pour enrichir la gamme d'outils construite autour du format PDF.

**Les formats à base de SGML ou XML** Nous pouvons ajouter à cette liste celle des DTD de l'information-documentation, que nous verrons page 89 dans la section D. *Les DTD de l'information-documentation* du chapitre 4 *Réflexions sur l'hypertexte*. Elles ont été développées pour le stockage de textes numériques, avant même que l'idée de livre numérique, tel que nous venons de le voir, n'ait pris réellement corps. Les raisons qui ont

poussé les fabricants à développer leurs propres formats sont sans doute commerciales, volonté de proposer le contenant et les contenus et d'en maîtriser la diffusion. En revanche, les formats ouverts qui ont été créés auraient pu l'être dans la continuité de ceux qui existaient déjà mais cela n'a apparemment pas été le cas.

Le deuxième aspect des nouvelles normes qui sont en train d'apparaître autour des livres électroniques concerne les technologies de sécurisation, dans ce cadre le choix est beaucoup plus restreint, mais la bataille commerciale Adobe-Microsoft est toujours présente, cette fois c'est EBX contre DRM.

- EBX (Electronic Book eXchange) Adobe invente le livre chronodégradable. Il s'agit d'un roman en ligne qui se désintègre au bout de trois jours, ou d'un ouvrage numérique qu'il est possible de prêter à un ami. Ceci est possible à travers la mise en place du système de protection des livres en ligne EBX (Electronic Book EXchange) mis au point par Adobe. L'intérêt de la technique réside dans la possibilité de prêter à un ami un ouvrage acheté.

*« La chronodégradabilité : un mot barbare certes, mais qui apporte certains espoirs chez les éditeurs pour qui la diffusion électronique d'ouvrages semblait impossible en raison du piratage...*

*(Jammamejian, 19 janvier 2001) »*

- DOI (Digital Object Identifier) Document de l'INIST présentant le DOI : INIST. DOI : système d'identification des objets digitaux.
- le DRM pour *Digital Rights Management* est un procédé de cryptographie particulier qui permet de rendre les contenus illisibles, sauf pour un lecteur ayant payé les droits de diffusion affectés à l'ouvrage. C'est un dispositif de sécurité très particulier, parce que habituellement ce type de système sert à protéger l'utilisateur qui est « bon ». Dans le cas de

DRM, l'utilisateur ayant accès au document par défaut, il est considéré comme « mauvais » et devra prouver sa bonne foi, par une clef valide, pour accéder au contenu. Le problème de cette technologie est donc de diffuser les clefs à ceux qui ont acheté un contenu, en les identifiant. Quand nous achetons un livre papier, c'est le libraire qui nous donne accès au livre, quand nous achetons un livre électronique, le libraire doit demander à un prestataire de service tiers de nous fournir une clef qui déverrouillera notre logiciel de lecture. Comment s'assurer de la fiabilité de ce prestataire, et comment assurer le respect de la vie privée des lecteurs dans ces conditions ? La question est posée mais ne dispose pas de vraie réponse.

Entre diffuseurs inquiets de perdre le contrôle de leur contenus copiés à l'infini à travers Internet et les lecteurs noyés dans une complexité technologique, qui semble de plus leur voler leur vie privée, le succès très mitigé des livres électroniques n'est peut-être pas une réelle surprise.

## 5. Papier et encre électroniques

L'avenir de l'écran pourrait se trouver dans le papier. Entamées il y a 20 ans dans un laboratoire de la firme Xerox à Palo Alto, les recherches sur l'encre électronique ont été relancées par le MIT, avec le rêve d'appliquer le texte et l'image à un autre support qui combinerait les avantages du papier et du numérique.

C'est donc parallèlement au développement de dispositifs « livres électroniques » ou « e-books » déjà traditionnels, que des laboratoires travaillent sur l'encre et le papier électroniques dans le but de rendre le livre électronique plus proche physiquement d'un livre papier. Ces dispositifs sont généralement présentés au public sous la dénomination de papier ou d'encre

électronique pour rappeler la similarité et la combinaison entre ancienne et nouvelles technologies d'écriture.

Le papier électronique est une des innovations qui pourrait sans doute donner plus de légitimité au livre électronique, en lui donnant plus de souplesse, en le libérant d'un support physique trop encombrant. L'objectif est d'être capable de présenter un livre électronique qui ressemble à un livre papier, avec des pages dont le contenu peut changer à loisir. Le livre devenant un dispositif électronique, de nouvelles fonctions vont pouvoir lui être données comme par exemple être capable de présenter plusieurs ouvrages différents, donc d'être une bibliothèque. Grâce à une « encre électronique » et une carte permettant de le connecter à un ordinateur ou à Internet, le livre pourra garder en mémoire des milliers d'ouvrages, en télécharger de nouveaux et donc d'être un livre qui contient virtuellement tous les livres.

Nous pouvons retenir quelques noms pour présenter la recherche et les projets industriels engagés dans cette voie. Les plus novateurs voudraient mettre au point un matériau à base de papier, mais ils sont encore loin de pouvoir présenter un produit au public, en revanche d'autres se sont *contentés* de perfectionner un affichage électronique classique, plus fin, plus léger et (relativement) souple. Ils proposent déjà ce type de dispositif pour des usages très variés, ce qui va depuis l'écran de téléphone portable jusqu'au panneaux d'affichage sur autoroute, le gain de poids étant le principal avantage mis en avant. Nous sommes alors dans des applications qui semblent très éloignées du livre électronique, mais elles ne sont que de nouvelles formes d'écriture électronique. Comme il y a des livres et des affiches papier, il y aura des livres et des affiches électroniques. Le support électronique, comme le support papier l'a fait avant lui, est capable de s'adapter aux différents usages que la société humaine voudrait lui voir adopter.

**Xerox PARC** Le projet de papier électronique réutilisable du *Xerox Palo Alto Research Center* a conduit à la mise au point d'un matériau, qui est plutôt en fait un dispositif d'affichage baptisé « Gyricon » et qui est visible sur la figure 3.1<sup>39</sup>. Les moyens de modifier l'image ou le texte affichés restent entre ceux du papier traditionnel, adaptés : imprimante électronique, crayon électronique et ceux de l'ordinateur, le Gyricon faisant alors office d'écran d'affichage.



FIG. 3.1: Prototype Xerox de papier électronique

**E ink** La société e-ink est propriétaire du nom « Eletronic Ink » qu'elle a utilisé pour baptiser le matériau qu'elle a mis au point. Le papier électronique est en fait une feuille qui ressemble en tout point au papier ordinaire, mais recouverte à sa surface de millions de microcapsules dont l'aspect est contrôlable électroniquement. Le procédé d'affichage utilise une encre polarisée à micro-particules. La société E-Ink travaille aussi à la commercialisation des premières affiches-écrans.

**Media Lab au MIT** Le projet de livre utilisant l'encre électronique du Media Lab du MIT (Massachusetts Institute of Technology) est mené dans le cadre du laboratoire MicroMedia<sup>40</sup>. Les efforts dans ce projet visent

---

39. Image provenant du site du projet Gyricon à l'adresse : <http://www2.parc.com/hsl/projects/gyricon/>

40. Une page sur le site du Media Lab MIT à l'adresse : <http://www.media.mit.edu/micromedia/elecpaper.html> décrit sommairement ce projet.

à mettre au point un affichage électronique sur un nouveau matériau principalement constitué de papier. C'est dans ce laboratoire que Joseph Jacobson, physicien, a inventé un nouveau procédé d'« encre électronique », qui permet d'imprimer et d'effacer des livres en une fraction de seconde. Avec son équipe, ils préparent un livre d'apparence classique mais dont on pourra modifier à loisir le contenu et qui conservera en mémoire tous les textes déjà lus et annotés. Chacun pourrait ainsi constituer sa bibliothèque portable et son carnet de notes perpétuel.

**IBM Electronic Newspaper** La compagnie IBM travaille aussi sur la mise au point d'un prototype de journal électronique qui ressemble à un livre traditionnel, mais dont les 16 pages sont d'un matériau flexible à base de papier et de fibre de verre. Les pages sont légères, cerclées de métal et couvertes d'une fine feuille de protection transparente. Le dispositif peut être roulé ou plié comme un journal traditionnel.

## **D. Livre électronique : Rupture ou continuité ?**

Avec l'arrivée d'une nouvelle technologie, apte à remplacer une ancienne, se pose souvent la question de la manière dont la transition va se faire. Est-ce que nous allons passer du livre papier au livre électronique avec de profonds bouleversement dans notre rapport à l'écrit, ou cela va-t-il se faire graduellement, progressivement, de nouveaux usages émergeant lentement avec l'adoption des nouveaux outils. En France, une commission du ministère de la culture, a étudié l'impact des nouvelles technologies sur la chaîne du livre. Ses conclusions ont été publiées en mai 1999 et sous la forme d'un document appelé rapport Cordier (Cordier, 1999).

Cette même question s'était déjà posée avec l'arrivée de l'imprimerie, comme nous le rapportent Febvre et Martin dans *L'apparition du livre* :

« Vers 1500-1510, en effet, l'imprimerie a gagné la partie. Dans les bibliothèques, les livres imprimés relèguent de plus en plus au second rang les manuscrits ; vers 1550, ceux-ci ne sont guère consultés que par des érudits. Pareil changement ne s'explique que par l'énorme activité des presses, qui multiplient les textes imprimés à un rythme toujours plus rapide : 30.000 à 35.000 éditions différentes antérieures à 1500 parvenues jusqu'à nous, et représentant 15 à 20 millions d'exemplaires [...] Mais bien plus encore au XVI<sup>e</sup> siècle : [...] à Paris plus de 25.000 éditions publiées au cours du XVI<sup>e</sup> siècle ; à Lyon, 13.000 peut-être ; en Allemagne, quelque 45.000 ; à Venise 15.000 [...] Tout cela nous permet de penser que l'on pourrait dresser une liste de 150 à 200.000 éditions différentes parues entre 1500 et 1600. Si l'on adopte, pour fixer les idées, le chiffre de 1.000 comme tirage moyen, 150 à 200 millions d'exemplaires ont donc été imprimés au XVI<sup>e</sup> siècle. (Febvre et Martin, 1958, page 396) »

Un demi-siècle plus tard, Henri-Jean Martin est toujours prudent quant au futur du livre électronique, en revanche il le situe dans une perspective historique qui tend à montrer que le véritable livre électronique reste à inventer, et que cela ne se fera qu'après une nécessaire période d'appropriation, dans laquelle nous sommes sans doute plongés :

« Ces tentatives me semblent trop récentes pour que je puisse avoir une opinion sur le futur livre électronique. Sans doute celui-ci évoluera-t-il pour trouver une nouvelle forme de présentation. Je me bornerai à rappeler à cet égard que le livre imprimé s'est d'abord efforcé d'apparaître

*comme le fac-similé du manuscrit<sup>41</sup> avant de trouver très lentement une mise en texte qui lui était mieux appropriée. (Martin, 2004, page 284) »*

Le futur du livre électronique devrait-il être nécessairement en rupture avec le livre papier ? Dans notre imaginaire collectif, sans doute parce que nous avons besoin de repères qui marquent, il y a l'invention de Gutenberg qui marque le passage du manuscrit au livre imprimé. Cette découverte marquerait-elle une rupture entre le livre imprimé et le livre manuscrit ? Nous verrons un peu plus loin qu'il y a eu en fait une longue période de transition et d'adaptation, à la fois de la technique, des éditeurs, des auteurs et des lecteurs. Les copistes n'ont pas tous abandonné leur travail, l'imprimerie elle-même a eu une longue évolution, l'art de la calligraphie s'est transformé en typographie. Pour penser le livre électronique il faudrait être capable de faire abstraction de tout ce que le livre véhicule : le plaisir de lire associé au plaisir de caresser le papier, la manipulation de l'ouvrage pour « sentir » son poids, la lecture des « métadonnées » que sont la page de titre, la quatrième de couverture. L'appareillage qui entoure le livre papier nous est maintenant devenu transparent mais il n'en est pas moins présent. Une fois ces signes qui viennent compléter notre compréhension du livre imprimé laissés de côté, il nous faudrait encore pouvoir aussi mettre de côté tous les signes habituellement véhiculés par l'ordinateur, alors seulement nous serions capables de penser un nouveau livre numérique.

C'est cette même question qui est posée en ces termes par Roger Chartier :

*« Une des difficultés pour penser ce phénomène est que l'imagination du futur reste toujours dépendante de ce que nous connaissons ; ce qui fait que, pour nous, la culture du texte électronique est forcément un monde*

---

41. Il est intéressant de mettre en parallèle les outils mis à disposition du lecteur sur *C*o*L*i*S*ci*ences*, sur lequel nous reviendrons page 106 dans la section 6.1. En effet nous proposons plusieurs « vues » du texte en parallèle, du traditionnel au novateur, c'est-à-dire du fac-similé au parcours notionnel.

*d'écrans. C'est l'ordinateur tel que nous le connaissons...[...] Mais sait-on ce que deviendront les supports matériels de la communication des textes électroniques ? (Chartier, 1997) »*

Nous pouvons par exemple constater la multiplication des dispositifs portables audio, qui ont été conçus avec pour objectif de permettre d'écouter, et donc aussi de vendre, de la musique téléchargée depuis Internet. Ces dispositifs ont vu leur capacité de stockage s'accroître et les plus puissants permettent maintenant de conserver plus d'un millier d'albums, c'est-à-dire des milliers d'heures d'écoute. Parallèlement, il s'est développé un marché du « livre audio », c'est-à-dire des ouvrages lus, souvent par une personnalité ou un comédien qui vient encore ajouter un nouvel attrait à l'ouvrage. Ceci correspond à une double attente des « audio lecteurs » : la simplicité de mise en œuvre et d'accès au texte : il suffit d'introduire une cassette ou un CD dans son auto-radio ou son baladeur pour écouter le (ra)conteur et le texte. Maintenant, ce type d'ouvrage est aussi disponible en téléchargement, il est donc possible d'emporter dans un dispositif pas plus grand qu'un livre de poche à la fois ses chansons mais aussi ses textes préférés, pour les écouter pendant le trajet en métro domicile-travail par exemple. Cette nouvelle pratique, *écouter* des ouvrages électroniques plutôt que lire, n'est sans doute par encore très répandue mais elle répond à plusieurs attentes : un dispositif simple à utiliser, qui permet plusieurs usages, dont le prix est amené à baisser du fait des gains de productivité et de capacité. Ce serait d'une certaine manière un retour à une société orale, ce qui soulève de nombreuses questions. Ce n'est de toute façon qu'un essai pour trouver une nouvelle réponse au futur du livre électronique et il ne faut pas rejoindre le camp des alarmistes qui prédisent la fin de l'écrit trop rapidement puisqu'en même

temps l'écrit continue de nous inonder, en particulier grâce à ces mêmes outils numériques.

### **1. Livre ou e-book, texte imprimé ou électronique**

Comme nous l'avons déjà abordé en introduction, le texte sous format électronique présente surtout la particularité d'être soit une transcription fidèle de l'imprimé, soit en rupture totale avec celui-ci. c'est-à-dire que nous avons d'un côté les textes électroniques destinés à être imprimés et qui seront finalement sans doute lus principalement sous format papier, et à l'opposé nous avons des ouvrages électroniques qui font un large usage de l'hypertexte, du son et de la vidéo, qui sont en totale rupture avec l'image traditionnelle du livre tome unique d'un roman qu'il faut lire de la première à la dernière page. Nous avons donc d'une part des textes électroniques qui respectent de manière très précise le format de l'imprimé. La suite de logiciels Adobe est exemplaire à ce titre puisqu'elle affiche comme principale qualité le fait d'être très fidèle, voire identique, au document papier. C'est là une des représentations de la formule *wysiwyg* dont nous avons déjà évoqué le caractère très partial d'utilisation du numérique comme simple brouillon de l'imprimé. Les documents électroniques qui sont en rupture avec l'ouvrage imprimé sont le CD-Rom, parce qu'il peut ajouter au texte vidéo et son, et bien entendu le World Wide Web, parce qu'en plus de l'interface hypertextuelle, il permet la communication, l'échange de données et il n'est pas figé mais au contraire en constante évolution.

## 2. Le futur du livre ?

Il y a encore peu de temps, de nombreuses questions émergeaient du fait du développement du livre numérique, par exemple en 2000, Jean Clément posait les questions :

*« Le e-book » est-il encore un livre ? L'expression « livre numérique » a-t-elle un sens ? Le livre traditionnel a-t-il encore un avenir ? (Clément, 2000, page 14) »*

Même si toutes ces questions ne sont pas entièrement résolues, nous pouvons déjà constater que le livre électronique n'est plus une nouveauté qui pose question mais un média de communication qui s'ajoute à ceux déjà existants. La première question est sans doute la plus importante : le livre électronique est-il un livre ? Suivant le sens que l'on fait porter à ce mot, la réponse peut être oui ou non. Par exemple, pour un bibliophile le livre électronique n'est pas un livre puisqu'il n'est pas cet objet de collection qu'il affectionne. En revanche, pour une personne qui voyage et qui est amatrice de romans, le e-book pourra être le moyen de lire en tous lieux : avion, train, hôtel, sans s'encombrer et dans ce cas le *dispositif électronique de lecture* sera considéré par cette personne comme un livre. Nous pouvons aussi déjà constater les premières conséquences de l'existence de ce nouvel objet sur le texte, plus que sur le livre lui-même. C'est donc encore un fois plus sur le fond que la forme que se situe l'enjeu.

*« Du livre objet, on est passé au livre-bibliothèque, au livre interactif, au livre au réseau, au livre multimédia. La dématérialisation du texte, sa dissémination sur les supports les plus variés et la convergence des différents médias laissent deviner que si le livre a un passé, le texte,*

*quant à lui, a un avenir dont le e-book n'est qu'une des figures possibles.*

*(Clément, 2000, page 19) »*

Un illustration de cette évolution pourrait être l'introduction du cartable électronique, qui pourrait être décrit comme une bibliothèque (portable) dans lequel les ouvrages numériques de l'élève sont stockés. Dans ce cas le livre électronique n'est même plus expressément cité et fait partie intégrante, *naturellement* du procédé ou système de bibliothèque numérique.

**Du Livre à la bibliothèque : le livre rechargeable** Contrairement au livre traditionnel, le livre électronique permet, par exemple, de naviguer, de faire des recherches, d'annoter ou de mettre en relief le texte. Cette interface de lecture numérique permet d'utiliser et de transporter avec soi une véritable petite bibliothèque de documents en ligne (livres, magazines, journaux, etc.). (OLF)

**Un livre interactif, multimédia ?** Interactif est synonyme de conversationnel dans le domaine de l'informatique. L'interactivité implique que l'utilisateur de la machine intervienne dans la sélection et le déroulement des messages. Le Grand dictionnaire terminologique de l'Office de la langue française donne un certain nombre de termes apparentés :

- Livre interactif ou livre multimédia
- termes apparentés de livre électronique : livre interactif, livre multimédia
- termes apparentés d'e-book : interactive book, multimedia book.

#### 1. Interactif

*« Se dit des matériels, des programmes ou des conditions d'exploitation permettant une interaction, en temps réel, entre l'utilisa-*

*teur et l'ordinateur. Note : l'interaction se traduit en un échange réciproque de questions et réponses entre l'utilisateur et l'ordinateur. »*

## 2. Multimédia

*« Technologie de l'information permettant l'utilisation simultanée de plusieurs types de données numériques (textuelles, visuelles et sonores) à l'intérieur d'une même application ou d'un même support, et ceci, en y intégrant l'interactivité apportée par l'informatique. Note : Le terme « multimédia » est utilisé comme nom ou comme adjectif et il prend un « s » au pluriel. »*

## 3. Hypermédia

*« Extension de l'hypertexte (présentation de l'information qui permet une lecture non linéaire grâce à la présence de liens sémantiques activables dans les documents) à des données multimédias, permettant d'inclure des liens entre des éléments textuels, visuels et sonores. Note : le terme « hypermédia » s'utilise aussi comme adjectif (ex. : des documents hypermédias, des liens hypermédias). »*

NB : ce dossier met l'accent sur le livre électronique (support de lecture) mais traite aussi plus largement du livre numérique (support de l'œuvre), multimédia et interactif.

## 3. Rupture

Dans quelle mesure le livre électronique tranche-t-il radicalement sur le livre et sur les autres supports de l'écrit, tant sur ceux qui ont existé avant le livre, que sur ceux qui ont coexisté avec lui, emblèmes, enseignes, insignes, affiches, tableaux, peintures d'églises, prédelles et retables, telle est la première des questions. Tous s'accordent pour voir là une évolution

capitale, mais des discordances se font entendre dès que l'on cherche à qualifier précisément cette évolution et à lui trouver des précédents. Certains y voient une révolution analogue, par sa portée, au bouleversement qui a accompagné la naissance de l'écrit ; d'autres, moins radicaux, considèrent que cette révolution s'apparenterait à celle qu'a opérée l'imprimerie : l'édition ne serait plus enfermée dans le cadre imposé à la page matérielle, support indispensable du livre imprimé. Des jeux de pointeurs commandant l'ouverture de multiples fenêtres sur les multiples facettes des textes, ajouteraient des dimensions supplémentaires transverses à la bi-dimensionnalité des pages. Certains se placent sur le plan des supports : tout cela aurait essentiellement à voir avec l'introduction du papyrus et l'abandon des tablettes de cire ou d'argile. Il y aurait là, avec l'électronique, une matière nouvelle, ductile à l'infini, qui se substituerait, du fait de sa souplesse, aux matières antérieures ; pour ce qui touche à la communication, l'histoire serait toujours allée dans le sens d'une flexibilité et d'une densité croissante, passant des cailloux peints et des ossements gravés, aux tablettes d'argiles, puis au papyrus, au parchemin, et enfin au livre imprimé. Le livre électronique, par la flexibilité résultante des propriétés computationnelles des ordinateurs qui le mettent en scène, continue à aller dans le sens de l'histoire.

D'autres enfin n'y voient qu'un changement quantitatif, un changement d'importance, certes, un changement d'échelle, mais un changement dans la continuité, un simple changement de vitesse d'accès et de taille... La grande coupure dont nous serions les héritiers se situerait au V<sup>e</sup> siècle avec le passage du rouleau au livre, c'est-à-dire avec la transition qui a permis d'aller au delà d'une lecture linéaire et exhaustive pour accéder à une lecture sélective et judicieuse, ce que le seul défilement du rouleau rendait impossible.

Selon les analogies évoquées, les perspectives et les préoccupations changent. D'un côté, pour ceux qui s'émerveillent du bouleversement au point d'y voir une véritable révolution, il convient, dès à présent, de réfléchir aux modes de pensée qui naîtront de la familiarité avec ces nouveaux supports. La passage du texte à l'hypertexte, c'est le passage d'une lecture suivie et imposée à un accès multiple, pris en charge par un appareillage, donc sans demander d'effort particulier au lecteur et surtout quasiment instantané. Ceci pourrait transformer l'espace mental de lecteurs qui, rompant avec la tradition intellectuelle classique, ne soumettraient plus leur esprit à des enchaînements linéaires de raisons. Il y aurait, derrière cette profusion de parcours de lecture, naissance d'une nouvelle forme de pensée, fille du *zapping* et de l'image, que certains vont jusqu'à qualifier d'hypertextuelle. Dans un ordre d'idées proche, certains se prennent à rêver de perspectives stupéfiantes : la séparation apparemment irréductible du texte et de ses lecteurs pourrait être abolie dans un futur proche, de sorte que seraient désormais possibles des voyages organisés entre les signes, devenus tangibles et palpables au sein de la page matérialisée grâce aux progrès des réalités augmentées, ou, plus prosaïquement, des incursions dans l'univers du récit rendu immédiatement perceptible par les techniques de création des mondes virtuels... Sans aller jusqu'à adhérer pleinement à cet enthousiasme visionnaire, plusieurs ont souligné qu'une porosité tendait à s'introduire dans la frontière auteur-lecteur grâce aux ressources de l'interactivité : le parcours original de chacun dans le livre devenu combinatoire s'apparenterait à une forme d'écriture. à cela s'ajouterait la possibilité de personnaliser le livre en introduisant des annotations ou en modifiant les parcours offerts selon le profil ou l'attitude des utilisateurs ; c'est ce que l'on appelle les documents actifs. Il faut aussi noter certaines des possibilités de l'écrit électronique liées aux possibilités de calcul

des ordinateurs qui les affichent et qui autorisent des traitements auparavant impossibles. Par exemple la transformation du texte en un message sonore permet aux personnes atteintes de cécité de pouvoir *lire* un document électronique au moyen d'un ordinateur. La traduction automatique n'a pas encore fait complètement ses preuves, mais là encore les possibilités de traitement automatique laissent envisager des applications futures intéressantes.

*« Luther s'est plaint en son temps que ses contemporains ne savaient plus lire. Qu'est-ce que cela voulait dire ? Tout simplement que ses contemporains avaient perdu l'habitude de lire des textes théologiques comme la Somme de saint Thomas, en s'accrochant phrase après phrase pour découvrir une construction absolument monolithique et directrice. Ils s'étaient d'une certaine manière libérés de l'emprise de l'écrit, du fait même qu'ils avaient de multiples écrits, et des écrits plus accessibles, sur leurs tables. (Martin, 2004, page 282) »*

Il se peut toutefois que l'organisation même du support papier soit modifiée par les habitus informatiques : le format dit « à l'italienne », auquel correspond la présentation cathodique usuelle, pourrait peut-être graduellement s'imposer... Á moins que les écrans ne continuent à grandir et à tendre vers le format du cinémascope. Parce que les ordinateurs personnels se transforment en centre de diffusion multimédia, pour être plus clair, parce qu'il est possible de visionner un film enregistré sur DVD sur l'écran de son ordinateur personnel, les constructeurs de matériel informatique proposent maintenant des écrans dont le rapport de la longueur sur la largeur rappelle le format du cinémascope. La convergence des nécessités entre le cinéma et le livre numérique pourrait paraître étrange, mais elle ne l'est pas véritablement pour qui veut se donner la peine de réfléchir puisqu'il s'agit simplement de proposer dans les deux cas un écran qui englobe autant que possible le champ

de vision d'un humain. Le cinémascope plonge le spectateur dans l'action et le livre dévoile son contenu au lecteur sans qu'il n'ait à faire de mouvements de tête. Le confort apporté pour la vision d'un film a apporté un nouveau confort pour la lecture électronique, sous ce format il est possible d'afficher deux pages cotes à cote, comme avec un livre ouvert. Alors que quelques années en arrière certains fabricants proposaient de basculer l'écran de 90° pour retrouver le format d'une page de papier, c'est le passage par la vidéo qui nous a permis de constater que nous lisons des livres dont les pages se font face. Même si nous lisons une ligne à la fois sur une page à la fois, la sensation de se retrouver en territoire connu quand les deux pages d'un ouvrage sont à nouveau affichées ensembles sur un écran apporte un confort indéniable. Au demeurant, à notre connaissance, aucune étude de psychologie sociale qui viendrait confirmer cette hypothèse n'a, à ce jour, été exécutée. à côté de ceux qui explorent les mutations vertigineuses qui se préparent, d'autres préfèrent souligner la continuité et mettre en relief la chaîne qui relie le nouveau à l'ancien. Il ne faudrait pas que les projecteurs braqués sur les techniques actuelles nous éblouissent au point de nous faire oublier que les pratiques anciennes, en particulier les stratégies d'écriture, sont, elles aussi, pour la plupart, non linéaires.

Les jeux d'indexation propres à l'hypermédia sont généralement totalement calqués sur les modes d'indexation inaugurés, à la fin du XII<sup>e</sup> siècle, par l'introduction de tables des matières et par la division en chapitres, puis perfectionnés par l'ordre alphabétique dans les encyclopédies. Ils reprennent aussi les modes d'indexation offerts par les bibliothèques et par leurs catalogues raisonnés, ce qui abolit les frontières entre ouvrages, faisant des hypertextes, de grandes bibliothèques.

Nous avons vu dans la section **A. Le livre traditionnel** (page 75) de ce même chapitre comment se sont construites les caractéristiques du livre, ce dernier repose sur des principes d'organisation nés il y a quinze siècles avec le passage du rouleau au livre. Ce sont ces mêmes principes qui se retrouvent dans le livre électronique, pourtant il y a bien plus dans l'hypertexte que dans le livre classique du fait des accès multiples, de la présence d'images, d'images animées et de son, et de la possibilité d'introduire des annotations...

*« La mutation numérique est donc à la fois une continuité et une rupture : c'est une continuité dans la mesure où elle ne fait qu'objectiver ce qui a toujours été déjà là. Un lecteur est toujours actif, un parcours de lecture est toujours dynamique et articulé au contenu, ce dernier est toujours organisé. Cependant, en objectivant ces structures en les abstrayant des documents et en leur donnant une contrepartie matérielle et manipulable, le numérique marque une rupture dans la manière de concevoir et d'élaborer le contenu. (Crozat et Bachimont, 2004) »*

Nous pouvons ajouter à cela ce que nous savons de ce qui s'est passé avec l'apparition de l'imprimerie : les lecteurs étaient habitués à une présentation, s'attendaient forcément à la retrouver dans les ouvrages imprimés ; certains copistes se sont transformés en imprimeurs, d'autres ont maintenu leur activité avec plus ou moins de succès ; le volume des ouvrages disponibles a soudainement augmenté dans des proportions difficiles à imaginer avec des manuscrits.

*« Tout d'abord, une constatation préalable : les premiers incunables ont exactement le même aspect que les manuscrits. En cette période de début, les imprimeurs, bien loin d'innover, poussent à l'extrême le souci de l'imitation[...] Que les premiers imprimeurs se soient efforcés de copier exactement, et parfois de reproduire servilement, les manuscrits qu'ils*

*avaient sous les yeux, il n'y a là, en vérité, rien qui doive nous étonner,[...] Comment les premiers typographes auraient-ils pu concevoir, pour les livres imprimés, un aspect différent de celui des manuscrits qui leur servaient de modèle ? Bien plus, l'identité de l'imprimé et du manuscrit ne devait-elle pas être à leurs yeux la preuve d'un triomphe technique en même temps qu'un gage de succès commercial ? Constatons donc que l'apparition de l'imprimerie n'entraîne pas de révolution soudaine dans la présentation du livre : elle marque seulement le début d'une évolution [...] avant qu'on arrivât, au milieu du XVI<sup>e</sup> siècle, à lui donner, à quelques détails près, la présentation qui est encore la sienne de nos jours. (Febvre et Martin, 1958, pages 106-108) »*

Le livre électronique et les questionnements qu'il suscite ne sont pas nouveaux comme en témoigne le mémoire de Delphine Rosenthal (Rosenthal, 1993) de 1993. En revanche, les technologies, les normes et même les acteurs ont énormément évolué et les questionnements d'hier peuvent ne plus avoir de rapport avec ceux d'aujourd'hui. Plusieurs des technologies mentionnées dans ce mémoire sont maintenant obsolètes : Data Diskman, CD-I, Vidéodisque, ce qui montre à quel point le problème de la pérennité peut être crucial. En effet, dans un domaine où l'innovation est la règle, de nombreux produits sont lancés sur le marché pour essayer de capter une clientèle avide de nouveauté. Le domaine bénéficie des innovations technologiques qui font que ce qui était totalement impossible ou hors de prix peut devenir la norme quelques mois plus tard. Les produits deviennent non seulement démodés, mais totalement obsolètes et trop souvent restent incompatibles entre eux. Avec les supports traditionnels comme le papier le problème de la pérennité d'un ouvrage est centré sur la conservation du support lui-même, il n'existe pas d'appareillage intermédiaire entre le message inscrit et le lecteur.

Les manuscrits de la mer Morte, compte tenu des conditions climatiques particulières de la région dans laquelle ils se trouvaient et du fait qu'ils ont pu rester hors d'atteinte pendant plusieurs siècles, ont pu bénéficier de relativement bonnes conditions de conservation. Ce témoignage a traversé les siècles, certaines parties sont perdues, d'autres sont très abîmées par le temps mais il reste suffisamment de matériaux pour permettre des interprétations. Quand ce n'est pas le papier qui a été utilisé mais un support bien plus pérenne comme la pierre, les écrits ont traversé les siècles et le message qu'ils renferment nous arrive presque intact (hiéroglyphes, pierre de Rosette, stèles gravées, ...). Bien que l'écriture utilisée ne nous soit pas toujours intelligible, encore un fois, le message est toujours présent. Le problème n'est *que* de traduire la langue ancienne dans laquelle ils avaient été écrits.

Avec le livre électronique les choses se compliquent. La détérioration reste bien évidemment possible, le vieillissement est lui bien réel, s'agissant de produits industriels correspondant à des normes de qualité précises, on sait aussi que la durée de vie de ces supports est limitée, mais le problème le plus grave est la disparition de la technologie de lecture. En effet, le support n'est pas lisible directement par un humain, il doit passer par une phase de traduction des codes binaires qui composent le document en caractères lisibles par un opérateur humain. Si la technologie matérielle disparaît, le support est illisible, si la technologie logicielle disparaît, le problème est quasiment le même, avec peut-être toutefois des solutions de traduction, mais qui resteraient sans doute difficiles à mettre en œuvre par des non-informaticiens.

En quelques décennies les disquettes sont passé par les formats 8", 5"1/4 puis 3"1/2, elles sont maintenant en train de disparaître au profit des « clefs USB ». Entretemps nous avons vu apparaître puis disparaître différents

formats de « super diskettes » de « très grande capacité », du type « Jazz ». Il y a eu aussi des améliorations du stockage sur bandes magnétiques, grâce aux cassettes ou autres cartouches. Les CD sont devenus inscriptibles (il est possible de les graver une fois), et même certains ré-inscriptibles, c'est-à-dire qu'ils sont utilisables avec presque la même facilité qu'une disquette. C'est maintenant le tour des DVD. Ce court paragraphe met en évidence une grande quantité de supports techniques, qui ont une fâcheuse tendance à l'obsolescence et qui demandent presque autant d'interfaces matérielles et logicielles pour pouvoir les lire et les écrire. Cela a été longtemps un frein à l'introduction dans les bibliothèques, espaces de conservation pour lesquels cette complexité technique est rédhibitoire.

« Les libraires réfractaires avancent aussi que le livre électronique n'a pas sa place dans une librairie, à côté des livres imprimés, mais plutôt dans le rayon des jeux vidéo ! (Rosenthal, 1993, page 84) »

Depuis l'époque de ce mémoire, la maturation et la convergence des technologies ont commencé à faire évoluer les bibliothèques. Les travaux sur DUBLIN-CORE, TEI, ... sont un exemple de cette évolution.

Avec l'apparition de ce nouveau substrat, beaucoup de questionnements se font jour pour déterminer son influence sur le sens de l'écrit qu'il contient. Nous pouvons légitimement nous poser la question de savoir si ces questionnements sont pertinents. Comme l'a écrit Jean Duvignaud dans *Le Monde*, nous pensons qu'il est surtout important d'utiliser ce nouveau substrat pour identifier *a posteriori* de nouveaux usages plutôt que de rechercher *a priori* son impact sur l'écrit.

« Si Montaigne ou Rabelais s'étaient attardés à réfléchir sur l'imprimerie, ils n'auraient jamais rien écrit ! Les techniques ne sont pas des objets de méditation. (Duvignaud, 18 janvier 1994) »

La manière habituelle de formuler ce type de question est de mettre en opposition électronique/numérique et création/pensée humaines, comme par exemple cette citation relevée toujours dans le même mémoire :

*« La technique est multiplicatrice d'anti-culture plus facilement car l'orientation « productiviste » banalisante et routinière de l'électronique la rend plus apte aux applications qu'aux créations individuelles essentiellement consommatrices de procédés manuels ou de pensée naturelle ». (Rosenthal, 1993, page 125) »*

Dans le cas de *C<sup>o</sup>L<sup>i</sup>S<sup>ci</sup>ences* la numérisation des ouvrages ne participe pas d'une orientation productiviste, ni banalisante et encore moins routinière. Notre projet est une création qui a demandé de mettre en œuvre de multiples procédés manuels : de la numérisation à la programmation du système hypertextuel, en passant par le balisage notionnel, c'est la main de l'homme et la pensée naturelle qui ont permis de faire avancer les choses. La définition du mot *livre* dans l'encyclopédie Wikipédia<sup>42</sup> donne un éclairage, qui à nouveau présente l'arrivée du numérique comme une rupture, mais qui en fait la situe au même niveau que celle qui s'est déjà produite avec l'imprimerie :

*« Aucune innovation majeure n'intervient dans la production du livre entre la fin du XV<sup>e</sup> siècle et la fin du XX<sup>e</sup> siècle. En revanche, de nouveaux types de documents apparaissent au XIX<sup>e</sup> siècle : photographie, enregistrements sonores et cinéma. La rupture se produit dans les années 1990. La généralisation du codage numérique multimédia, qui code sous une forme unique et simple (0 ou 1) des textes, images fixes, images*

---

42. Wikipédia est une encyclopédie internationale, dont la particularité est de pouvoir être enrichie par n'importe lequel de ses lecteurs. C'est en fait un projet communautaire inspiré des Logiciels Libres, sur lesquels nous revenons au chapitre 8. Wikipédia est disponible gratuitement et librement en ligne à l'adresse <http://fr.wikipedia.org/wiki/Accueil>. Nous reviendrons plus précisément sur ce projet dans le cadre de l'écriture hypertextuelle au cours du chapitre 4, en particulier avec ce qui touche à l'écriture hypertextuelle collaborative, section C., page 76.

*animées et sons est probablement une invention du même ordre que l'écriture. L'hypertexte améliore encore l'accès direct à l'information, comme l'a fait le codex en son temps. Enfin, l'Internet fait baisser les coûts de production et de diffusion, comme l'imprimerie à la fin du Moyen-Age.*

*Il est difficile de prédire l'avenir du livre. Une part importante de l'information de référence, destinée à un accès direct et non à une lecture séquentielle, comme par exemple les encyclopédies, existe de moins en moins sous forme de livre et de plus en plus sous forme de site Web. En revanche, les livres électroniques, ou e-books, n'ont, pour l'instant pas connu, un grand succès. On peut penser que la forme codex a encore un long avenir pour tout ce qui nécessite une lecture séquentielle, ou qui est autant un bel objet qu'un support d'information : les romans, les essais, les bandes dessinées ou les livres d'art. »*

#### **4. Les métadonnées**

Nous reviendrons sur les métadonnées dans le cadre de la section 4.4 (page 84) traitant de la Gestion Électronique de Documents (GED) plus en détail. Elles interviennent assez directement pour deux raisons dans le dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* : d'une part parce qu'à titre de bibliothèque numérique, *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* nécessite qu'elles soient mises en œuvre comme toutes les techniques habituelles, et d'autre part elles sont indispensables dans la construction des parcours savants.

Depuis toujours, les bibliothécaires ont rédigé des notices, ou des index, pour décrire les documents disponibles. Pour reprendre le vocabulaire utilisé en informatique, ces notices constituent des données, qui servent à décrire d'autres données (le contenu des livres) : on parle alors de métadonnées. Pour rester dans le cadre de cette section, parmi les spécificités fortes du

livre numérique nous pouvons citer, outre l'interactivité, la présence des métadonnées. Elles ne sont pas une invention du numérique, puisque les bibliothécaires n'ont pas attendu d'avoir des ordinateurs pour constituer des notices ; en revanche avec le numérique les notices et autres informations accompagnent directement l'ouvrage sous forme de métadonnées. C'est pour cette raison que nous pensons que les métadonnées sont sans doute là où la différence est la plus marquée avec le livre conventionnel. Par métadonnées nous entendons toutes les informations qui s'appliquent au livre, mais qui ne font pas partie de son contenu stricto-sensu comme par exemple les notices que nous avons déjà mentionnées, les mot-clefs, etc. Avec l'utilisation d'un ordinateur les métadonnées peuvent aller plus loin et renseigner sur les étapes de construction du document, une partie d'entre elles ne sont plus directement contrôlées par l'auteur du document<sup>43</sup>, elles sont automatiquement ajoutées par l'éditeur de texte en fonction du contexte : nom de l'auteur, son organisation, le nombre d'éditions et le temps passé à éditer le document, cela peut aller jusqu'aux modifications apportées au document avec l'historique des ajouts et des ratures.

Dans le cas qui nous intéresse ici, les métadonnées sont surtout un moyen d'automatiser un certain nombre de traitements. En rendant accessible à une interface programmatique des informations telles que le nom de l'auteur, le titre de l'ouvrage, le nom de l'éditeur, la date d'édition et tout ce qui permet de décrire le contexte dans lequel a été écrit l'ouvrage, nous pouvons utiliser un agent intelligent pour rassembler les ouvrages ayant certaines caractéris-

---

43. Les informations accumulées de cette manière peuvent être interprétées de manière anodine dans le contexte personnel, donner lieu à des interprétations sur la genèse de l'œuvre dans le contexte artistique, en revanche dans le contexte des affaires elles permettent quelquefois de mettre en évidence un changement radical dans une communication commerciale, un repentir dans un contrat qui peuvent entraîner un problème de relation client-fournisseur. D'où la question, à laquelle nous ne répondrons pas, de savoir s'il est pertinent que ces informations soient enregistrées à l'insu de l'auteur.

tiques. L'agent n'a pas à être codé pour reconnaître tel ou tel ouvrage, il doit simplement savoir interpréter les métadonnées pour reconnaître celles qui correspondent à la requête sur laquelle il a été lancé.

## **5. Lecture sur papier et lecture sur écran**

Une problématique importante qui motive notre travail de recherche est donc celle de la lecture et de la navigation dans une double perspective :

- Il s'agit d'abord d'interroger l'acte de lire dans son nouveau rapport au texte électronique. Cette question est en fait très imprécise parce qu'elle ne pose aucune contrainte sur ce qu'est un texte électronique. Nous venons de voir à travers tous les dispositifs qui ont déjà été développés et ceux qui sont en préparation que le texte électronique n'est pas un. Suivant que le dispositif le supportant sera portable ou pas, qu'il permettra uniquement de lire ou aussi de visionner des films, d'écouter de la musique, voire de gérer ses finances personnelles, le rapport du lecteur au dispositif sera différent et le texte sera plus ou moins accessible. Avant de pouvoir se poser la question du changement des modalités de la lecture avec le support électronique, il faudra que nous ayons catégorisé ces supports et alors nous verrons peut-être que pour certains le changement sera radical et pour d'autre ce ne sera qu'une adaptation. Cette étape franchie, nous pourrons ensuite réfléchir aux nouvelles spécificités introduites dans l'acte de lire, par la médiation électronique, par les possibilités collaboratives et aussi par l'introduction de l'hypertexte.
- Nous pourrons ensuite réfléchir à comment spécifier ces nouvelles conditions de l'offre de lecture, ne serait-ce que pour rentabiliser la puissance des dispositifs ainsi mis en jeu. Le mot « rentabiliser »

pouvant être pris dans son double sens puisqu'il s'agit d'une part d'explorer toutes les nouvelles possibilités techniques offertes par le texte électronique, mais il s'agit aussi d'autre part de pointer vers les problèmes économiques qui pourraient faire jour par la multiplication d'ouvrages électroniques dont la distribution ne serait pas contrôlée. Pour rester dans le cadre des possibilités nouvelles qui peuvent être offertes au lecteur, nous devons principalement nous pencher sur la façon de maîtriser l'hypertextualité au travers des parcours offerts et anticiper ces navigations, mais il est vrai que nous pourrions aller jusqu'à explorer les possibilités de transformation du texte vers l'audio, rendues facilement automatisables grâce au média numérique.

Finalement, ce travail sur l'acte de lire lié au texte électronique et à sa mise en scène hypertextuelle doit nous amener à voir comment se construit le sens dans un hypertexte, dans la manière de le construire et aussi pendant que la navigation du lecteur. Le lecteur focalise souvent la réflexion sur l'évolution de l'imprimé vers l'écrit, parce qu'il est le destinataire final d'un long processus de mise à disposition du texte. Il ne faudrait oublier que de l'auteur au bibliothécaire, toute la chaîne de l'édition est de près ou de loin impactée par cette évolution vers le numérique.

*« On est en droit de s'interroger sur les effets dus aux changements des dispositifs de lecture, à savoir le passage de la feuille imprimée à l'écran. De la même façon, dans son rôle de médiateur, le bibliothécaire est à l'origine des métadonnées : les fichiers matières et auteurs qui sont les outils cartographiques de l'organisation spatiale des bibliothèques. Là encore cette instrumentation implique une certaine vision du monde dont l'origine est attribuée à Callimaque, qui fut, semble-t-il le premier à structurer – classifier – la bibliothèque d'Alexandrie.[...] Les techniques*

*numériques effacent ces multiples repérages du savoir, mais en proposent d'autres. Elles invitent en conséquence à revisiter les méthodes de conservation du patrimoine intellectuel des individus et de la collectivité.*

*(Dupoirier, 1999, page 11) »*

En fait, nous sommes au cœur de l'évolution et comme nous la vivons il nous est souvent difficile de prendre le recul nécessaire. La tendance est de rendre le texte, à travers lui des idées, disponible toujours plus facilement à toujours plus de personnes. Dans ce sens, le numérique n'est pas une révolution mais la continuité logique de ce qui a commencé avec Gutenberg.

*« Nous sommes à l'aube de l'après livre ! Il y a eu les manuscrits médiévaux, puis est venue la révolution Gutenberg grâce à laquelle le livre a cessé d'être un objet unique, archi-élitiste, la prochaine étape verra la disparition du livre comme tel mais pas du texte, évidemment. (Butor, 11 décembre 1993) »*

**Lecture de documents papier** La communication écrite est largement influencée par le support papier, ainsi que par l'utilisation de l'imprimerie, mais cela a été complètement assimilé dans nos usages et ce n'est que rarement que cela est remis en cause. La taille des pages correspond à des règles très strictes, pour des raisons d'industrialisation de l'impression et d'optimisation du rangement des volumes sur les étagères des bibliothèques. Les polices de caractères utilisées sont très influencées par le fait qu'elle doivent être reproduites en plomb, les tailles sont fixées assez rigide<sup>44</sup> et les fantaisies sont généralement interdites. Même si depuis maintenant plusieurs années ce ne sont plus des caractères de plomb qui sont utilisés mais des polices

---

44. La police de caractères Times par exemple, tire son nom du quotidien britannique qui l'a mise au point. Sa caractéristique est un compromis entre la lisibilité et l'étroitesse des caractères pour permettre un remplissage satisfaisant des colonnes.

de caractères numériques utilisées en photocomposition, les mêmes règles continuent de contrôler la plus grande partie de la typographie, qu'il est possible de retrouver par exemple dans (Perrousseaux, 2000). A ces règles sont d'ailleurs venues s'ajouter les contraintes induites par l'utilisation de claviers d'ordinateurs aux possibilités restreintes, c'est ainsi que par exemple les majuscules accentuées sont devenues interdites dans les cours de dactylographie, alors qu'elles étaient simplement devenues impossibles à réaliser à cause du dispositif technique de frappe. François Richaudeau dans (Richaudeau, 1999) fait une analyse intéressante du conservatisme des usages typographiques et propose une nouvelle façon de présenter le texte, présentée par son éditeur.

*« Il est important de bien comprendre que les propositions typographiques de François Richaudeau répondent à une autre façon de lire : celle dite « en mosaïque », « en écrémage », celle permettant des « repères » au lecteur survolant Internet ou un programme multimédia. (Richaudeau, 1999, page 128) »*

Ces propositions comportent l'utilisation du souligné pour mettre en valeur des mots-clés importants, une nouvelle ponctuation utilisant le point-virgule pour découper la phrase autour de l'articulation des idées qu'elle permet de développer, une capitale grasse précédée de deux espaces pour baliser le début des phrases et permettre au lecteur de sauter plus facilement de phrases en phrases. Ces « outils » se sont développés petit à petit, avec les progrès du livre : manuscrit, imprimé traditionnellement, imprimé depuis un support informatique, ... et ils se sont rendus indispensables parce qu'ils participent de l'appropriation qu'un lecteur fait de l'ouvrage qu'il a entre les mains. Mais ces outils ne sont pas tout, la qualité du papier, l'éditeur, la date d'édition mais aussi l'usure d'un ouvrage peuvent aussi faire sens.

### **Lecture de documents électroniques : De nouveaux usages à inventer?**

La première question que nous sommes amenés à nous poser consiste à mesurer l'impact du document numérique sur la lecture. Il faudrait pour cela d'abord définir en quoi la lecture d'un livre papier procède de mécanismes particuliers, mécanismes induits par le support et que nous ne pourrions pas retrouver tels quels sur un support électronique. Mais à chercher à qualifier les spécificités de la lecture d'un ouvrage papier, nous allons devoir considérer les différents supports papiers disponibles aujourd'hui, comment ils ont évolué et comment ils continuent d'évoluer. Il paraît évident que nous n'abordons pas la lecture d'un magazine comme celle d'un livre savant. Le papier est lui aussi, plus faiblement, multimédia : les gravures ou les photos, le colonage, le format, la qualité du papier, tout cela contribue à placer le lecteur dans un certain état d'esprit, à le préparer à la lecture. Le magazine sera volontiers feuilleté à la recherche d'un indice qui « accroche » l'œil et donne envie de commencer la lecture à cet endroit. La lecture sur écran ne change pas *vraiment* quelque chose, c'est plutôt la tournure d'esprit du lecteur qui se prépare à entrer dans un document qui est l'aspect prédominant. Le roman sera bien entendu lu séquentiellement (nous mettons à part les impatientes et les romans policiers). Lors du Colloque virtuel : « *Écrans et réseaux, vers une transformation du rapport à l'écrit* »<sup>45</sup> qui s'est tenu du 15 octobre 2001 au 15 mars 2002, des conférenciers ont exploré l'impact d'Internet sur la lecture, l'écriture et la diffusion du savoir. L'impact le plus fort qu'aura le livre numérique sur la lecture ne sera sans doute pas du à la nouveauté du médium numérique mais plutôt à l'utilisation d'Internet pour publier, diffuser et communiquer des ouvrages électroniques. En particulier parce qu'Internet permet une conver-

---

45. Colloque organisé dans le cadre de « Lire en fête sur Internet » par la BPI (Bibliothèque Publique d'Information) avec l'Institut Jean Nicod (CNRS et EHESS) et l'association Euro-Edu. Les actes sont disponibles en ligne à l'adresse <http://www.text-e.org/>.

sation simultanée, comme dans les salons virtuels par exemple, ou sur les Wikis dont nous reparlerons, ce qui n'existe pas avec les autres moyens de la communication écrite. Utiliser la communication simultanée permet de faire tomber les distances et de permettre à tous les internautes de participer aux débats et à la construction d'un nouveau savoir ensembles. La nouveauté est donc bien la possibilité de cette simultanéité, qui rapproche l'écrit de l'oralité. Nous pourrions alors continuer notre raisonnement en constatant que l'écrit est en évolution permanente, il n'y a peut-être de différences que dans la vitesse de l'évolution. La même évolution qui a accompagné l'écrit jusqu'à présent est sans doute celle qui contribue à le pousser vers le numérique. L'évolution du support, papyrus, parchemin, livre manuscrit puis imprimé a influencé notre rapport à l'écrit ; l'électronique n'est peut être que la suite logique de cette évolution.

*« La succession de l'oralité, de l'écriture et de l'informatique comme modes fondamentaux de gestion sociale de la connaissance ne s'opère pas par simple substitution, mais plutôt par complexification et déplacement de centres de gravité. (Lévy, 1990, page 10) »*

Ces trois étapes que Pierre Levy cite comme une succession de modes de gestion sociale ne sont pas antagonistes mais complémentaires. Comme l'écrit n'a pas remplacé le discours mais a permis de le perfectionner, de l'appuyer, nous voyons avec l'informatique un nouveau moyen de renforcer l'écrit et de renforcer l'oralité. L'écrit numérisé est en train de se trouver une nouvelle place et la lecture se réorganise autour. Alors que le e-book semble être un échec, nous assistons à une débauche de moyens de communication écrite, surtout orientés vers la communication personnelle : les blogs (journaux intimes en ligne), les sites personnels, le courrier électronique et à la limite

les SMS<sup>46</sup> sont autant de façons d'écrire et surtout de lire. De plus, l'écriture collective est elle aussi facilitée par l'utilisation de dispositifs particuliers, tels que les forums ou encore les wikis que nous prendrons précisément pour exemple de dispositif d'écriture hypertextuelle en groupe dans la section C., page 76 du chapitre 4, ce qui sera l'occasion de mener une réflexion plus approfondie sur le travail collaboratif et ses outils. L'impact de ces dispositifs sur la lecture est sans doute important, parce que les écrits sont plus spontanés, fragmentés. Ils incitent à passer rapidement d'un sujet à un autre et les imprécisions orthographiques ou typographiques sont plus facilement tolérées, sous prétexte de spontanéité.

**Instrumentation matérielle et logicielle** Avec un livre électronique les repères ne sont plus les mêmes, mais le lecteur cherche encore trop souvent à s'orienter par rapport aux repères qu'il connaît, qu'il maîtrise : ceux du livre papier, d'où incompréhension et réticence. L'écran n'est qu'une fenêtre sur l'ouvrage électronique, au départ plus petite qu'une page imprimée traditionnelle, elle commence à être souvent plus grande. Mais ce qui rend peut-être le livre imprimé toujours plus souple, c'est que nous pouvons en ouvrir plusieurs à la fois sur notre bureau ! Le bureau « virtuel » de notre ordinateur est encore bien trop petit.

La taille des pages doit être modifiée, faute de quoi l'ergonomie de lecture est mise en péril par la taille de l'écran, comme l'avait déjà noté Delphine Rosenthal

*« L'écran offre une surface d'autant plus réduite et limitée de visibilité qu'il n'est pas seulement dédié à la présentation des données, mais*

---

46. Ce qui pourrait être un sujet d'étonnement, ces mini-messages écrits sont envoyés à l'aide d'un téléphone portable, outil *a priori* de communication orale.

*qu'il tend conjointement à servir d'instrument de commande. (Rosenthal, 1993, page 67) »*

Les techniques de navigation ne sont plus les mêmes et les outils changent, par exemple l'index devient obsolète puisqu'il est généralement possible de faire des recherches sur n'importe quel mot, dans la totalité du texte. C'est aussi dans ce cadre que s'inscrit notre démarche, parti de la recherche de nouveaux usages et de nouvelles pratiques de lecture, nous nous trouvons finalement à étudier à la fois de nouveaux instruments de création, mais aussi plus simplement de nouveaux outils d'édition.

Car n'est-ce pas finalement le métier de l'édition qui devrait se retrouver au centre des préoccupations liées à la mise en ligne de contenus sur le Web. Un texte sur Internet, en dehors des sites institutionnels, est souvent suspecté d'être peu fiable parce qu'il n'est pas toujours possible de connaître la source des informations présentées, leur date d'édition, de modification, ... Mais si nous décidons de faire abstraction de tout ce qui n'est pas officiellement publié, nous passons à côté de millions, peut-être de milliards de pages. Nous reviendrons sur ce sujet avec les travaux du Web sémantique, qui propose une nouvelle manière d'aborder ce problème. Puisqu'il n'est plus possible humainement de faire face à la quantité d'information mise en ligne sur le Web, il s'agit de mettre au point des outils et des agents logiciels qui le feront à notre place. Plutôt que d'aller tout les jours vérifier les nouveautés sur ses sites de référence, l'internaute demande à un agent de scruter ces sites pour lui et de ne lui rapporter que les résumés des nouveautés. Ce qui lui permet d'avoir une vue d'ensemble et d'aller lire uniquement les articles qui l'intéressent.

*« L'outil se développe indépendamment et au détriment des besoins culturels et crée lui-même ses propres besoins pseudo-culturels. Il est rare que les gadgets et fausses innovations s'implantent durablement*

*lorsqu'elles ne correspondent pas à des attentes du public. Cependant ce qui s'implante durablement c'est l'habitude de consommer des gadgets, l'un remplaçant l'autre et ainsi de suite. (Rosenthal, 1993, page 124) »*

**Séquentialité contre hypertextualité** Ce qui oppose de manière flagrante la séquentialité à l'hypertextualité c'est la navigation et la possibilité de création d'un parcours à travers le texte, qui devient hypertexte. Le problème qui découle de la navigation, est qu'il nécessite la présence d'un fil d'Ariane pour que le lecteur puisse retrouver son chemin à travers le corpus.

*« Le cas du document littéraire est emblématique de cette nécessité de « fil d'Ariane ». La génétique textuelle représente un savoir spécialisé de la littérature qui précisément a besoin que les bibliothèques soient parfaitement aménagées en ce qui concerne les mentions d'éditions, les mentions de transformation et le repérage des annotations d'auteur ou de lecteurs célèbres, ce qui devient possible dès lors que se mettent en place des bibliothèques virtuelles sur lesquelles peuvent s'appuyer une moderne « lecture savante ». (Rouet, 2000, page 44) »*

Ces mentions peuvent être ajoutées au texte de plusieurs manières, visibles ou invisibles, en fait la façon qui l'emporte est de faire usage des métadonnées pour enrichir le texte de manière à ce que des agents logiciels soient capables de les interpréter. C'est ensuite à la charge de ces agents logiciels de tisser un fil d'Ariane au fur et à mesure que le lecteur s'enfonce dans le labyrinthe hypertextuel.

## **6. Modélisations sémantiques et cognitives**

*« Grâce aux technologies de l'information, l'écrit n'a jamais été aussi omniprésent dans notre vie quotidienne, à la fois sous forme papier et sous*

*forme électronique, de sorte que les nouvelles technologies apparaissent aussi comme un formidable moteur de développement de la production écrite. (Vignaux, 2002) »*

La place de l'écrit est quelquefois remise en question par les nouveaux médias, elle demeure pourtant prépondérante. Cela ne doit pas toutefois nous empêcher de nous interroger sur les formes et suivant quelles modalités l'écrit garde-t-il sa place ? Pour certains, l'écrit est en révolution depuis l'arrivée du support numérique, et surtout de sa médiatisation rendue possible grâce à ce support. À terme ce sera soit le Web soit le livre électronique qui marqueront de façon certaine cette révolution. C'est cette mutation qui est annoncée dans les propos de Roger Chartier :

*« [La mutation électronique du livre sera] une violence faite aux textes dès lors séparés des formes classiques, qui ont contribué à construire leurs significations historiques. [...] L'univers des textes électroniques signifiera nécessairement un éloignement vis-à-vis des représentations mentales et des opérations intellectuelles spécifiquement liées aux formes qu'a eu le livre en Occident depuis dix-sept ou dix-huit siècles. (Chartier, 1994)(Chartier, 1997) »*

Qu'en sera-t-il donc des nouveaux usages de l'écrit qui ainsi, vont apparaître et quels types d'appropriation vont-ils permettre, susciter ou même inhiber ? Cette question des modalités et des formes de l'appropriation par le lecteur ou l'internaute est essentielle et elle conditionne la mise en place de stratégies futures de diffusion des savoirs. Or nous n'avons que peu d'observations, d'analyses et de synthèses sinon fragmentaires, disponibles sur ces processus d'appropriation.

Nombre de problèmes demeurent à explorer en ce qui concerne les rapports entre l'écrit électronique et l'appréhension et l'acquisition de connaissances. Nous pouvons citer :

- l'influence des supports (électronique versus papier, base de données, hypertexte, etc.),
- l'impact des types de textes (littéraires, scientifiques, argumentatifs, etc.),
- les formes stylistiques et les degrés de complexité des textes,
- les modalités du raisonnement et de l'argumentation impliquées selon les textes,
- les rapports images-textes,
- les processus de navigation et les parcours de lecture selon les textes.

Le papier comme l'électronique ne sont « que » des formats de stockage, pourtant la similitude s'arrête quasiment là. Pour être utilisé le papier ne nécessite que de l'encre. En revanche l'électronique n'est pas un support matériel et nécessite de mettre en œuvre un support physique : ordinateur de bureau ou portable, PDA (Personal Digital Assistant, ordinateur « de poche ») ou même matériel dédié. De plus, ce matériel va nécessiter des logiciels, pour l'animer. Nous avons là une complexification des modalités de lecture, une multiplication des interfaces entre les idées et le lecteur, qui va aller en s'intensifiant au fur et à mesure que nous chercherons à définir une nouvelle médiation de la lecture. Pourtant, l'objectif de base est le même, conserver les idées et les transmettre en les mettant à disposition dans un dispositif de lecture, la complexité de ce dispositif ne devrait pas avoir d'impact sur les idées qu'il véhicule.

*« L'utilisation des systèmes hypermédias est à la fois très proche et très différente de la lecture de documents conventionnels. [...] dans les*

*deux cas il s'agit de lire et de comprendre des informations essentiellement verbales. (Rouet, 2000, page 10) »*

Le livre électronique n'est sans doute que l'évolution naturelle du livre papier et il reposerait à ce titre sur des principes d'organisation du savoir nés il y a plusieurs siècles. Il y a malgré tout indéniablement du nouveau avec le livre électronique. Cette nouveauté apporte sans aucun doute une charge cognitive qui n'est pas nulle. En revanche, il est pour l'instant difficile de la mesurer, en fait il est même difficile de l'identifier clairement et de qualifier ce qui est plutôt dû au dispositif de lecture (l'écran), ce qui est plutôt du ressort de l'ergonomie et de l'interface, et enfin ce qui est la conséquence de la numérisation du texte et de la malléabilité qui en découle. Le passage du livre papier au livre électronique devrait se faire avec l'appropriation de ces nouveaux dispositifs, qui est liée à l'éphémérialisation<sup>47</sup> des techniques utilisées dans les réalisations d'ouvrages numériques. En fait ces techniques tournent autour de l'évolution d'un certain nombre de concepts, barbarismes de la modernité, parmi lesquels nous pouvons citer :

**Hypertextualité** Nous reviendrons longuement sur ce concept dans le chapitre 4 - **Réflexions sur l'hypertexte** (page 163) qui lui est entièrement consacré.

**Interactivité** L'interactivité peut être considérée comme la mise en relation structurée du lecteur avec un dispositif électronique. Alors que le livre n'autorise *a priori* que la lecture séquentielle mais autorise le feuilletage de manière complètement aléatoire, l'interactivité qui donne en quelque sorte la vie à l'hypertexte n'autorise finalement qu'un choix réduit de parcours. Nous

---

47. Ce concept inventé par l'architecte Buckminster Fuller sera détaillé dans le chapitre 8, plus précisément dans la section consacrée aux textes fondateurs, page 176.

pouvons surtout l'interpréter comme un dialogue lecteur-dispositif. Nous reviendrons là aussi sur ce concept dans le chapitre 4, plus particulièrement dans la section A. **Interactivité** (page 73) parce que nous avons besoin de plus que quelques lignes pour l'analyser et c'est surtout son utilisation dans le cadre de l'hypertexte qui nous intéresse.

**Multimodalité, ou multimédia** En revanche, la multimodalité peut nous amener à réfléchir sur les aspects sémiotiques introduits par le support numérique et pour cette raison nous allons l'aborder dès à présent, même si ce concept aurait pu trouver lui aussi sa place aux côtés des deux premiers dans l'étude de l'hypertexte.

La multimodalité est intéressante par rapport au lien qu'elle établit entre support et contenu, même si tout est fait pour permettre d'affranchir le fond de la forme, il faut admettre, et ce n'est pas nouveau, que le support a une influence non négligeable sur ce qu'il transporte. Déjà sur les tablettes d'argile des mésopotamiens anciens, le support et les outils pour le travailler avaient leur importance. Sans doute pour faciliter l'interprétation, pour mieux codifier les signes porteurs de sens et pour éviter les ambiguïtés, seuls quelques poinçons étaient autorisés, ce qui dessinait des graphismes élémentaires à partir desquels tous les messages pouvaient être construits. Avec le parchemin, des enluminures pouvaient bien orner les lettrines, mais les impératifs de la copie limitaient nécessairement l'empire du décor. La presse de Gutenberg, dans ses débuts tout au moins, réduisit plus encore la part de l'image. Seules quelques gravures, généralement grises ou monochromes, sur bois, plomb ou cuivre, étaient autorisées. Ce n'est qu'au XX<sup>e</sup> siècle, lors de l'avènement de la quadrichromie, que l'image prit son plein essor dans le livre devenu désormais coloré grand teint. Avec les nouvelles techniques de l'information,

c'est maintenant un foisonnement : non seulement l'image en couleur orne le texte partout, mais on peut y adjoindre à loisir, son, musique ou voix, dessins animés qui deviennent parfois même interactifs, et enfin séquences filmées. Bref, plusieurs modalités de présentation de l'information, textuelle, visuelle fixe, visuelle animée, auditive, peuvent coexister simultanément dans la même œuvre. C'est ce que l'on appelle communément la multimodalité.

Dans les feux d'artifice colorés et bruyants qui éclatent ici comme le tonnerre dans l'atmosphère sereine et recueillie des bibliothèques silencieuses, beaucoup donnent le rôle fort à l'image : elle serait le pivot qui aurait fait basculer de l'ère du livre d'école ennuyeux et passéiste à l'ère de la modernité résolue et polychrome qui gît derrière l'électronique. Or, ce statut magnifié de l'image, si souvent invoqué comme révélateur de la modernité présente, demande à être reconsidéré, ou, plutôt, à être mis en regard de ce à quoi il s'oppose, ne serait ce que parce que l'image est antérieure à la diffusion massive de l'écrit par le livre imprimé : l'église, par ses chants, ses prêches, ses sermons, ses peintures, ses vitraux, ses ornements, son encens, ses hosties et son vin de messe fut elle aussi multimodale. Elle a su s'adresser à presque tous les sens, visuel, auditif, gustatif, olfactif, à l'exclusion du seul toucher ; encore que toute une tradition de romans licencieux en ferait douter ; pour éduquer aux mystères de la foi une société majoritairement analphabète qui n'avait donc pas un accès direct aux écritures. Avec l'électronique, nous courons derrière, loin derrière l'église pour ce qui touche à la multimodalité !

**Sémiotique** Sous l'image, célébrée et magnifiée, il y a différentes modalités de communication : visuelle fixe (photographie), par dessins, par dessins interactifs où des zones « chaudes » sont associées à des procédures, par dessins animés, par films muets, par films parlants... Au plan technique, ces

modalités se ramènent toutes, plus ou moins, au même, ou presque, modulo quelques changements de représentation. Ainsi, si un film demande 24 images pleines par seconde, un dessin animé demande moins car d'une image à l'autre, les plans glissent les uns contre les autres, sans que l'ensemble soit redessiné. En revanche, au plan cognitif, cela ne revient pas au même : un texte est visuel, ce n'est pourtant, en rien, l'équivalent cognitif d'une photo. Nous avons recensé au moins quatre oppositions qui se font jour sans que l'on sache, pour l'instant, laquelle est la plus pertinente pour rendre compte des dimensions cognitives du livre électronique et de la rupture d'avec le livre papier :

**Le visuel (photo, film, schéma...) opposé au verbal (texte, parole...)** Ce serait l'opposition classique géométrie contre algèbre, image contre langage que l'on retrouverait ici ; mais, dans ce cadre, où placer la musique ? où mettre la poésie faite d'euphonie, de rythme, et, surtout, d'images à la fois au sens propre de disposition dans la page, et au sens figuré, de métaphore ? Autant de questions non encore résolues.

**Le visuel opposé au sonore** le texte serait alors mis aux côtés de l'image et du film, tandis que la parole prononcée serait placée aux côtés des sons avec la musique. Au plan formel, une telle distinction est plus satisfaisante que la précédente, mais elle ne rend en rien compte de la rupture produite par l'irruption de l'image et de la vidéo dans le monde du texte imprimé.

**L'image opposée à l'informationnel.** L'image couvrirait alors l'information brute, non structurée, non interprétée telle la photo ou la parole. L'information, quant à elle serait représentée à l'aide de structures algébriques sur lesquelles des opérations syntaxiques seraient possibles.

De la sorte, un texte et un schéma seraient informationnels ; une interprétation musicale serait image.

**L'image visuelle ou sonore opposée au scriptural.** Dans cette perspective, l'écrit serait une modalité toute particulière d'accès à la connaissance qui a transmis, pendant des siècles, les ouvrages de l'esprit intacts, ou presque, alors que les voix s'étaient à jamais tues et que les tableaux avaient disparu. L'image, visuelle ou sonore s'opposerait donc à cette modalité comme le futur au passé : le scriptural demeurerait dans le passé le seul véhicule de la présence, ou presque, si tant est que la sculpture se brise, que la peinture se fane et que les fresques se pulvérisent ; or aujourd'hui, l'image numérisée dans ses modalités visuelle et sonore devient écriture, elle-même et, par là, elle rejoint le scriptural.

En conclusion de cette section sur la multimodalité, il apparaît que tout peut être réinterprété en invoquant le statut sémiotique des objets manipulés. Ce sont tous des signes, mais le mode de donation du sens et la fixation de la référence s'opèrent différemment selon le statut de ces objets ; les transformations autorisées sur ces objets sont elles-mêmes révélatrices de ce statut sémiotique : une photo ne se manipule pas de la même façon qu'un mot, ou un schéma. En se conformant aux catégories sémiotiques de Peirce, on distingue les icônes, les indices et les légisignes selon qu'il y a un rapport d'analogie entre le signe et le signifié (cas d'une photographie ou de la poubelle du Macintosh), que le signe indique ce qu'il signifie (ce que fait un curseur sur un traitement de texte lorsqu'il indique une position), ou enfin qu'il y a une loi externe et arbitraire qui cheville le signe à sa référence (par exemple, lorsque le mot « fichier » désigne telle entité informatique). Ces catégories nous semblent pertinentes dans une première approche, encore qu'il conviendrait de les affiner encore et d'en préciser le contenu. Dans le

cadre écrit, ou plus exactement, le scriptural relèverait de la catégorie du légisigne, les photographies, les films et les paroles enregistrées relèveraient des icônes, enfin, les schémas et les sons pourraient relever soit des indices, c'est le cas d'un signal d'alerte, par exemple du bip sonore d'un répondeur téléphonique, ou d'icônes.

## 3.5 Conservation numérique

### A. Les formats

Dans le cadre de la numérisation d'un corpus, comme dans un cadre plus restreint de constitution d'une bibliothèque numérique, la conservation numérique doit être mise en place de manière à assurer la pérennité des œuvres numérisées. Étrangement, le numérique est souvent associé à une idée de conservation améliorée dans le temps, parce que le signal enregistré est numérique au lieu d'analogique il est susceptible d'être contrôlé, vérifié voire corrigé si une altération est détectée. C'était un des arguments commerciaux véhiculés avec l'arrivée du disque compact (CD) pour remplacer les fragiles disques vinyles par exemple. C'est toujours un argument pour accompagner les lecteurs de DVD et autres caméscopes numériques. Derrière ce message rassurant, il se cache en fait une réalité qui montre que le support numérique est au contraire très fragile. Il n'est pas fragile en lui-même, mais parce qu'il doit intégrer une chaîne complexe d'outils et de dispositifs régulièrement frappés d'obsolescence. Pour ce qui concerne la conservation numérique, le problème majeur apporté par ce nouveau format est qu'il crée une distance entre le contenu et le lecteur, comme nous le rappelle Christain Rossi :

*« Un document numérique n'est pas directement accessible à un utilisateur humain Une caractéristique importante du numérique est*

*qu'il existe de nombreux intermédiaires entre le support où est stocké l'information numérique et l'utilisateur :*

- le support*
- un lecteur*
- le système d'exploitation*
- un logiciel applicatif*
- les périphériques*
- un mode d'emploi*

*En premier le support (disque magnétique, cdrom...) où se trouve l'information physique, ensuite un lecteur qui permet de passer de l'information physique à une information binaire, puis le système d'exploitation qui permet de regrouper ces 0 et ces 1 en fichiers et en répertoires. Pour finir un logiciel applicatif qui rend ce fichier accessible à l'être humain par l'intermédiaire d'un périphérique audio, vidéo ou d'une imprimante, et ce pour les cas les plus classiques. Sans oublier, bien sur, une personne qui sache utiliser, si ce n'est réparer, tous ces intermédiaires. Il s'agit là une différence majeure par rapport au papier où l'accès à l'information est direct. (Rossi, 2005, page 1) »*

Un même ouvrage électronique peut se présenter sous différents formats, depuis les formats sources et jusqu'à des formats visualisables. Les éditeurs de texte *Wysiwyg* nous masquent cette répartition parce qu'avec ce type d'outil le format source est le même que celui qui sert à visualiser le document. En revanche, si nous regardons ce qui se passe dans le cas d'un document HTML, par exemple, nous avons des outils différents pour créer le document et pour le visualiser.

## **B. La numérisation comme changement de support**

La numérisation permet de changer de support sans changer la forme, nous passons d'un document papier à un document électronique, et ce document électronique peut prendre lui-même différentes formes : image, texte brut, texte formaté ou encore structuré par XML et pourquoi pas, si nous ajoutons encore une nouvelle étape et un nouvel outil, un fichier sonore.

Le changement de support peut s'opérer de plusieurs manières, en fait d'autant de manières que le format électronique peut en supporter et aussi d'autant d'outils que les ingénieurs sont capables de développer pour numériser un texte.

Nous pouvons retrouver sous format électronique la copie « conforme » (au support près) de l'original papier, c'est-à-dire pour simplifier, une liasse de photographies électroniques. A l'inverse, nous pouvons retrouver le texte enrichi de liens hypertextuels vers diverses bases de connaissance, dictionnaires ou encyclopédies et qui n'aura plus qu'un lien assez lointain avec son original papier. Bien entendu, entre les deux nous pouvons imaginer toutes les combinaisons, entre texte électronique et images.

## **C. La chaîne de numérisation**

La technologie exigée pour pouvoir naviguer d'un bout à l'autre de la chaîne de numérisation consiste principalement en matériel, logiciels et réseaux. Ils seront au centre de notre attention dans cette partie. Une vision complète de l'infrastructure technique comprend également les protocoles et normes, politiques et procédures (concernant le déroulement du travail, la maintenance, la sécurité, les mises à jour, etc. . .) et les niveaux de compétences et responsabilités de travail de personnel de l'organisme.

Néanmoins, même les écrous et les boulons ne peuvent être évalués en isolation totale. Parmi les actions et les considérations liées pouvant affecter les décisions concernant l'infrastructure figurent :

- Le choix des exigences de qualité requises, basées sur les attributs de documents (Etalonnage)
- L'évaluation des forces et des faiblesses institutionnelles, le calendrier et le budget (Gestion)
- La compréhension des besoins utilisateurs (Présentation)
- L'évaluation des projets à long terme (Conservation numérique)

Les décisions concernant l'infrastructure technique nécessitent une planification précise car la technologie de l'imagerie numérique évolue rapidement. Le meilleur moyen de minimiser l'impact de la dépréciation et de l'obsolescence est d'effectuer une estimation minutieuse, et d'éviter les solutions propriétaires uniques. Si les choix d'équipement correspondent parfaitement aux utilisations prévues et résultats prévus, et s'ils sont synchronisés avec des calendriers réalistes, le retour sur investissement sera maximal.

### **1. Soin continu**

Le Soin continu doit être considéré comme la base d'une stratégie de conservation numérique. La supervision du bon état des ressources numériques commence par la gestion de la sécurité et le contrôle des accès à ces ressources. Cela va bien sur bien au delà puisqu'il faut être capable de préserver physiquement ces ressources. Une gestion vigilante de la collection comprend l'hébergement des images, des textes et une gestion intelligente des sauvegardes en assurant l'accompagnement des fichiers vers des destinations et médias de stockage sûrs et fiables. L'entreposage et la manipulation des médias doit être conforme aux recommandations de l'industrie afin d'optimiser

leur durée de vie, le choix des lieux de stockage doit permettre de compléter la sécurisation des médias tout en optimisant les délais de restauration en cas de besoin. Dernier point, la mise en place de vérifications et de sauvegardes périodiques et systématiques doit être cohérent avec le rythme des mises à jours effectuées sur les ouvrages numériques.

## **2. Rafraîchissement**

Le Rafraîchissement implique la copie du contenu d'un média de stockage vers un autre pour contourner les problèmes d'obsolescence des médias. En ce sens il ne constitue pas une stratégie de conservation à part entière, en pratique il pourra être combiné au soin continu mais il faudra alors veiller à ne pas avoir de conflit entre type de média et date de sauvegarde, le problème typique étant la situation où l'on dispose d'une sauvegarde mais plus du matériel pour l'utiliser. Pour cette raison, le rafraîchissement devrait être considéré comme une partie intégrante d'une politique de soin continu. Un exemple de rafraîchissement est la copie de fichiers de CD-Rom vers des DVD.

## **3. Migration**

La Migration correspond au processus de transfert des informations numériques d'une plate-forme logicielle et système vers une autre, ou d'une génération d'ordinateurs vers des générations suivantes. Il s'agit d'une action complémentaire de celle de rafraîchissement, cette fois ne se limitant pas au support physique mais étendue à l'architecture technique qui donne accès aux textes numérisés. Par exemple, le déplacement de fichiers d'un système basé HP vers un système SUN implique l'adaptation à la différence entre les deux environnements. La migration peut également être basée sur le format, de la

transformation de fichiers images d'un format obsolète ou afin d'améliorer leur fonctionnalité, de maximiser l'espace utilisé ou de se mettre à l'abri de contraintes liées à l'utilisation de formats propriétaires protégés. À ce titre, l'utilisation de formats ouverts doit être encouragée pour minimiser les opérations de migration dues à la disparition ou au verrouillage d'un format propriétaire.

#### **4. Émulation**

L'Émulation sera rendue nécessaire à la suite d'une faille dans les processus de migration ou de rafraîchissement, qui aurait entraîné la disparition de l'environnement technique nécessaire à l'accès à certaines informations avant qu'elles aient été migrées vers un nouveau format. Cela implique donc la création à nouveau de l'environnement technique requis pour visualiser et utiliser une collection numérique. Ceci ne pourra être réalisé que si ont été conservées les informations et les spécifications techniques concernant les besoins de matériel et d'applications afin de pouvoir reconstruire le système.

#### **5. Conservation de la technologie**

Pour synthétiser ces réflexions autour de l'aspect technologique de la conservation, il faut bien prendre garde de considérer le problème dans son ensemble. Nous avons parlé de chaîne de numérisation et comme avec l'image de la chaîne et du maillon faible, il est inutile de payer des technologies coûteuses pour assurer une pérennité d'un document à travers les siècles si dans la chaîne un des maillons n'a qu'une durée de vie limitée. De plus ce n'est pas toujours la disponibilité technique ou physique qui est en cause, mais il s'agit le plus souvent de la disponibilité commerciale. Dans un monde soumis aux lois du marché, les ordinateurs, leurs périphériques et leurs logiciels sont

très régulièrement renouvelés et rendus obsolètes. C'est ce qui fait parler (Rossi, 2005) d'obsolescence technico-commerciale.

*« Pour la conservation, une conséquence importante de l'existence de ces intermédiaires est que ce n'est pas seulement le support qui doit être disponible mais l'ensemble de cette chaîne de consultation, si ce n'est l'originale une équivalente.[...] Assurer la conservation de documents numériques c'est donc actuellement choisir entre migrer en permanence l'ensemble des supports et des chaînes matériels et logiciels ou bien créer des musées vivants de l'informatique où sont conservés en état de marche ces intermédiaires. En fait on en arrive à devoir migrer non pas à cause du support qui a atteint sa limite de vie physique mais à cause de son obsolescence technico-commercial : le lecteur n'est plus vendu, plus maintenu et les caractéristiques techniques du support sont dépassées. Dans cette optique il est inutile d'utiliser un support dont l'espérance de vie est de cent ans puisque dans trois ans il faudra le migrer. L'espérance de vie technico-commerciale est plus faible que l'espérance de vie physique. (Rossi, 2005, page 1) »*

La Conservation de la Technologie est basée sur la préservation de l'environnement technique supportant le système, y compris les logiciels et le matériel tels que systèmes opérationnels, logiciels et applications originaux, supports médias, et similaires. La gestion efficace de collections numériques exige des organismes qu'ils développent et suivent un projet commercial pour évaluer les besoins de préservation à long terme et d'accessibilité, identifier les coûts et les bénéfices et évaluer les risques.

Parmi les exemples d'initiatives supportant une telle approche figurent :

- L’AHDS, le Arts and Humanities Data Service <sup>48</sup>(Service des Données Artistiques et Sciences Humaines), situé au Royaume-Uni, met au point un arbre de processus décisionnel pouvant être utilisé dans l’analyse coût-bénéfice des options de conservation numérique (Cadres administratifs et de gestion des contenus numériques).
- Le projet Risk Management of Digital Information<sup>49</sup> de Cornell (Gestion de Risque des Informations Digitales) a examiné les risques liés à la migration de formats de fichiers (p.ex. du format TIFF 4.0 au TIFF 6.0) et a mis au point un outil d’évaluation afin d’estimer les risques liés à la migration. Cet outil aide également à évaluer le degré de préparation des organismes avant toute opération de préservation numérique.

Les initiatives suivantes offrent des exemples d’approches pratiques et prometteuses de la conservation numérique :

Le modèle de référence de l’OAIS<sup>50</sup> (Open Archival Information System) fournit un cadre pour la préservation et l’accessibilité à long terme, y compris la terminologie et les concepts de description et de comparaison des architectures d’archivage. Les projets NEDLIB et Cedars 1 ont tous deux adopté le modèle de référence de l’OAIS comme base pour leurs recherches.

Le projet Cedars 1<sup>51</sup> (CURL Exemplars in Digital Archives) vise à créer un cadre stratégique concernant les politiques de gestion des collections numériques, et de promouvoir les méthodes appropriées à la préservation à long terme de différents types de ressources numériques, y compris la création des métadonnées appropriées.

---

48. <http://ahds.ac.uk/>

49. <http://www.clir.org/pubs/abstract/pub93abst.html>

50. <http://ssdoo.gsfc.nasa.gov/nost/isoas/us/overview.html>

51. <http://www.leeds.ac.uk/cedars/>

Le Networked European Deposit Library <sup>52</sup> (NEDLIB) est un projet commun des bibliothèques nationales Européennes afin de développer un cadre pour une bibliothèque de dépôt en réseau. Parmi les questions clefs qu'il aborde figurent les procédures de maintenance des archives et les liens entre les besoins de métadonnées et les stratégies de conservation.

Le projet PANDORA <sup>53</sup> (Preserving and Accessing Networked Documentary Resources of Australia) a mis au point avec succès une archive de publications australiennes en ligne sélectionnées, développé plusieurs politiques et procédures de conservation numérique, jeté les bases d'un modèle de données logique pour la conservation des métadonnées et a dessiné les contours d'une proposition concernant une approche nationale pour la conservation de ces publications à long terme.

### 3.6 Les corpus électroniques

*« Le corpus remplace le livre. Constitué des textes d'un même auteur ou de ceux d'un ensemble plus large (une anthologie d'un genre littéraire, une collection des ouvrages publiés dans une période donnée, etc.), le corpus est la nouvelle unité de lecture sur support électronique. (Clément, 2001) »*

De nombreux corpus électroniques existent, leur création peut être motivée par plusieurs raisons :

**Économiques** Parce que le support électronique permet de stocker un grand nombre d'information dans une place très restreinte, parce que le coût de la « mémoire » est de plus en plus faible.

---

52. <http://www.kb.nl/coop/nedlib/index.html>

53. <http://pandora.nla.gov.au/index.html>

**Diffusion** Un ouvrage sous format électronique peut être facilement transporté d'un ordinateur à un autre, que ce soit parce qu'il est stocké sur un support amovible, ou mieux parce qu'il est diffusé à travers Internet. De plus ce format offre aussi l'avantage de la versatilité : le même ouvrage électronique pourra s'afficher de manière différente dans des applications différentes et donner lieu à des usages variés (lecture « simple », lecture érudite dans une encyclopédie, ...)

**Pérennisation** Encore que ce point soit discutable et nous allons y revenir dans la section 1. suivant, la numérisation d'un corpus peut être utilisée pour sauver des ouvrages papiers en train de se détruire.

En fait, ces trois raisons sont intimement liées, si la pérennisation passe par la numérisation, c'est aussi pour des raisons économiques. La possibilité de diffuser facilement le corpus se combine de la même manière puisqu'elle va permettre de « vendre » l'accès au corpus ou à certain de ses ouvrages. Nous pouvons d'ailleurs voir avec cet aspect un dangereux glissement, puisque maintenant ce n'est plus un ouvrage qui est vendu, mais le droit de le lire.

Nous reviendrons sur cette préoccupation dans le chapitre 8 (page 367) sur les Logiciels Libres, qui traite en fait plus largement du libre accès aux ressources électroniques.

## **A. Pourquoi un corpus en ligne ?**

### **1. Pérennisation et patrimoine**

La question de la conservation de l'information n'est pas neuve, mais elle se renouvelle. Quand il n'est pas possible de garder intact un ouvrage, c'est par la copie que son contenu a été préservé.

*« Avant l'invention de l'imprimerie, les systèmes d'information se déployaient avec les générations de « moines copistes » qui recopiaient les manuscrits de siècle en siècle. Ces manuscrits ne nous sont souvent parvenus que sous forme partielle : il arrive que l'original soit perdu et que l'on ne dispose que d'un ensemble incomplet de copies divergentes. (Boydens, 2004, page 14) »*

Il semblait qu'avec la dématérialisation un progrès allait être fait dans la conservation des documents écrits, mais en fait ce terme masque complètement la réalité. Même sous format électronique, un document a besoin d'un substrat matériel pour le contenir. La dématérialisation, au lieu de s'affranchir de ce substrat, l'a rendu technique. Comme nous l'avons vu avec la théorie du support numérique (page 58) en plus du vieillissement du matériau contenant les signaux électronique qui servent à décrire le document, nous devons maintenant aussi compter avec l'obsolescence des technologies qui permettent de lire ces supports ainsi que celles qui permettent d'en afficher le contenu.

*« La numérisation de documents patrimoniaux et l'accès qui en est proposé via l'Internet s'inscrivent dans le prolongement des nombreuses initiatives prises depuis une quinzaine d'années dans le domaine de la valorisation du patrimoine. En ce sens, et dans bien des cas, il faudrait plutôt parler d'expositions virtuelles (des objets accompagnés de légendes) que de mise à disposition de corpus indexés et structurés. Le caractère mal défini initialement de nombreuses entreprises fait qu'elles proposent à un large public des documents dont la singularité ou la complexité d'appréhension relève davantage des préoccupations de la recherche. Comme l'œuvre d'art présentée aux visiteurs d'un musée, le document patrimonial (ou le fragment) exposé sur la Toile est sorti de son contexte d'origine. En effet, le livre prend sens le plus souvent par rapport à ceux qui l'entourent*

*sur les rayons ; il appartient à une collection, à un fonds précis qui a son unité. Parfois, sa reliure importe autant sinon plus que son contenu. Le lieu dans lequel il est conservé n'est pas indifférent et fournit un autre élément de contexte. Tout se passe donc comme si la virtualisation de l'objet patrimonial s'accompagnait d'une « décontextualisation ». Ce phénomène est d'autant plus accentué que s'affiche sur l'écran non pas un volume entier dans son épaisseur tridimensionnelle, mais un extrait très bref, une sorte d'indice iconographique, de trace colorée inapte à traduire le sens cohérent de l'œuvre. L'objet ainsi présenté voit donc sa signification réduite à l'anecdote et affadie. Platement reproduit, il devient autre : une image de plus dans le bazar de couleurs et de sons des média. (Arot, 1998, page 125) »*

## **2. Travail collaboratif**

La construction d'un grand corpus électronique en ligne nécessitant de nombreuses ressources, ce n'est que rarement une institution isolée qui en prend la charge. Nous verrons dans la description des projets de la section B. suivante que les acteurs de ce type de projets sont généralement de grandes institutions universitaires ou nationales, elles mènent à bien le projet de numérisation et de mise en ligne grâce à plusieurs partenariats, chacun des partenaires apportant son savoir faire.

Il existe toutefois de grands projets lancés par des organisations non-gouvernementales et avec des moyens beaucoup plus restreints. Dans ce cas, la construction du corpus passe généralement par le travail bénévole de milliers d'internautes, qui mettent en ligne chacun un ouvrage. Pour éviter tout problème de droits d'auteur et parce que les ouvrages mis ainsi en ligne

sont ensuite librement accessibles de tous les internautes, les ouvrages retenus doivent être libres de droits. Ce sont donc :

- des ouvrages du domaine public,
- ou des ouvrages dont l'auteur est la personne qui le met en ligne et dont il cède les droits,
- ou des ouvrages sous licence « *creative commons* ».

Nous reviendrons sur les problèmes de droit qui entourent les ouvrages numériques dans la partie III et plus précisément dans le chapitre 8 consacré aux Logiciels Libres, où nous détaillerons le contenu et les enjeux de *creative commons*. A ce stade de notre étude, nous pouvons toutefois faire un parallèle entre ce mode de diffusion des livres électroniques et celui des « hommes-livres » du roman de Ray Bradbury *Fahrenheit 451*. Comme dans ce récit, chaque personne peut choisir un ouvrage (la seule restriction est qu'il soit du domaine public pour éviter les problèmes de droits d'auteur) et décider de le transmettre aux générations futures. Pour cela, au lieu de le mémoriser il faut le numériser c'est-à-dire le scanner, le transcrire et l'ajouter à la liste des ouvrages disponibles en ligne. Le plus ancien et le plus ambitieux de ces projets est né aux États-Unis en 1971, il s'agit du projet « Gutenberg » que nous verrons dans la section B. **Les grands projets de numérisation de texte** suivante.

Un autre possibilité de mettre en route un chantier basé sur le travail collaboratif à travers Internet pour constituer un corpus de textes électroniques est de demander aux internautes de mettre en ligne leur propre production. C'est ce que propose d'une certaine façon Wikipédia, en proposant à n'importe quel visiteur d'écrire ses propres articles dans cette encyclopédie qui se trouve librement accessible en ligne à l'adresse <http://www.wikipedia.org>. Nous avons déjà mentionné cette encyclopédie précédemment lorsque nous

avons eu besoin de faire référence à certains de ses articles. Nous reviendrons plus amplement sur ce projet dans la section **3. - Wikipédia**, page 78 du chapitre 4 consacrée à l'hypertexte, mais nous pouvons déjà relever qu'il paraît très ambitieux. C'est en particulier l'absence apparente de tout contrôle officiel, de comité de rédaction et de revue systématique des articles mis en ligne qui peut se faire poser des questions quand à la validité des articles par rapport à une encyclopédie traditionnelle. Nous verrons que le contrôle est en fait le fruit du travail collaboratif, la logique en ayant été poussé très loin, sans doute au maximum de ce qu'il est possible de faire dans ce domaine. Le succès de Wikipédia a poussé ce projet à essaimer le même concept dans d'autres domaines et il existe aussi d'autres projets, en particulier un dont l'objectif est la mise en ligne d'ouvrages pédagogiques, écrits sur ce même modèle collaboratif par des internautes motivés. Il s'agit de WikiLivres<sup>54</sup> qui annonce ses objectifs ainsi sur sa page d'accueil :

*« Wikilivres (en anglais Wikibooks) s'est donné pour but de mettre gratuitement à la disposition de tous, des «textes pédagogiques» au contenu libre. Démarré le 10 juillet 2003, le projet compte à l'heure actuelle<sup>55</sup> 705 sections dans plusieurs dizaines de livres, dont certains demandent encore beaucoup de travail. Tous peuvent être améliorés grâce à vos contributions. »*

## **B. Les grands projets de numérisation de textes**

Avec Internet, le besoin de la mise à disposition de grands corpus de textes sous format électronique s'est fait plus pressant et de grands projets de numérisation ont démarré dans plusieurs pays. Un premier recensement a

---

54. Consultable à l'adresse <http://fr.wikibooks.org/>

55. mars 2005

été fait en 1999 dans (Jacqueson et River, 1999, pages 171 à 184). Nous l'avons repris ici et mis à jour, bien qu'il soit pratiquement impossible de faire une liste exhaustive de ces sites. Ce travail est résumé dans une table en annexe B qui permet d'avoir une vue d'ensemble de la taille et du contenu de ces bibliothèques électroniques.

### 1. Les projets Français

**ABU** <sup>56</sup> L'Association des Bibliophiles Universels propose le libre accès au texte intégral d'œuvres du domaine public francophone sur Internet depuis 1993. Ces textes sont produits et diffusés par les membres bénévoles de l'Association des Bibliophiles Universels.

**CNAM** <sup>57</sup> Le Conservatoire numérique des Arts & Métiers est une bibliothèque numérique consacrée à l'histoire des sciences et des techniques, constituée à partir du fonds ancien de la bibliothèque du CNAM.

**Lamarck** <sup>58</sup> Le site Lamarck est consacré à faire mieux connaître la vie et les œuvres de l'un des plus illustres naturalistes français. Il retrace les principales étapes de la vie et de la carrière du savant, avec des liens vers les documents qu'il est possible de consulter sur le site, dont certains sont peu connus ou inédits : témoignages de contemporains de Lamarck, documents d'archive, historiographie.

Une rubrique « œuvres de Lamarck » permet d'accéder directement aux ouvrages de Lamarck en texte intégral : 7200 pages (la presque totalité de ses œuvres théoriques) ont été numérisées et peuvent être téléchargées ou consultées sur ce site. Les manuscrits et l'herbier invitent à découvrir deux importantes collections conservées au Muséum national d'histoire

---

56. <http://abu.cnam.fr/>

57. <http://cnum.cnam.fr/>

58. <http://www.lamarck.net/>

naturelle de Paris, au fur et à mesure de l'entreprise de numérisation de ces collections, qui a débuté en février 2003.

**Bibliothèque Nationale de France (BNF)** <sup>59</sup> La Bibliothèque Nationale de France a, dès sa conception, intégré le fait qu'elle devrait proposer des collections numériques en plus des ouvrages qu'une bibliothèque met traditionnellement à disposition du public et des chercheurs. En revanche, pour plusieurs raisons plus ou moins liées, les possibilités de consultation à distance n'ont pas été particulièrement développées dès le départ. Nous pouvons même aller jusqu'à dire que cet accès est resté longtemps bridé par rapport à d'autres projets du même type et de cette ampleur. La première raison est sans doute historique, nous pouvons en effet noter que le projet a démarré avant qu'Internet ne soit devenu un outil omniprésent comme nous le connaissons aujourd'hui et surtout avant que le haut débit n'ait fait son apparition dans le grand public. Quand le site de la BNF a ouvert ses portes au printemps 1997, il faisait partie des précurseurs, mais il n'avait pas pris en compte l'évolution très rapide qu'allait connaître Internet. Pour les concepteur du projet, il paraissait inutile de permettre la consultation à distance de documents numériques dont la taille impliquait des temps de transfert rédhibitoires. Il aurait fallu intégrer la vitesse des progrès dans ce domaine pour prévoir que ce qui n'était pas alors possible techniquement deviendrait rapidement la norme. De plus, la BNF a pu négocier la numérisation de certains ouvrages qui ne sont pas encore dans le domaine public, sous condition que leur consultation ne pourrait se faire que sur place, pour éviter que les copies prolifèrent en dehors de tout contrôle à travers Internet. Finalement, nous pouvons aussi remarquer que la BNF a voulu

---

59. <http://www.bnf.fr/>

prendre l'initiative de fournir à ses lecteurs un environnement adapté à la lecture des ouvrages numériques. Il faut encore une fois se rappeler que quand ce projet est né le contexte du livre numérique en était vraiment à ses tout débuts, les dispositifs de lecture n'existaient pas ou peu et le e-book était un prototype de laboratoires. Dans ces conditions, la BNF ne pouvait se résoudre à mettre à disposition des lecteurs de simples fichiers de texte ou d'images sans un dispositif pour aider à la navigation et à la consultation. Dans ce contexte, il est donc paru nécessaire de mettre au point un *Poste de Lecture Assistée par Ordinateur* (PLAO) pour permettre la lecture des ouvrages numériques, ce qui au final a interdit l'accès aux ouvrages à travers un simple navigateur Web, donc à travers Internet. Avec l'utilisation de ce dispositif, une particularité très importante de l'accès aux ouvrages numériques de la BNF était qu'il ne pouvait avoir lieu que sur le site même de la bibliothèque pour une grande partie.

La BNF ayant fait l'objet de nombreuses études, nous pouvons retrouver de nombreux chiffres sur la teneur de son contenu et aussi son évolution. Nous trouvons ainsi par exemple les objectifs affichés au départ :

*« Cent mille documents sont numérisés par la BNF. À ce nombre s'ajoutent les éditions numériques acquises auprès des éditeurs, telles des ouvrages sur CDROM. La BNF prévoit un accroissement annuel de la collection numérisée de 25 000 documents par an, à partir de l'ouverture. (Bouchard, 1995, page 15) »*

Ces objectifs vont évoluer rapidement, notamment au niveau de l'accès à distance qui suivait l'évolution générale d'Internet.

« Dès la fin de l'été 1998, la Bibliothèque nationale de France (BNF) offrira à ses lecteurs universitaires une ambitieuse collection de 100 000 documents « numérisés ». Ceux-ci seront consultables sur des postes de lecture assistée par ordinateur (PLAO)<sup>60</sup>, développés par la Bibliothèque elle-même à l'intention des lecteurs savants. En principe, ils seront aussi diffusables à distance : les initiateurs du projet semblent avoir anticipé le succès public d'Internet.[...] L'objectif central était la transmission à distance des documents ; au finish, la collection numérisée risque de rester prisonnière des murs de la Bibliothèque parisienne et des fameux PLAO. Pourquoi et comment est survenu ce « détournement » majeur ? L'ambition du projet élaboré dans une pure logique d'offre semble victime d'elle-même, ayant fait l'impasse sur un ensemble de considérations socio-techniques. (Bouchard, Octobre 1996, page 1) »

Si nous avançons dans le temps et que nous utilisons ces prévisions pour nous livrer à un petit calcul, nous devrions trouver en 2002, soit quatre ans plus tard :

$$100.000 + 4 \times 25.000 = 200.000$$

Même en ne considérant simplement que le seul mode image, cet objectif est atteint et même dépassé :

« Fin 2002, la Bibliothèque numérique est composée à ce jour de 75 000 monographies et 20 000 fascicules de périodiques numérisés en mode image (fac-similé), 2 600 documents en mode texte et 250 000 images fixes.

---

60. Pour une description du PLAO, voir l'article de Jacques VIRBEL, « La lecture assistée par ordinateur et la station de lecture de la Bibliothèque de France », in A. VUILLEMIN, *Les banques de données littéraires*, Paris, Pulim, 1993, pp. 19-27.

*Conçue comme une collection patrimoniale et encyclopédique, elle est constituée d'un large corpus de monographies et de périodiques qui, de l'Antiquité à la veille de la première guerre mondiale, propose des textes d'auteurs classiques mais aussi d'auteurs moins connus, des dictionnaires ainsi que des outils bibliographiques et critiques. Certains documents peuvent y figurer dans plusieurs éditions permettant d'en diversifier les niveaux d'accès.*

*Cette bibliothèque numérique est donc aujourd'hui toujours accessible sur place et à distance :*

*Sur place, dans les salles de lecture sur des postes de consultation dédiés, sont offerts aux lecteurs les documents libres de droits mais aussi les documents pour lesquels les ayants droit ont donné une autorisation de communication interne. Ainsi 250.000 images fixes sont disponibles en salles audiovisuelles alors que dans les différentes salles de la bibliothèque sont offerts 55.000 monographies et périodiques sur les postes dédiés à la consultation du fonds numérique. (Van Cuyck, 2003, pages 5-6) »*

Le dispositif de PLAO a finalement été abandonné, avec l'accroissement et l'amélioration des possibilités des accès à distance, il devenait indispensable d'uniformiser les modalités d'accès à la fois locales et distantes. Un site spécifique, le serveur Gallica, présenté dans la section suivante a été développé pour présenter la partie libre de droit des collections de la BNF, soit plus de 50.000 monographies et périodiques et 80.000 images. La Bibliothèque Nationale de France, garde aussi un site de présentation, qui permet de fédérer les autres projets plus spécifiques comme Gallica.

**Gallica** <sup>61</sup> L'appellation « Gallica » depuis 1997 désigne l'offre numérique de la BnF destinée au public distant. Au 1er janvier 2004, Gallica offrait sur la Toile :

- 70.000 volumes imprimés en mode image,
- 1.200 volumes imprimés en mode texte,
- 500 documents sonores,
- 80.000 images fixes.

Gallica est un instrument de diffusion du patrimoine écrit et iconographique.

En conséquence, on proposera au public des corpus à thème articulés autour d'une ou plusieurs collections réparties dans plusieurs départements, avec des accès documentaires adaptés. Il faut pouvoir travailler sur plusieurs centaines de documents dans une thématique et donc proposer des ensembles fournis sur un auteur, un courant de pensée ou un sujet donné.

Il ne s'agit pas, là encore, de mettre à disposition des collections dans leur totalité. Gallica n'a pas de prétention à l'exhaustivité.

**Frantext** <sup>62</sup> Frantext est à la fois un corpus de textes littéraires et un outil de navigation à travers ce corpus, dont l'accès est réservé aux abonnés. Voici la description qui en est donnée sur le site de l'ATILF :

*« L'outil de consultation de ressources informatisées sur la langue française (1992) peut être définie comme l'association d'une part d'un vaste corpus de textes littéraires français, et d'autre part d'un logiciel offrant une interface Web avec des possibilités d'interrogation de consultation et d'hyper-navigation. Historiquement, le*

---

61. <http://gallica.bnf.fr/>

62. <http://www.atilf.fr/frantext.htm>

but premier de ce corpus textuel était de permettre la constitution d'une base d'exemples destinée aux rédacteurs des articles du TLF.

L'outil de consultation Frantext contient 3737 textes appartenant aux domaines des sciences, des arts, de la littérature, des techniques, qui couvrent 5 siècles de littérature (du XVI<sup>e</sup> au XX<sup>e</sup> siècle). Il est accessible sur Internet, moyennant un abonnement.

Deux versions de Frantext sont proposées :

- L'intégralité de la base (3737 textes, environ 210 millions d'occurrences, environ un millier d'auteurs). Les œuvres se répartissent pour 80% d'œuvres littéraires et 20% d'œuvres scientifiques ou techniques. Il est possible d'effectuer des recherches à différents niveaux : simples ou complexes.

- Une sous-partie constituée de 1940 œuvres en prose des XIX<sup>e</sup> et XX<sup>e</sup> siècles, soit environ 127 millions d'occurrences, qui ont fait l'objet d'un codage grammatical selon les Parties du Discours. Aux fonctionnalités du Frantext intégral, ont été ajoutées des possibilités de requêtes portant sur les codes grammaticaux. »

**TLFi** <sup>63</sup> Le Trésor de la Langue Française Informatisé est la version informatisé d'un dictionnaire des XIX<sup>e</sup> et XX<sup>e</sup> siècles en 16 volumes et 1 supplément.

Il est en accès libre et propose :

- 100 000 mots avec leur histoire,
- 270 000 définitions,
- 430 000 exemples,
- 350 millions de caractères.

---

63. <http://www.atilf.fr/tlfi.htm>

**La bibliothèque électronique de Lisieux** <sup>64</sup> La bibliothèque de Lisieux a rendu une partie du fonds normand disponible en version électronique, ainsi qu'une riche documentation principalement consacrée aux auteurs du XIX<sup>e</sup> siècle. Le site permet aussi l'accès à travers des moteurs de recherche : l'incontournable Google, mais aussi un moteur spécialisé ALEPH-LITTERATURE<sup>65</sup>. La BM Lisieux donne aussi accès à LexoTor<sup>66</sup> qui a été créé à Toronto sur le site de Langue du XIXe siècle de Jacques-Philippe Saint-Gérard (Université Blaise Pascal Clermont-Ferrand II) et Russon Wooldridge (University of Toronto).

**Persée** <sup>67</sup> Les revues françaises en sciences humaines et sociales constituent un riche patrimoine scientifique dont la valorisation représente un enjeu stratégique pour la communauté des chercheurs, soucieuse d'une meilleure diffusion de leurs publications.

De ce constat et des attentes manifestes des chercheurs en sciences humaines et sociales est né le projet PERSEE avec un objectif triple :

- préserver et valoriser ce patrimoine
- participer au rayonnement de la recherche en langue française par la diffusion d'un vaste corpus d'articles intégré dans l'offre scientifique internationale
- permettre une exploitation enrichie de ces revues en termes de recherche et de consultation

---

64. <http://www.bmlisieux.com/>

65. Moteur spécialisé qui n'indexe que des sites spécialisé, il était disponible en ligne à l'adresse <http://aleph.ens.fr/> et réalisé par les équipes bénévoles de [fabula.org](http://fabula.org), de [revues.org](http://revues.org) et de marges linguistiques. Il est fermé depuis fin 2004 sans indication sur une prochaine réouverture.

66. Visible en ligne à l'adresse <http://www.chass.utoronto.ca/epc/langueXIX/lexotor/>

67. <http://www.persee.org>

**Revel** <sup>68</sup> Site de mise en ligne de revues électroniques, Revel@nice est né sous l'impulsion du Conseil Scientifique de l'Université Nice-Sophia Antipolis. Ce projet novateur veut promouvoir la recherche universitaire niçoise par un affichage clair de sa recherche en Sciences Humaines et Sociales.

Outil d'édition électronique simple d'utilisation et site de publication en ligne destiné à toute la communauté des chercheurs, Revel@nice s'adresse aussi bien aux étudiants s'initiant à la recherche qu'aux universitaires étrangers. Revel@nice est un site conçu en deux volets :

- un site ouvert au public, sur lequel les internautes ont accès au contenu des revues,
- un espace privé d'administration grâce auquel l'éditeur effectue la mise en ligne de sa publication. Le site s'appuie en partie sur le logiciel Lodel développé par l'équipe de Lodel.org.

**Cyberthèses** <sup>69</sup> Ce site regroupe les références aux thèses soutenues dans les différents établissements participant à Cyberthèses et les liens hypertextes menant au texte intégral de chacune d'elles. Sont aussi disponibles gratuitement plusieurs études et rapports sur l'édition électronique des thèses, les ressources pédagogiques et techniques développées dans ce but.

**CoLisciences** <sup>70</sup> C'est l'objet de notre recherche, nous reviendrons bien entendu plus longuement pour en décrire les dispositifs dans la section 6.1.

## 2. Les projets étrangers

### Les projets anglophones

---

68. <http://revel.unice.fr/>

69. <http://www.cybertheses.org/>

70. <http://www.colisciences.net/>

**Le projet Gutenberg aux Etats Unis (1971)** <sup>71</sup> C'est la première réalisation d'un réservoir de textes numérisés. Les textes retenus doivent être susceptibles d'intéresser le plus grand nombre de lecteurs, c'est-à-dire une audience à 99% dans le grand public. Les textes sont donc d'intérêt très général, présentés en « ASCII brut », sans aucune fioriture.

**ETC** <sup>72</sup> L'Electronic Text Center de l'Université de Virginie (1992) Les textes sont balisés en SGML conformément à la TEI et gardent une trace de la source imprimée. Le format original est d'ailleurs scrupuleusement respecté. Quand les textes sont consultés sur le Web, la conversion stockage SGML vers affichage HTML est effectuée à la volée.

L'objectif premier est de servir les étudiants et chercheurs de l'Université, en particulier pour le stockage des thèses et mémoires.

**The Universal Library de l'université Carnegie Mellon** <sup>73</sup> L'objectif est de rendre disponible sur Internet toutes les œuvres de l'humanité, et comporte donc des œuvres provenant de tous les domaines du savoir. Les livres sont disponibles sous plusieurs formats : html, ASCII, PostScript, RTF et même GIF.

**Center for Electronic Texts in the Humanities, CETH** <sup>74</sup> Ce site héberge plusieurs projets de textes numérisés, généralement spécialisés dans les sciences humaines, de plus il recense de nombreux projets de bibliothèques virtuelles aux États-Unis de manière à servir de centre de ressources. Les textes sont proposés au format propriétaire DjVu, ce qui nécessite l'installation d'un logiciel spécifique pour pouvoir y accéder.

---

71. <http://promo.net/pg/>

72. <http://etext.lib.virginia.edu/>

73. <http://www.ulib.org/>

74. <http://www.ceth.rutgers.edu/>

**The Oxford Text Archive, OTA** <sup>75</sup> L'*Oxford Text Archive* (OTA), ou archive textuelle de Oxford, héberge les ouvrages de littérature, langues et linguistique de l'AHDS (Art and Humanities Data Service<sup>76</sup>). L'OTA travaille étroitement avec plusieurs membres de la communauté académique des Arts et Humanités pour rassembler, cataloguer et préserver des textes électroniques de grande qualité pour la recherche et l'enseignement. L'OTA distribue plus de 2000 ressources en plus de 20 langues différentes et travaille pour étendre son fonds<sup>77</sup>.

**Google** Bien que le projet n'ait pas encore abouti, en fait il vient tout juste d'être lancé à l'heure où ces lignes sont écrites, il est difficile de ne pas le mentionner. En effet, le géant de la recherche sur Internet a décidé de lancer une campagne de numérisation à une échelle qui n'avait encore jamais été imaginée, cela sur un fond de campagne médiatique qui a trouvé un écho assez négatif en Europe qui voit là une dangereuse hégémonie de la culture nord-Américaine. Nous reviendrons en détail sur ce projet et les réactions qu'il a suscité dans le chapitre 9 dans lequel nous analyserons *le processus de concrétisation d'une bibliothèque numérique*, ce qui sera l'occasion de réfléchir sur la genèse du projet de numérisation Google en particulier.

### Les projets francophones

---

75. <http://ota.ahds.ac.uk/>

76. Le Arts and Humanities Data Service (AHDS) est un projet soutenu par le Royaume-Uni pour aider à la découverte, création et préservation de ressources digitales et pour la recherche, l'enseignement et l'apprentissage dans les arts et humanités.

77. The Oxford Text Archive hosts AHDS Literature, Languages and Linguistics. The OTA works closely with members of the Arts and Humanities academic community to collect, catalogue, and preserve high-quality electronic texts for research and teaching. The OTA currently distributes more than 2000 resources in over 20 different languages, and is actively working to extend its catalogue of holdings.

**Athena** <sup>78</sup> Athena est un projet Suisse de bibliothèque électronique qui rassemble des textes et des liens vers d'autres sites mettant à disposition des textes électroniques. C'est à la fois un projet de collection et un projet d'indexation. Les ouvrages sont classés en trois catégories :

- Littérature.
- Sciences.
- Arts.

**Grand Dictionnaire Universel** <sup>79</sup> La Base GDU a pour mission, d'une part de servir de vitrine analytique au projet d'informatisation du Grand Dictionnaire universel de Pierre Larousse, et d'autre part d'offrir aux chercheurs et aux curieux, d'une part la Préface, pièce-clé de l'histoire de la lexicographie française (elle est donnée ici en modes image, lecture et base de données), et d'autre part une sélection d'articles du GDU (en modes lecture et base de données).

**ARTFL** <sup>80</sup> Le projet *American and French Research on the Treasury of the French Language* (ARTFL) combine les efforts de l'ATILF (*Analyse et de Traitement Informatique de la Langue Française*), du CNRS, du département des Sciences Humaines, Sociales et de l'Electronic Text Services (ETS) de l'Université de Chicago. De nombreux textes sont proposés en ligne, dans un assez grand nombre de langues, dont le Français, qui vont du Grec ancien à la langue Apache. Il est entre autres possible de trouver plusieurs textes d'auteurs féminins français, de poésie Provençale, ainsi que l'Encyclopédie de Diderot et D'Alembert. Dernière remarque, une partie du corpus n'est pas librement accessible au public mais réservée

---

78. Athena est visible à plusieurs adresses, par exemple <http://un2sg4.unige.ch/athena/html/athome.html> ou <http://un2sg4.unige.ch/athena/html/francaut.html>

79. <http://www.chass.utoronto.ca/epc/langueXIX/gdu/>

80. <http://www.lib.uchicago.edu/efts/ARTFL/>

aux seules institutions qui participent financièrement au projet (266 sont recensées en ligne).

**Corpus de la littérature narrative du moyen âge au xx<sup>e</sup> siècle** <sup>81</sup> Le Corpus de la littérature narrative est le premier cédérom de B.A.S.I.L.E. (Base internationale de Littérature électronique). Coproduit par les éditions Champion et le CNED - Ministère de l'éducation Nationale, le Corpus de la littérature narrative met à la disposition des Institutions (bibliothèques, établissements scolaires et universitaires, centres de recherches, etc.) une bibliothèque patrimoniale de 1 000 œuvres narratives (romans, contes, nouvelles). Ce corpus est géré par un logiciel simple et puissant, étudié pour la gestion des textes littéraires. Il est à la fois une bibliothèque virtuelle de base et spécialisée qui couvre, en particulier, tous les programmes de français des collèges, des Lycées et des Universités (examens et concours). Le spécialiste et le chercheur y trouveront de leur côté des textes référencés et édités selon les règles de l'édition scientifique.

**Wikibooks** <sup>82</sup> Nous avons déjà présenté ce projet dans le cadre des écritures hypertextuelles en mode collaboratif. Wikilivres est la partie francophone du projet Wikibooks qui propose de mettre gratuitement à la disposition de tous des « textes pédagogiques » au contenu libre. La particularité de ce site est de demander aux internautes qui le visitent de participer à l'écriture des ouvrages. Il ne s'agit donc pas de numériser d'anciens ouvrages mais plutôt d'en écrire d'autres.

---

81. <http://www.burioni.it/news/novita/cdbasile.htm>

82. <http://fr.wikibooks.org/wiki/Accueil>

CHAPITRE 4

**RÉFLEXIONS SUR L'HYPertexte**

---

## Sommaire

---

<b>4.1</b>	<b>Introduction à l'hypertexte . . . . .</b>	<b>165</b>
A.	Historique . . . . .	166
B.	Définition . . . . .	170
<b>4.2</b>	<b>Fondements théoriques de l'hypertexte . . . . .</b>	<b>176</b>
A.	Modèles théoriques hypertextuels . . . . .	177
<b>4.3</b>	<b>Les notions autour de l'hypertexte . . . . .</b>	<b>180</b>
A.	Interactivité . . . . .	180
B.	Hypertexte et connaissances . . . . .	183
C.	Hypertexte collaboratif . . . . .	187
D.	Les systèmes hypertextes . . . . .	199
<b>4.4</b>	<b>La Gestion électronique de Documents . . . . .</b>	<b>204</b>
A.	Définition . . . . .	204
B.	Le traitement du document . . . . .	204
C.	Les normes de balisage . . . . .	213
D.	Les DTD de l'information-documentation . . . . .	216

---

## 4.1 Introduction à l'hypertexte

L'hypertexte est une notion qui demande d'être précisée en fonction du contexte dans lequel il est mis en œuvre. Même si au départ nous trouverons une définition sur laquelle la plupart des spécialistes s'accorderont, en pratique nous trouverons plusieurs manières de décrire un système hypertextuel et certaines ne sont pas toujours compatibles. Pour exemple de cette difficulté à se mettre d'accord sur un système hypertextuel opérationnel il suffit de rappeler que d'une part certains voient dans le Web la réalisation de la vision de Ted Nelson, une bibliothèque des bibliothèques quand lui même affirme que le Web est la plus mauvaise implémentation d'un système hypertexte. Il faut aussi se rappeler que l'hypertexte est très intimement lié aux technologies qui l'animent, ces technologies étant en évolution permanente, les systèmes hypertextes évoluent eux-aussi très rapidement.

Par ailleurs, le Web est aux yeux de beaucoup l'interface hypertextuelle idéale, souvent du fait que c'est la seule représentation qu'ils ont de l'hypertexte et notre système a été mis en place sur le Web. Il nous paraît donc important de faire un mise au point sur ce choix, le Web n'a pas été retenu principalement pour son interface hypertexte mais pour la possibilité de mise en réseau de documents à l'échelle mondiale. Comme Ted Nelson nous ne pensons pas que le Web, tel qu'il est mis en œuvre avec les outils traditionnels (pages html pointant les unes sur les autres), soit un système hypertexte complet. Nous présenterons donc, après notre vision de l'hypertexte comment celle-ci a pu être mise en œuvre sur le Web et quelles sont les concessions que nous avons dû faire pour que la technologie réponde à notre modèle, qui est une évolution d'un modèle à 3 modules, comme nous l'avons présenté dans la section **1. Hypertexte**, page 5 du chapitre 1 d'introduction.

## A. Historique

### 1. Un mot

Si nous nous en tenons à l'invention du mot, l'hypertexte daterait de 1965 où il est apparu pour la première fois dans un article de Theodor Holm Nelson écrit pour *Dream Machines*. Le projet qui anime Ted Nelson depuis lors est la construction de Xanadu<sup>83</sup>, une bibliothèque universelle dans laquelle n'importe qui pourrait publier et consulter les documents de chacun en payant une somme modique qui rétribuerait l'auteur. Mais surtout, puisque c'est l'hypertexte notre sujet, tous les ouvrages seraient connectés entre eux par un système dont les prétentions vont bien au delà du Web actuel : gestion des versions, pas de liens brisés, gestion des droits, etc... Cela fait bientôt quarante ans que le projet est actif sans qu'une réelle réalisation permette d'en prouver la faisabilité, mises à part la maquette d'un système hypertexte (CosmicBook) dans lequel les liens entre documents sont visibles sur l'écran sous la forme d'un trait épais qui va d'une fenêtre à l'autre. C'est ce qui fait dire à certains que Xanadu est le plus long projet de *vaporware* de l'histoire de l'informatique (Wolf, 1995) et alimente une certaine controverse autour du personnage de Nelson, qu'il mentionne et pour laquelle il s'explique sur son site<sup>84</sup>.

### 2. Un concept

Mais ce concept n'est pas nouveau, nous pouvons le trouver évoqué de manière plus ou moins explicite dans plusieurs travaux ou ouvrages. La

---

83. Bien que se défendant d'avoir voulu créer un système tel que le Web, Nelson n'a pas pu faire l'impasse d'un site pour expliquer ses travaux et essayer de proposer une maquette adaptable au modèle du Web, celui-ci se trouve à l'adresse : <http://xanadu.com/>

84. Lire pour cela sa page « What they say » <http://ted.hyperland.com/whatsay/> ce qu'ils disent à propos de Xanadu. Il explique entre autres que la controverse a démarré avec cet article de Wired.

paternité du concept est très unanimement attribuée à Vanavar Bush, qui a décrit dans un article prophétique « *As we may think* » paru dans le numéro 176 de la revue *Atlantic Monthly* en juillet 1945 (Bush, 1945) un appareillage électromécanique « **Memex** » (Memory Extender), présenté comme

*« sorte de « bibliothèque personnelle mécanisée », dans laquelle une personne stocke tous ses livres, ses documents et ses informations. Cedispositif fonctionne par des mécanismes d'associations plutôt que des formes d'indexation artificielles, numériques, alphabétiques ou hiérarchiques, de manière à être consulté à la fois rapidement et souplement. Conçu pour un utilisateur unique, il constitue une sorte d'agrandissement intime de sa propre mémoire. Pour Bush, ce mécanisme de liaison par associations est, bien qu'il n'en soit pas une copie conforme, analogue au fonctionnement de la pensée humaine et son objectif est de soutenir et d'améliorer les processus de pensée des humains, ceci en étendant leur mémoire. (Bruillard, 1997, page 201) »*

Cet appareillage n'a jamais été construit, il devait améliorer les performances cognitives du chercheur qui l'utilise en calquant son fonctionnement sur celui d'un cerveau. Pour la première fois quelqu'un imaginait qu'il serait possible de construire une machine qui travaillerait par associations d'idées plutôt que par simple indexation.

C'est ce même souci d'améliorer les performances intellectuelles qui a sous-tendu les travaux de Douglas Engelbart au sein de l'ARC (*Augmentation Research Center*) pour mettre au point son système NLS (*oN Line System*), qui mettait à disposition des chercheurs des possibilités de travail collaboratif. C'est notamment à lui que nous devons la souris comme système de pointage sur l'écran et le fenêtrage qui permet d'afficher plusieurs documents simultanément sur un même écran. Ce système, fonctionnel en 1968, peut-

être considéré comme la première implémentation d'un système hypertexte électronique.

Un autre aspect du système hypertextuel est sa capacité à gérer de grandes quantités d'informations, tout en les gardant cohérentes. Paul Otlet avec son *Traité de documentation* (1934) nous a apporté la première approche systématique de ce que nous appelons aujourd'hui les Sciences de l'Information. Avec Henri La Fontaine il est le co-fondateur de « l'Office International de Bibliographie » et du **Mundaneum** « l'Internet de papier » qu'il a commencé à imaginer dès 1918 (Otlet, 1989). Pour Guy Teasdale, historien d'Internet, Otlet

*« en est arrivé à construire un immense instrument de recherche hypertexte manuel et multimédia. Otlet consignait sur des fiches de grandeur standardisée des informations concernant des ouvrages ; ces informations peuvent être assimilées aux nœuds de l'hypertexte. Les liens étaient assurés par la Classification Décimale Universelle (Teasdale, 1995) »*

Le **Mundaneum** est un projet visant à réunir dans un même lieu toutes les connaissances du monde. Centre de documentation à caractère universel, il se voulait un lieu de rencontre de tous les hommes de bonne volonté et de toutes les disciplines du savoir humain. Le site <http://www.mundaneum.be/> apporte des informations sur ce projet qui n'a malheureusement pas su trouver les ressources nécessaires à son fonctionnement, ni même à sa réelle survie. Obligé de déménager plusieurs fois, les archives patiemment rassemblées par Paul Otlet et son équipe ont en grande partie disparu : perdues, détruites ou mises au pilon. Une partie des archives qui ont survécu sont depuis 1998 abritées au cœur de la ville de Mons, au 76 de la rue de Nimy.

Sous son appellation éponyme, le **Mundaneum** est aujourd'hui un centre d'archives de la Communauté française. Au-delà de la conservation des collections et de l'édition de publications, le **Mundaneum** organise régulièrement des événements (expositions, conférences, colloques...) et des formations.

### 3. Les ancêtres

Si nous continuons à chercher des systèmes hypertextuels non électroniques il faut bien entendu citer aussi l'Encyclopédie de Diderot et d'Alembert, ou des ouvrages moins connus comme le *Tractatus Logico-philosophicus* de Ludwig Wittgenstein (Vienne 1889 - Cambridge 1951), qui parut pour la première fois en 1921 dans la revue *Annalen der Naturphilosophie*. Ce texte n'a jamais cessé de susciter l'intérêt des philosophes, logiciens et littéraires, il se compose de cinq cent vingt trois propositions, toutes numérotées suivant un système dont Wittgenstein, lui-même donne la clef dans une note préliminaire :

« Les nombres décimaux assignés aux propositions individuelles indiquent le poids logique des propositions, l'importance qu'elles ont dans mon exposé. Les propositions  $n1$ ,  $n2$ ,  $n3$ , etc. sont des commentaires à la proposition  $n$ , les propositions  $n.m1$ ,  $n.m2$ , etc. des commentaires à la proposition  $n.m$  et ainsi de suite. (Wittgenstein, 1961) »

La numérotation des propositions peut constituer le point de départ de nombreux parcours de lecture, chaque lecteur étant invité à lire sa propre version du Traité. La rubrique 69 du Cahier Brun l'y invite :

« Une personne lit lorsqu'elle tire directement sa propre version du modèle qu'elle est en train de copier. (Wittgenstein, 1961) »

Il serait difficile de dresser une liste exhaustive de ce type d'ouvrage, notre propos étant principalement de montrer que le concept d'hypertexte

est depuis très longtemps sous-jacent dans l'écriture, c'est l'ordinateur et sa capacité à manipuler le texte pour nous qui a démultiplié la puissance du concept. Pour terminer sur ce rapide tour d'horizon de l'hypertexte et de sa genèse, il faut mentionner tout simplement la bibliothèque. C'est un endroit où les ouvrages ont été classés par associations d'idées, avec des index et des fiches bibliographiques pour pointer directement sur ceux-ci. Elle pourrait donc être interprétée comme un ancêtre du concept d'hypertexte.

## B. Définition

L'hypertexte pourrait se définir simplement comme un texte dont la lecture n'est pas prédéfinie de la première à la dernière page mais peut se faire par fragments de texte, ces fragments étant lus dans un ordre qui est entièrement sous le contrôle du lecteur. Dans ce cas, un dictionnaire ou une encyclopédie pourraient être considérés comme des hypertextes. En fait pour être plus précis nous pouvons reprendre la définition que nous donne (Bruillard, 1997), qui est elle-même reprise de (Shneiderman et Kearsley, 1989 ; Nielsen, 1990 ; Berk et Devlin, 1991) :

*« Un système hypertextuel se présente comme un dispositif informatisé permettant l'interconnexion de documents de divers types, ceci non sur la base d'un modèle hiérarchique ou relationnel, mais par des mécanismes associatifs sous contrôle de l'utilisateur.(Bruillard, 1997, page 198) »*

Cette définition fait ainsi référence à la navigation, c'est-à-dire au déplacement du lecteur à travers les nœuds d'un réseau de liens. Mais généralement, le système hypertextuel n'est pas véritablement compris ainsi (Lebrave, 1994) ; (Lebrave, 1997) ; (Baron et Bruillard, 2002). La définition qui en est faite tradi-

tionnellement tourne autour du texte numérique et du procédé informatique pour y accéder.

*« Un hypertexte est un ensemble de données textuelles numérisées sur un support électronique, et qui peuvent se lire de différentes manières. Les données sont réparties en éléments ou nœuds d'information - équivalents à des paragraphes. Mais ces éléments, au lieu d'être attachés les uns aux autres, comme les wagons d'un train, sont marqués par des liens sémantiques qui permettent de passer de l'un à l'autre lorsque l'utilisateur les active. Les liens sont physiquement « ancrés » à des zones, par exemple à un mot ou une phrase. (Laufer et Scavetta, 1992, page 3) »*

En revanche, quand on demande à des utilisateurs au moins occasionnels si ce n'est réguliers d'Internet et du Web, d'expliquer cette notion, ils en donnent rarement une définition cohérente. Le plus souvent, ils font comme les enfants, ils se placent mentalement dans une situation habituelle mais fictive et disent ce qu'ils font. En gros, « je clique et j'accède à un nouveau document » (Bruillard, 2002). Cette définition est si imprécise qu'elle donne une vision complètement erronée de ce qu'est un hypertexte. En effet si nous la rapprochons de l'utilisation quotidienne que nous avons de nos ordinateurs personnels, il est impossible de faire la différence entre une navigation hypertextuelle et l'utilisation banale d'un ordinateur au moyen d'une interface graphique. Par exemple, « je clique et j'ouvre mon courrier électronique », « je clique sur le sujet d'un courrier et je peux le lire ». Cette définition basée sur le « clic » de déclenchement du parcours se résume à une action dans une vision locale : des pages sont reliées entre elles et on se « déplace » de l'une à l'autre en cliquant sur des mots ou expressions mises en exergue ou des zones de l'écran « réactives ». La vision dominante n'est pas tant celle

d'un dispositif que d'une action locale (ponctuelle). Cela n'existe que dans l'ici et maintenant de la consultation, ce qui est soit l'explication soit la raison de la désorientation, parce qu'il n'y a pas création d'un parcours réfléchi mais simplement ouverture de « portes ». Ce phénomène a certainement de multiples causes, mais il est renforcé par la manière dont se présentent les systèmes d'exploitation des ordinateurs actuels et la généralisation de l'accès à des ressources et des programmes par simple clic sur des icônes. Pour un utilisateur, le fait de cliquer l'emporte sur les raisons qui l'ont amené à cliquer telle icône, tel mot ou telle notion.

*« La fonction hypertextuelle, très généralement, a comme tâche de contextualiser un signe textuel lato sensu à l'intérieur d'un champ de signes textuels qui, ensemble, constituent un système sémiotique [...]. Autrement dit, un signe sémiotique n'acquiert son sens qu'en fonction du champ de signes qui se positionne par rapport à lui ou encore par rapport auquel se positionne le signe textuel en question. C'est cela, en fait, l'intérêt, l'importance de l'hypertexte non pas au sens technologique, mais au sens sémiotique ou conceptuel. Au sens technologique, l'hypertextualité se résume en la possibilité de créer des liens [...] entre différentes ressources textuelles [...] et cela, selon un certain point de vue, un besoin, un objectif, une idée, etc. (Stockinger, 2001, page 70) »*

L'unicité de l'interface permet aux usagers de facilement savoir manipuler le dispositif dans ce qu'il a de plus simple : l'activation de lien, en revanche cela va lui masquer l'opportunité qui s'offre de créer des parcours à travers un corpus pour avancer dans une réflexion en même temps qu'il avance dans la lecture.

*« On en reste au niveau perceptif immédiat sans pouvoir développer le moindre langage, on a besoin de voir (s'aider de ce que montre*

*l'écran) pour refaire les actions faites, sans pouvoir se construire de représentation globale cohérente (Normand et Bruillard, 2001). Dans ce contexte, on peut rechercher des informations, on est loin d'avoir l'idée que l'hypertexte puisse aider à en développer de nouvelles. C'est une technologie d'accès, pas de structuration. (Bruillard, 2002) »*

Nelson n'est pas du tout satisfait de la façon dont le Web s'est approprié la notion d'hypertexte, en la déshabillant pour lui permettre une implémentation technique malgré de nombreuses contraintes. C'est cette simplification qui a rendu possible la création du Web, sans contraintes fortes, sans vérifications, choses que les chercheurs du Web essaient maintenant de réparer à grand renforts de normalisation, par exemple en introduisant XML pour séparer données et présentation. Ces préoccupations de normalisation sont devenues majeures parce que le rythme auquel le Web se développe nous fait perdre son contrôle. Autre exemple, il y a quelques années le portail de recherche et d'information Yahoo! faisait son apparition et sa renommée était basée sur le fait qu'il proposait un catalogue de liens triés par catégories. Les Webmasters proposaient leur site à Yahoo! dans une catégorie particulière, la cohérence de ce souhait était vérifiée et ensuite les internautes pouvaient consulter ce catalogue et accéder par exemple à la liste des sites proposant de la littérature française du XIX<sup>e</sup> siècle. Aujourd'hui, ce type de recherche se fait plutôt à travers un moteur de recherche du type Google, en lui fournissant des mots-clefs qui vont lui permettre de générer une liste de sites correspondants à ces mots-clefs. Il n'est plus question d'essayer de consulter un catalogue préparé à l'avance. Trop de site qui évoluent trop vite rendent ce type de liste impossible à gérer correctement, d'autre part l'arborescence du catalogue est devenue si complexe qu'il est devenu très complexe de naviguer à l'intérieur. C'est là que les propositions du Web Sémantique vont permettre de continuer à avancer.

Nous allons les présenter dans la section 2., page 101 du chapitre 5 suivant. Rappelons simplement ici qu'il s'agit entre autre de construire des ontologies qui vont permettre de mettre d'accord au niveau mondial sur des formes de description de ressources (Grandbastien, 31 janvier et 1er février 2002). L'enjeu est de taille, parce qu'il doit permettre de faire des recherches plus pertinentes et qu'il va permettre à des agents logiciels de faire ces recherches pour nous, mais cela tend à laisser de côté la question plus complexe des définitions de parcours d'interprétation humains, l'un des aspects régulièrement oubliés de l'hypertexte (Baron et Bruillard, 2002).

Pour un projet tel que *Co<sup>L</sup>i<sup>S</sup>ciences*, l'indexation n'est pas une fin en soi. Il est bien entendu important de pouvoir rechercher des paragraphes particuliers, mais c'est loin de suffire pour reconstituer la « pensée » d'un auteur comme cela est notre objectif. Si on s'accorde sur l'idée que l'hypertexte ne peut se limiter à faciliter l'accès à des ressources, il faut essayer d'identifier ce qui permet d'aller plus loin. Il faut penser en termes d'aide à la compréhension et à l'analyse autant qu'en termes d'aide à la navigation et multiplier les objectivations, les représentations (Ganascia, 2001). Pour cela, deux éléments sont à prendre en compte :

**Faciliter toutes les possibilités de visualisation globale du corpus** Sur ce premier point, donner des formes d'appréhension globale d'un corpus apparaît essentiel car cela permet d'y projeter des éclairages et de retrouver la perception qu'un lecteur a face à un ouvrage papier. Il le pèse, il le feuillette. Quand il le lit, il perçoit à travers l'épaisseur des pages lues et de celles non lues à quel moment il se situe dans son voyage à travers le texte. Pour cela il pourrait suffire d'ajouter des repères visuels tels que les ascenseurs qui habillent traditionnellement les côtés des fenêtres des interfaces graphiques de nos ordinateurs, mais rendre

uniquement cette perception est réducteur. Puisque les textes numérisés ouvrent l'accès du contenu à son contenant, la visualisation doit en bénéficier. Il s'agit : (i) d'aider à visualiser tous les paragraphes ayant une caractéristique donnée (traitant d'une notion, explicitant une relation, etc.), le spatial donnant à la fois le quantitatif et la localisation, avec une facilité d'accès au contenu ; (ii) de procurer des modes d'interrogation multiples : quelles sont les relations les plus nombreuses, les plus rares, etc.

**Aider à garder, représenter et organiser les parcours** Il s'agit d'essayer de traduire le dynamique (temporel) par du spatial. Sur ce second point, représenter un chemin sous la forme d'un objet manipulable est une piste essentielle à creuser. Conserver l'historique d'une consultation est une première étape, il faut également pouvoir organiser cet historique, y insérer des notes, etc. Il faut aussi pouvoir comparer deux parcours, les projeter sur la visualisation du corpus, les qualifier vis-à-vis des notions présentes et des relations retenues, etc. L'idée est de développer une forme d'«*exercice*» proprement hypertextuelle. Alors que dans le cursus scolaire, on passe beaucoup de temps à acquérir des formes de présentation écrite très spécifiques comme la dissertation ou le commentaire composé, il y a certainement d'autres formes nouvelles à développer.

L'informatisation introduit la possibilité de mettre en place d'autres moyens, plus graphiques (telles les cartes de concepts), dans des types d'hypertextualisation. Dans *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, cela a abouti à la réalisation d'un dispositif particulier que nous avons baptisé **lectochromie** et que nous verrons dans la section 7.4, page 151 du chapitre 7.

« Tout énoncé est porteur d'une "orientation" déterminée du fait d'une certaine mise en relation qu'il opère entre différents "repères" linguistiques renvoyant à des acteurs, des états, des processus, des situations, des domaines. Ce qui importe donc [...], c'est de travailler sur ces types de mises en relation (thématisation, prédication, modalités) grâce auxquels, à chaque fois, des auteurs vont "tisser" un jeu structuré de références repérant des "domaines" et des "significations" (Vignaux et Khadiyatoula, 1990) »

Un texte construit et se construit donc par l'établissement de certaines frontières établissant « l'intérieur » d'un domaine en regard d'un « extérieur », orienté ainsi à chaque fois, vers un certain point de vue cognitif. Ce travail sur des ontologies (des notions centrales, des « idées premières »), apparaît essentiel aujourd'hui, tant sur les questions de la représentation que sur celles de l'hypertextualité.

## 4.2 Fondements théoriques de l'hypertexte

Dans notre mémoire de DEA, nous avons fait un tour d'horizon de l'état de la théorie en commençant par la recherche d'une définition pratique de l'hypertexte comme nous venons de le faire. C'est en introduisant la notion de réseau, dont chaque élément est appelé un « nœud », ceux-ci étant liés entre eux par des relations que la dimension logique du modèle hypertextuel trouve une correspondance avec la réalité physique à laquelle le modèle doit se rattacher. Pour compléter ce modèle, nous devons aussi définir que chaque nœud est relié à un document (texte, image son, vidéo, ...) par une « ancre ». De cette façon, naviguer dans le système hypertextuel consiste à naviguer au milieu de ces nœuds et à travers eux en suivant les ancres. Cette disposition

se retrouve déjà dans le schéma d'un modèle « à deux couches » (cf. (Nanard, 1993), (Nanard, 1994) et (Simoni et Fluhr, 1997)) comme représenté sur la figure 4.1.

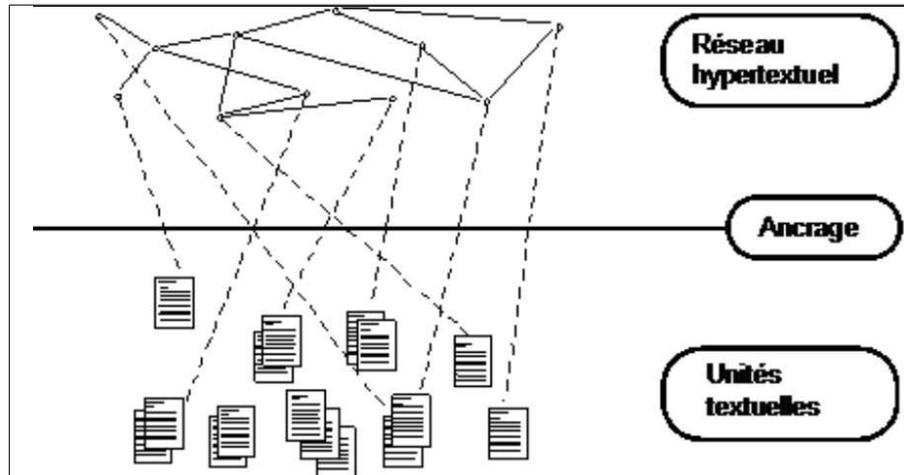


FIG. 4.1: L'ancrage dans le modèle à deux couches de Marc Nanard

## A. Modèles théoriques hypertextuels

### 1. Modèle de Marc Nanard

Pour Marc Nanard, un système hypertextuel s'articule en deux couches autour de trois concepts : les documents, la connaissance et la navigation. Dans ce modèle, le système hypertexte se présente comme la superposition par voie d'ancrage d'un ensemble de documents, cette structure dite faible constituant la première couche et d'une structure de connaissance, dite forte, constituant elle la deuxième couche Nanard (1993, 1994). La métaphore de la navigation étant utilisée pour exploiter cette structure. C'est ce modèle que nous avons déjà présenté figure 4.1 en introduisant la notion d'ancrage. En effet, c'est ce mécanisme qui permet d'associer aux documents de la structure faible les qualifications de la structure forte de connaissances. De cette façon, le lecteur peut parcourir et choisir les documents, non pas d'après leur propre

contenu mais d'après les indications de lecture apportées par la couche des connaissances.

## 2. Modèle de Dexter

Le modèle de Dexter (Halasz et Schwartz, 1990, 1994) tire son nom du *Dexter Inn* de l'état du New Hampshire où s'est tenu le séminaire qui a vu le lancement d'une série d'études et de confrontations qui ont conduit à la mise au point d'un modèle générique d'architecture hypertextuelle. La figure 4.2 donne une vue schématique des trois couches du modèle de Dexter :

- La couche d'exécution (*run-time layer*) au sommet décrit les mécanismes supportant les interactions de l'utilisateur dans l'hypertexte.
- La couche d'enregistrement (*storage layer*) en position médiane renferme la structure « statique » de l'hypertextualité, c'est-à-dire les mécanismes par lesquels les composants liés et non liés sont « collés » ensemble pour former le réseau hypertexte.
- La couche du contenu (*within-component layer*) à la base renferme les contenus spécifiques de chaque hypertexte. Tous types de documents peuvent être utilisés et la problématique de cette couche tient plus de la disponibilité physique de ces types que de l'hypertexte.

La figure 4.2 permet aussi de mettre en évidence comment les couches sont raccordées et peuvent interagir. La couche d'exécution et la couche d'enregistrement sont raccordées par des instructions de présentation (*presentation specification*), et la couche de contenu est rattachée à la couche d'enregistrement suivant des mécanismes d'ancrage (*anchoring*).

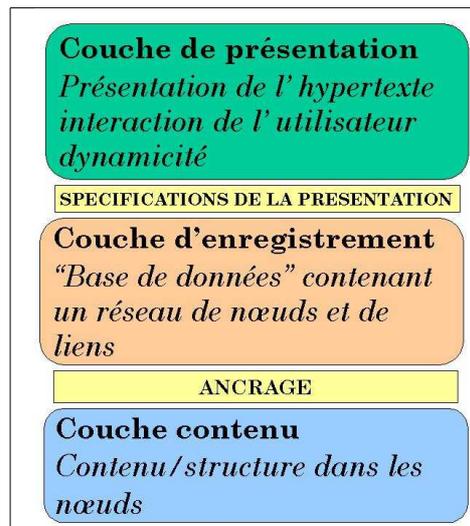


FIG. 4.2: Le modèle à trois couches de Dexter

### 3. Modèle à trois modules

Ce modèle à trois modules a été mis au point par David Piotrowski dans le cadre de l'équipe « *Hypertextes et textualité électronique* ». Nous l'avons déjà présenté dans l'introduction générale et nous avons signalé à cette occasion qu'il a servi de base pour la mise en œuvre du système hypertextuel *C<sup>o</sup>L<sup>i</sup>s<sup>c</sup>iences*. De façon très schématique ce modèle peut être présenté comme sur la figure 4.3, nous y retrouvons les trois couches du modèles de Dexter, mais cette fois les couches ne sont plus superposées mais mise en relation, ce qui permet de faire apparaître le lecteur et ses interactions avec le système hypertextuel.

En effet, dans le modèle de Dexter le lecteur se trouvait relégué tout en haut, au-dessus de la couche de présentation et ses seules interactions avec le système hypertextuel doivent être formalisées à travers celle-ci. Avec le modèle à trois modules, le lecteur peut s'intercaler entre la couche de présentation et celle de consultation, ce qui permet de mettre en évidence qu'il accède aux données à travers la couche consultation mais qu'elles sont

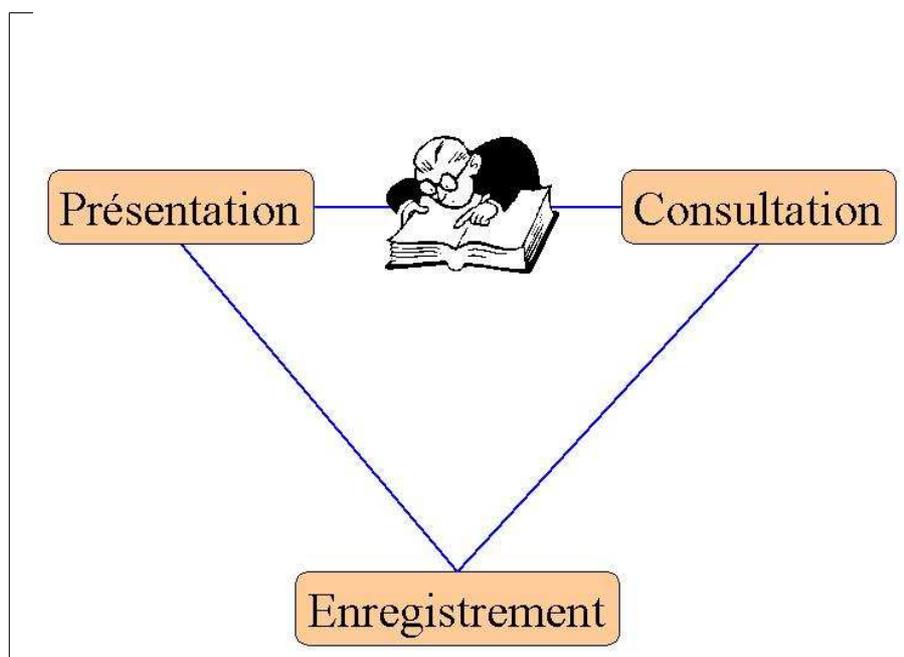


FIG. 4.3: Le modèle à trois modules

prises à sa disposition à travers la couche présentation. Ces interactions sont représentées sur la figure 4.4 qui permet de mettre en évidence les trois modules en interactions fonctionnelles à travers les trois doubles flèches rouges, mais surtout qui fait apparaître cadre de présentation (P) et grille de consultation (C), qui sont en interaction entre elles et avec une nouvelle entité nommée système de qualification (Q). Ce système, extérieur au système hypertextuel proprement dit, est un dispositif qui donne au lecteur des outils d'annotation raisonnée, qui vont l'assister dans sa navigation.

## 4.3 Les notions autour de l'hypertexte

### A. Interactivité

Une sculpture de bronze ou de plâtre, un monument aux morts sur une place d'un village français, un tableau naturaliste n'ont généralement rien

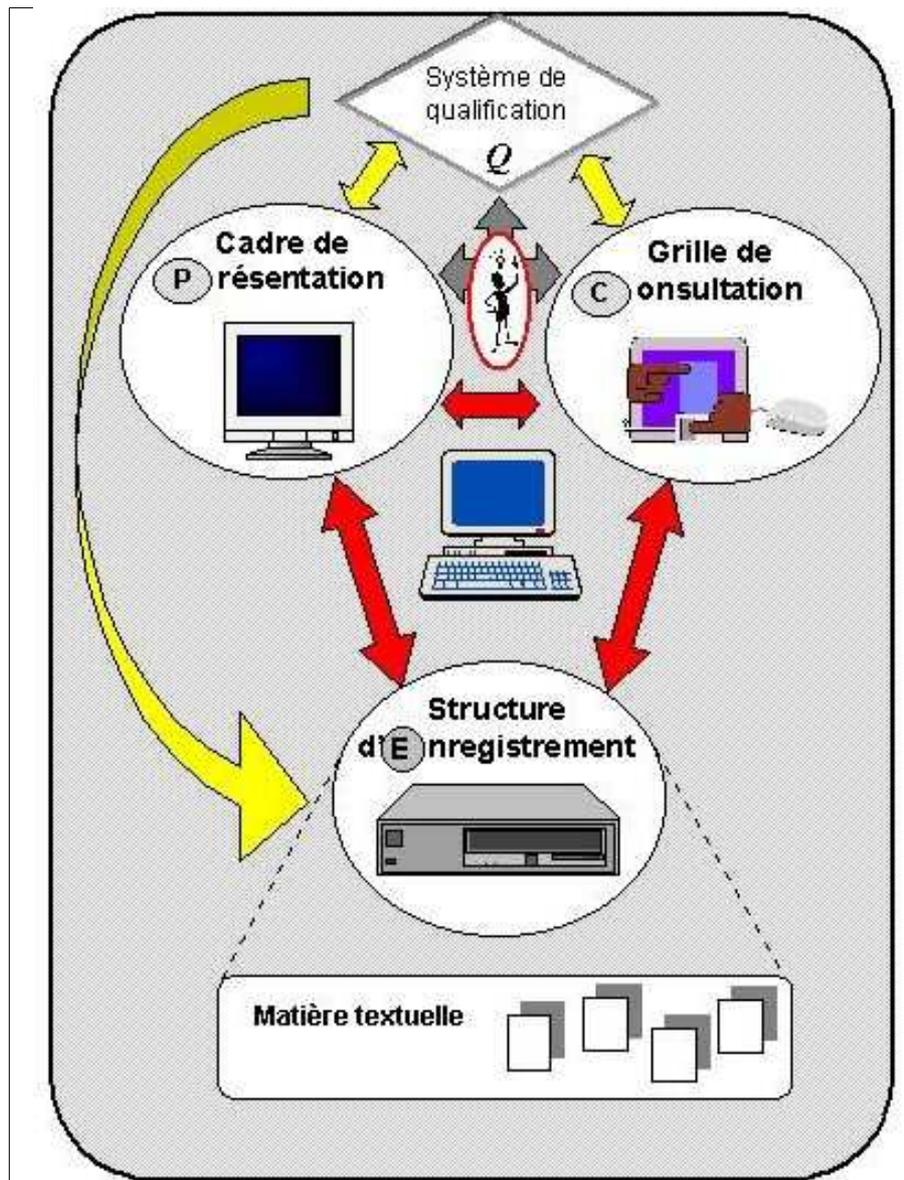


FIG. 4.4: Le modèle à trois modules en détail

d'interactif : l'observateur ne les modifie en aucune manière, il se contente d'y apposer son regard, sans y toucher aucunement... Le livre classique présente déjà une forme d'interactivité dans la mesure où il y a action réciproque du livre sur le lecteur dont l'esprit va être modifié au cours de la lecture, et du lecteur sur le livre, dont la configuration spatiale est modifiée lorsque les pages sont tournées. Toutefois, cette interactivité est limitée du fait que le support imprimé impose une linéarité primitive que le lecteur ne suit pas

forcément, comme c'est le cas dans un dictionnaire ou une documentation technique, mais qui structure l'espace du texte et le circonscrit... Avec ses *Mille milliards de poèmes* (Queneau, 1961), Raymond Queneau inaugure sur le papier ce qui fera, à notre avis, l'essence même de l'interactivité, à savoir que le nombre de parcours possibles est combinatoire et que chacun de ces parcours affecte le lecteur d'une façon différente...

Il pourrait être intéressant de noter que dans ce cas la version électronique n'apporte peut être pas un réel progrès ou même une réelle nouvelle dimension. En effet, sous format papier le lecteur peut être surpris par le contraste entre le faible nombre de pages et l'étendue de la combinatoire des parcours de lecture. Sous format électronique, ce contraste disparaît, l'ouvrage est réduit à une simple page qui s'affiche toujours différente sans que le lecteur ne puisse prendre conscience des subtilités qui sont à l'origine de cette page.

L'interactivité introduite par le support numérique n'apporte que peu de choses par rapport à la désorientation et la perte de référence qu'elle introduit. Paradoxe d'un ouvrage papier dont le système hypertextuel semble supérieur à l'ouvrage électronique. L'interactivité, littéralement, l'activité réciproque de l'œuvre sur celui qui la contemple signifie, en outre, que l'œuvre regardée va être renseignée sur le regard qui la lit, ou, plus exactement, sur les manipulations auxquelles elle est soumise, et qu'elle va se transformer au fil de ces manipulations pour offrir un nouveau visage. Deux versions de l'interactivité sont proposées concurremment de nos jours, une version faible et une version forte.

Selon la version faible, l'observateur se contente de construire son propre parcours parmi les mille milliards qui existent potentiellement et qui s'offrent à lui, mais l'œuvre reste intacte ; elle ne fait que se déguiser, se poudrer et se maquiller pour le plaisir de son lecteur sans changer de nature.

Selon la version forte, l'œuvre elle-même se transforme par les adjonctions des lecteurs ; elle devient, de ce fait même, œuvre collective. Dans la première version, on peut se demander ce qui, dans le kaléidoscope des possibles, permettrait de privilégier certains parcours où les épisodes seraient reliés entre eux par un fil narratif, où une intrigue se nouerait et tiendrait en haleine, où un sens d'ensemble serait à même de se constituer, où une liaison intelligente des épisodes répondrait à l'esprit qu'aurait voulu mettre le concepteur du livre. Rappelons à ce propos que lecture vient de *legere* qui signifie lier en latin, ce qui a donné *inter-legere* d'où l'on a tiré intelligence... Dans la seconde, les différents lecteur-auteurs doivent tous collaborer à une œuvre commune, sans que l'intrusion d'un imposteur malveillant ne condamne les autres. Il faut donc que des règles de coopération préviennent les débordements et facilitent les échanges.

## **B. L'hypertexte comme réseau de connaissances**

Nous avons déjà dit que le concept de l'hypertexte n'était pas nouveau. En particulier, nous pouvons voir dans l'index d'un livre qui rassemble un ensemble de termes auxquels sont associées les pages où ces termes sont apparus comme une première étape dans la création de liens entre différentes parties d'un texte. Cet index autorise le rapprochement des différents contextes d'apparition d'un même mot, ce qui est une opération classique dans la tradition herméneutique et qui permet d'affiner la définition qui pourrait être donnée de ce mot par la mise en évidence de l'usage effectif qui en est fait. La numérisation, avec l'utilisation de moteurs de recherche plein texte va permettre de mettre en place ce type de procédure pour tous les mots d'un corpus, sans qu'il n'y ait besoin d'un travail préalable de mise en forme ou de recherche. Les occurrences d'un mot dans le texte seront

rendues accessibles par l'intermédiaire d'un pointeur, ce que l'on appelle un lien hypertextuel, et qui est l'analogie exact d'un index à la différence près que le lien hypertextuel porte en lui la dynamique qui permet au lecteur d'aller automatiquement du mot vers ses occurrences, alors que l'index demande de feuilleter l'ouvrage. C'est l'informatique qui facilite et rend possible l'automatisation des procédures d'accès.

En fait, l'hypertexte va bien au delà, puis qu'il est possible de construire n'importe quel lien hypertextuel non seulement d'un mot vers ses différentes occurrences, mais aussi d'un mot vers sa définition, d'un mot vers sa traduction dans une autre langue, d'un mot vers ses synonymes, d'un mot vers sa transcription phonétique, les possibilités sont infinies.

Si nous ajoutons à cela la disponibilité sur la même plate forme informatique de documents de provenance diverses et enregistrés sur plusieurs types de médium, nous arrivons à l'hypermédia qui consiste à mettre en place le même type de liens mais plus seulement entre des textes mais entre des médium différents. Ainsi, l'hypermédia permet d'associer à une zone d'un schéma ou d'une image, des éléments textuels ou d'autres images, et à un mot, une photographie, un son, une vidéo...

Quand, en 1945, Vanevar Bush définit le principe de l'hypertexte, sans le nommer, c'est en fait un réseau de connaissances qu'il décrit :

*« L'esprit humain opère par association. On ne peut pas espérer dupliquer pleinement ce processus mental artificiellement [...] Mais il serait possible de dépasser la puissance et la permanence par le stockage de l'information aux fins de mieux stocker les informations et les mettre en interaction. (Bush, 1945) »*

L'hypertexte est donc d'abord, un système d'organisation et de classification des connaissances. Il modélise des fonctionnements qui tentent

de s'approcher de ceux du cerveau, tels les mécanismes associatifs de la mémoire ou l'emploi de métaphores visuelles pour accéder à un concept (principe de l'hyper-image), etc. Ce réseau de connaissances qui se veut structurellement différent du livre, s'apparente à une mémoire artificielle apte à entrer en interaction avec l'intelligence humaine. En effet, à la différence du texte traditionnel composé d'une suite séquentielle de paragraphes, les documents qui composent l'hypertexte ne sont pas reliés les uns aux autres de manière continue, mais visent à s'organiser selon un réseau hiérarchique cognitivement ordonné.

En raison des renvois multiples qui le constituent, l'hypertexte n'a d'une certaine façon, ni début ni fin. C'est une structure évolutive à laquelle peuvent être ajoutés selon les besoins, de nouveaux liens et documents. L'écriture hypertextuelle nécessite donc la décomposition du domaine de savoir considéré en unités d'information interdépendantes. Cette décomposition permet le développement de nouvelles mises en relation.

Nous avons rapidement posé les bases de l'hypertexte, sur lesquelles de nombreux ouvrages se sont déjà étendus, sans pour autant qu'ils soient allés jusqu'à la mise en place. Le mode de création et l'auteur ne sont bien entendu pas neutres, comme c'est le cas pour toutes les œuvres intellectuelles. En revanche ce qui conduit la plupart du temps à ignorer la place de l'auteur dans le cas d'un hypertexte, c'est sans doute que la médiation à travers un dispositif informatique semble « neutraliser », c'est-à-dire « rendre neutre », l'auteur. Comme si l'ordinateur était capable d'effacer les décisions, les choix et les idées de l'auteur, pour les remplacer par une opération de logique binaire, sans nuance.

Pour être plus précis et aller plus loin dans notre argumentation, il faut analyser cette propriété qui veut que les nœuds soient organisés suivant un

réseau hiérarchique cognitivement ordonné : comment laisser supposer que cette propriété n'est pas étroitement liée à l'auteur qui a créé ces liens, d'après ses propres réflexions et son jugement personnel. Bien que l'hypertexte soit mis en œuvre par une technologie informatique, son mode de création montre une faiblesse directement liée à son enjeu. Si un réseau a été construit par une seule personne, ou même un petit groupe d'individus, ce réseau pourrait n'être adapté uniquement qu'aux processus de pensée de cette seule personne ou ce petit groupe. Il reste le cas où l'hypertexte n'est pas directement le fruit d'un créateur, mais plutôt la mise en œuvre d'une opération computationnelle, dans laquelle les nœuds sont mis en relation de manière automatisée, en suivant des règles syntaxiques, sémantiques ou autres, établies au préalable par un programmeur. Dans ce cas la pertinence des liens dépend de l'efficacité des algorithmes utilisés pour les construire et là encore nous ne sommes pas du tout assurés que les capacités cognitives du lecteur seront en adéquation avec ceux-ci.

Ainsi, la constitution de données hypertextuelles autorise un processus de recherche constructiviste et émergent au cours duquel les hypothèses s'élaborent et se reconfigurent selon l'effet des informations rencontrées chemin faisant par l'internaute. L'hyperlien permet de la sorte, la co-construction d'un « scénario » sous la forme d'enchaînements grâce auxquels le lecteur construit sa propre logique de parcours.

Il n'est pas toujours possible d'anticiper les orientations de lecture qu'adopteront les différents utilisateurs. Selon le contexte, chacun peut décider de son cheminement dans la gamme des parcours possibles. En fonction des degrés de liberté qui lui sont accordés par la structure hypertexte, l'utilisateur peut élaborer son propre programme de lecture par concaténations d'un fragment à un autre, par ajouts et retraites de nœuds et de liens. C'est de cette manière que

le lecteur prend à son tour un statut d'auteur, mais statut bien éphémère étant donné que la plupart du temps ses interactions sont perdues et ne servent pas à participer à la construction de nouveaux parcours de lecture.

En résumé, l'hypertexte n'est pas seulement à lire, mais aussi à écrire. Il ne s'agit pas d'un moyen de connaissance au sens ordinaire, mais bien d'un système de métaconnaissances : on actualise les programmes conçus par d'autres pour construire ses propres parcours de sens, ses circulations dans des architectures cognitives, ses consultations de banques d'informations, etc. C'est ce que nous entendons observer et modéliser : recueil des parcours, analyse des types d'enchaînements selon les étapes, modes de légitimations et d'explications, argumentations.

## C. Hypertexte collaboratif

### 1. Définition de la collaboration

Dans le contexte professionnel, « collaboration » est un terme très répandu, bien qu'il soit pas toujours très bien défini. Il est aussi souvent fait référence à des « outils collaboratifs »<sup>85</sup>, voire même à une ou des technologies collaboratives. Avant d'aller plus loin, pour ne pas tomber dans le même travers, il nous faut préciser ce que nous entendons par collaboratif dans le contexte technologique en général et hypertextuel en particulier. Il est possible de décomposer le fait d'écrire à plusieurs un même document en 4 sous types :

« 1. Deux personnes ou plus écrivent conjointement le texte complet d'un document.

2. Deux personnes ou plus contribuent en apportant des composants à un document.

---

85. aussi connu sous leur nom anglais de *groupware*

3. Une personne ou plus modifie, ou relit, le document écrit par une ou plusieurs personnes ; et
4. Une personne travaille de façon interactive avec une personne ou plus pour faire l'ébauche d'un document basé sur les idées de la ou des personnes.

(Farkas, 1991, page 14) »

## 2. Écriture hypertextuelle : Wiki

Il existe de nombreuses manières de construire un hypertexte, plusieurs logiciels sont disponibles, certains s'adressent à un unique auteur, d'autres sont utilisables en équipe. Généralement tous ces outils fonctionnent suivant un schéma classique qui consiste à avoir le ou les auteurs travaillant à la constitution de l'hypertexte. Il est alors possible de faire toutes les écritures et modifications voulues. Puis l'hypertexte est déclaré « terminé » et mis en ligne. Il peut y avoir des versions créées successivement, les unes sur les autres, mais on retrouvera toujours les mêmes phases qui s'enchaînent et l'hypertexte mis en ligne « saute » de version en version. Il existe toutefois un système hypertextuel qui ne ressemble pas du tout à ce modèle : il permet la construction d'un site Web de manière collaborative et directement en ligne. L'hypertexte n'est jamais figé sur une version, toujours en construction, cela fait même partie de sa « philosophie ». Ce système s'appelle un Wiki, il a été inventé et mis au point la première fois par Ward Cunningham en 1995 pour le site du Portland Pattern Repository<sup>86</sup>. D'après Cunningham, « wiki wiki » signifie « vite » en Hawaïen, ce qui dans son esprit signifie qu'un **WikiWikiWeb** est conçu pour être édité vite et facilement. Il dispose d'une syntaxe de marquage très simple, dont le détail tient sur une simple page

---

<sup>86</sup>. Toujours visible en ligne à l'adresse : <http://c2.com/cgi/wiki?WelcomeVisitors>

Web, comme celle dont le début est visible figure 4.5 et il permet de mettre rapidement en ligne sur le Web des idées structurées sous la forme d'un hypertexte.

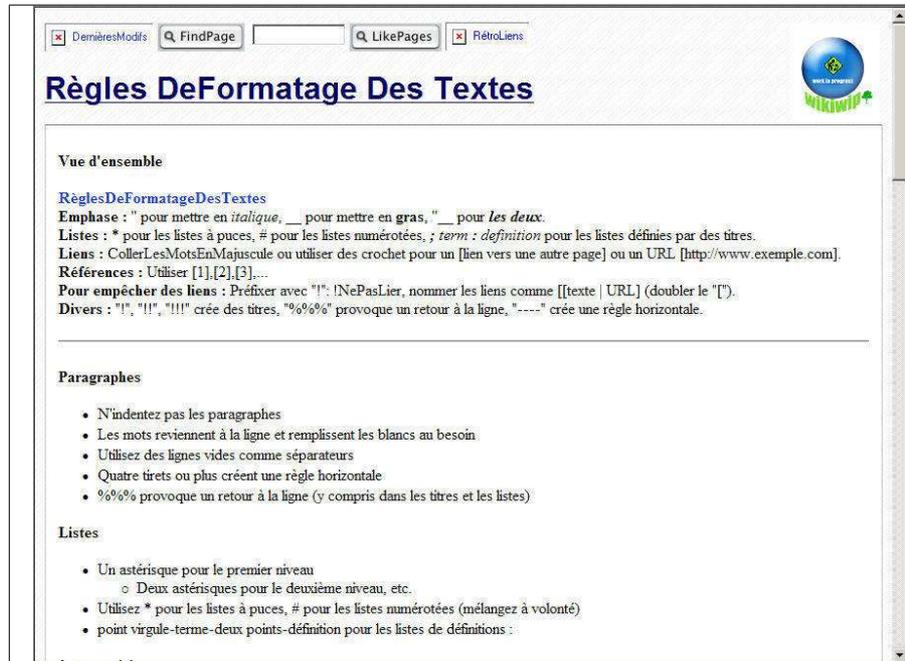


FIG. 4.5: La page RèglesDeFormatageDesTextes sur un Wiki

Ward Cunningham décrit les wikis comme

« *Freely expandable collections of interlinked Web 'pages', a hypertext system for storing and modifying information - a database, where each page is easily editable by any user with a forms-capable Web browser client.* <sup>87</sup> (Leuf et Cunningham, 2001) »

Le wiki gère automatiquement les versions de chaque page, à tout moment il est possible de reconstituer l'histoire de l'écriture d'une page : à quel moment qui a ajouté ou supprimé quoi. Le wiki gère aussi tout seul les liens entre les pages, ce qui rend très simple l'écriture hypertextuelle puisqu'il suffit

87. Une collection facilement extensible de 'pages' Web liées entre elles, un système hypertexte pour conserver et modifier des informations - une base de données, dans laquelle chaque page est facilement modifiable par n'importe quel utilisateur simplement muni d'un client navigateur Web capable de manipuler des formulaires

de baliser un mot suivant la syntaxe adéquate pour qu'il soit transformé en lien par le wiki. Si la page cible a déjà été créée, le lien pointera vers elle. Sinon la création d'un lien vers une page qui n'existe pas encore créera en fait un lien vers le formulaire de création de ladite page, ce qui fait qu'un internaute qui navigue sur un wiki n'atteindra jamais la fameuse « page erreur 404 » qui signale une page non existante, il recevra à la place une invitation à écrire la page manquante.

De nombreux « moteurs » wikis sont maintenant disponibles, utilisant des solutions techniques différentes mais toujours basés sur les mêmes concepts de base :

- N'importe quel visiteur peut créer ou modifier une page.
- N'importe quel navigateur Web peut faire l'affaire
- Tout est très simple à réaliser (voire rustique)

L'idée fondatrice étant que le wiki ne nécessite ni outil, ni compétence, ni autorisation pour pouvoir participer et modifier une page. Dans la conception de ce dispositif, Ward Cunningham s'est inspiré d'un concept d'esthétique Japonaise appelé *wabi sabi*. *Wabi* pourrait être traduit par simplicité et calme, *sabi* par élégance et simplicité. Le wabi-sabi se plaît dans ce qui est incomplet, fugace, et imparfait, les valeurs en sont la modestie et l'humilité. Wabi-sabi est très lié à la philosophie Bouddhiste Zen, en particulier celle du *wu-wei*, qui est le concept de modération ou « ne pas faire ».

### 3. Wikipédia

Wikipédia<sup>88</sup> est une encyclopédie internationale, libre, gratuite et écrite coopérativement par ses lecteurs. Elle est construite techniquement sur un moteur de type wiki : le logiciel libre MediaWiki. Cette information et ce choix

---

88. Wikipédia est consultable en français à l'adresse : <http://fr.wikipedia.org/>

technique de Wikipédia ne sont pas neutres, cela veut dire que le contenu et le contenant sont distribués et mis à disposition sous les mêmes modalités libres. Il est donc tout à fait possible pour un autre projet encyclopédique de reprendre le même dispositif et de l'utiliser, voire de le modifier, pour ses besoins propres.

**Historique** Un court historique de ce projet permettra de mieux percevoir son succès et sa dynamique. La puissance éditrice que permet de mettre en œuvre un wiki n'a pas tout d'abord été réellement exploitée à grande échelle, parce que les sites qui avaient mis en place un wiki n'avaient pas de projet suffisamment fédérateur pour rassembler une large communauté. C'est finalement début 2001 que voyait le jour le projet Wikipédia qui allait mettre la technologie wiki à la portée de tous. Wikipédia est à l'origine issue de Nupédia, un projet d'encyclopédie en ligne gratuite fondée en mars 2000 par Jimmy Wales et soutenue par la société Bomis, dont il était l'actionnaire majoritaire). L'encyclopédie Nupédia fonctionnant traditionnellement avec un comité scientifique, sa progression se révéla très lente et, en janvier 2001, le rédacteur en chef Larry Sanger proposa la création d'un wiki dans le but d'accroître la vitesse de développement des articles. Il proposa alors à Jimmy Wales d'allier le potentiel du format wiki à la mise en place d'une encyclopédie plus souple, ce qui donna lieu à la création de l'encyclopédie Wikipédia. À partir de ce moment, Larry Sanger travailla parallèlement à l'encyclopédie Nupédia et à l'encyclopédie Wikipédia, pour laquelle il a participé à la formulation de la majorité des politiques d'origine. Le projet Wikipédia opère par consensus, utilisant les règles et recommandations adaptées au cours du temps par les différents contributeurs. Nupédia cessa ses activités

à cause de la lenteur de sa progression fin 2003 et Wikipédia, elle, continua au contraire son expansion.

**Une réussite quantitative** L'encyclopédie Wikipédia a ainsi vu le jour en anglais le 15 janvier 2001, l'encyclopédie francophone a été la première déclinaison dans une autre langue et elle a été fondée officiellement le 23 mars 2001. En mars 2005, Wikipédia existait dans exactement 200 langues<sup>89</sup>. Vingt encyclopédies possédaient plus de 10 000 articles. Les versions qui comportaient le plus d'articles étaient la Wikipédia en anglais (495.000 articles), en allemand (208.000 articles), en japonais (105 000 articles) et en français (Le samedi 23 avril 2005, Wikipédia a franchi la barre symbolique des 100.000 articles et elle contenait 110.802 articles le 26 mai 2005). Wikipédia est aujourd'hui la plus grosse encyclopédie du monde mais également celle qui grandit le plus vite. Chaque jour, quelque 2.500 articles sont ajoutés à Wikipédia, et plus de 25.000 sont modifiés. La figure 4.6 rend compte de cette progression sur la partie francophone et met très bien en évidence l'exponentialité de celle-ci.

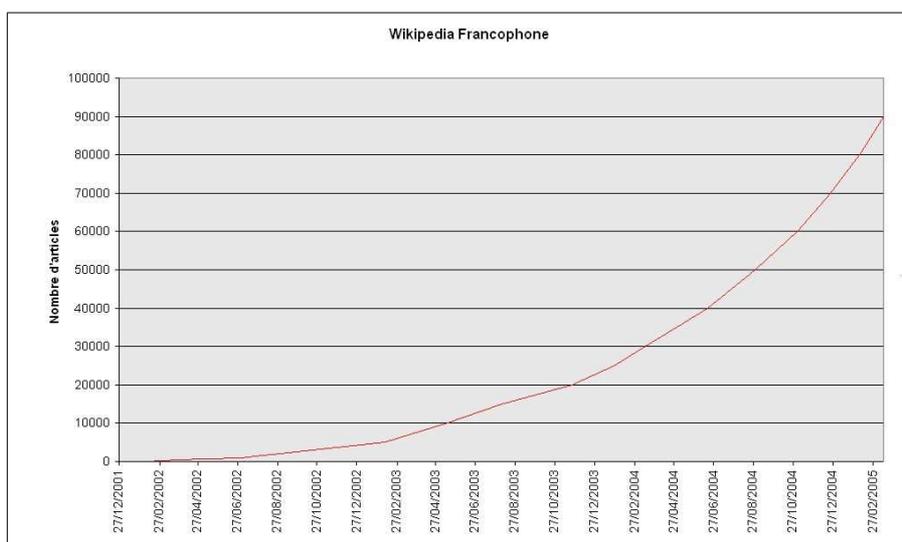


FIG. 4.6: Courbe de progression du Wikipédia francophone

---

89. Cette valeur actualisée est disponible en ligne à l'adresse : [http://meta.wikimedia.org/wiki/Complete\\_list\\_of\\_language\\_Wikipedias\\_available](http://meta.wikimedia.org/wiki/Complete_list_of_language_Wikipedias_available)

A première vue, ce projet international paraissait au départ utopiste. Pourtant, déjà de part le volume son succès est maintenant acquis. Le tableau 4.1<sup>90</sup> permet de comparer Wikipédia à d'autres encyclopédies plus connues ou plus traditionnelles. Wikipédia est de loin celle qui comporte le plus d'articles, tout en maintenant une moyenne du nombre de mots par article honorable, parce qu'il pourrait être facile d'ajouter une grande quantité d'articles creux ou presque vides, ce qui n'aurait aucune valeur.

Encyclopédie	Édition	Articles (milliers)	Mots (millions)	Mots/article moyenne
Encyclopédie de Diderot		75		
Wikipédia	English	458	166	362
Nationalencyklopedin	–	183		
Encyclopædia Britannica	2002	85	55	650
	En Ligne	120	44	370
Microsoft Encarta	Encarta Deluxe 2002	70	40	600
	Encarta Deluxe 2005***	63	40	600
	2002 Encarta Encyclopedia	40	26	600
Grolier Multimedia Encyclopedia Online		39.2	11	280
Columbia Encyclopedia		51	6.5	130

TAB. 4.1: Tableau comparatif entre encyclopédies

**La question de la pertinence** Il faut toutefois ne pas se laisser emporter par trop d'enthousiasme et s'attacher à vérifier la qualité du contenu. L'absence de

90. Source Wikipédia [http://en.wikipedia.org/wiki/Wikipedia:Size\\_comparisons](http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons)

comité de rédaction pourrait conduire à des errements, la taille de l'encyclopédie, sa disponibilité dans de nombreuses langues ont leur importance mais sans vérification sur la validité des articles cela conduirait à rien de solide. La manière dont est assurée ce contrôle est tout à fait intéressante à étudier, parce qu'il est fondé sur la collaboration à distance d'un grand nombre d'individus qui ne se connaissent pas, mais qui pourtant se font confiance.

Le contrôle de la qualité, l'objectivité et la neutralité de point de vue sont clairement affichés dans les règles de fonctionnement de Wikipédia. Mais, encore une fois, est-ce vraiment suffisant pour assurer ce niveau de qualité indispensable? Ces déclarations d'intention sont bien évidemment louables et servent à rassurer le lecteur. En fait, elles s'adressent de manière encore plus précise aux rédacteurs et dans ce contexte elles s'avèrent indispensables compte tenu du fonctionnement de Wikipédia : s'agissant de recruter à distance des collaborateurs dont il est difficile de contrôler à priori les bonnes intentions, il s'agit de clairement leur signifier les règles du jeu.

L'historique des modifications permet à n'importe quel lecteur/rédacteur de vérifier ce qui a été modifié, quand et accessoirement par qui, cette dernière information permettant aux habitués de se reconnaître et de construire petit à petit un réseau de confiance. La figure 4.7 montre l'historique des modifications apportées à un article particulier (la poulie). Nous pouvons retrouver la genèse de cet article, créé le 6 janvier 2004 par un utilisateur qui se fait appelé « Haha », mis à jour le soir même par un autre appelé « Dom », le lendemain « Haha » est revenu apporter quelques modifications. Nous pouvons aussi noter les deux passages d'un utilisateur particulier appelé « MedBot<sup>91</sup> qui est

---

91. L'article correspondant sur Wikipédia indique « *MedBot est – comme son nom l'indique – un bot dont la principale fonction est d'ajouter automatiquement des liens InterWikis aux pages de Wikipédia.* ». Bot est en fait une abréviation de Robot pour désigner un agent logiciel automatisé. »

un des dispositifs qui permet d'assurer la cohésion dans Wikipédia, en plus des administrateurs humains, en nombre relativement réduit.

The screenshot shows the Wikipedia article page for "Poulie". At the top, there are navigation tabs: "article", "discussion", "modifier", "historique", "renommer", and "suivre". Below the title "Poulie", there is a section for "Versions précédentes" (previous versions). A legend explains the symbols: "(actu)" for current version, "(dem)" for previous version, and "m" for minor modification. The history table lists 20 revisions, with columns for version number, date, time, and user. The most recent revision is 81.247.65.109 by user "MedBot" on May 28, 2005. Other notable users include "Cdang", "Dom", and "Haha". The page also features a search box, a "Consulter" button, and a footer with "WIKIPÉDIA L'encyclopédie libre" and "Powered by MediaWiki".

FIG. 4.7: Historique de modification d'un article

Il n'est pas utile de continuer à détailler cet historique, il suffira de noter que les noms des utilisateurs sont aussi des liens, qui permettent d'accéder à la page/article que chacun aura écrit sur lui-même pour se présenter. Si un internaute a fait une modification sans s'identifier c'est son adresse IP – c'est-à-dire l'adresse physique de sa machine dans le réseau Internet – qui est affichée. Sans que cela fasse complètement tomber l'anonymat, cela permet d'identifier les mauvais plaisants et de leur interdire de vandaliser l'encyclopédie si besoin était.

A la différence des autres encyclopédies qui ne sont revues au mieux qu'une fois l'an, Wikipédia est continuellement en relecture, modification et

mise à jour. Cela permet par exemple une réactivité sans précédent dans le monde des encyclopédies. Quelques minutes après un évènement, celui-ci peut être déjà mis en ligne comme par exemple l'élection du nouveau pape Benoit XVI le 19 avril 2005, la fumée blanche apparaît sur le toit de la chapelle Sixtine à 17h56. À 18h35, le cardinal protodiacre annonce publiquement sur la place Saint-Pierre son élection. Il n'a pas fallu plus de 13 minutes pour qu'une première personne n'ajoute le texte suivant à l'article sur le cardinal Joseph Ratzinger :

*« Le 19 avril 2005 à 17h50, il devient le 265ème pape de l'histoire et succède à Jean-Paul II sous le nom de Benoit XVI (Wikipedia, 2005a) »*

Les modifications se sont enchaînées à un rythme très rapide, remaniant complètement le texte et le migrant sous un nouveau titre d'article de manière à rendre compte de la nouvelle identité du cardinal.

Bien entendu, les responsables du projet sont bien conscients des critiques qui peuvent être émises quand à la pertinence des informations qui peuvent être trouvées en ligne. Ces critiques mettent en évidence la collision entre les objectifs d'une encyclopédie et la philosophie *wabi sabi* du Wiki, que nous avons citée précédemment et qui se résume à ces trois termes : incomplet, fugace et imparfait.

Comment faire confiance à une encyclopédie incomplète, changeante et imparfaite ? Un début de réponse consiste à rappeler que ces caractéristiques sont avant tout celles du dispositif. Ce problème était déjà en débat sur le premier wiki, il est simplement devenu plus visible et se pose maintenant à plus de monde avec Wikipédia. En plus des réponses qui proviennent des discussions autour du concept de Wiki, il existe plusieurs réponses spécifiques à Wikipédia : un article sur lequel les internautes ne sont pas d'accord ou si un problème de pertinence a été identifié par l'un d'eux peut être mis en

évidence comme cela est visible sur la figure 4.8 par exemple. Mais il existe aussi un article de Wikipédia (Wikipedia, 2005d) qui détaille les raisons de faire confiance à Wikipédia. Bien entendu, même cet article est développé en communauté, il est donc modifiable par tous les internautes et criticable au même titre que les articles qu'il est sensé défendre. Nous nous trouvons dans un paradoxe difficile à résoudre, la dernière chose qui pourrait nous donner confiance est la similarité avec les développements de Logiciels Libres. Au départ ceux-ci étaient critiqués et qualifiés de « non-sérieux », voire pire, par les professionnels du logiciel. Aujourd'hui la qualité de bon nombre de Logiciels Libres est reconnue. Il resterait à valider que la méthode pour produire des programmes peut être appliquée à la production d'un savoir, de plus tous les Logiciels Libres ne sont pas de bons logiciels.



FIG. 4.8: Problème de pertinence d'un article

Pour le lecteur qui peut se trouver désorienté à séparer le bon grain de l'ivraie dans la masse d'informations accumulées par Wikipédia, il est important de noter que les articles faisant référence à un savoir objectif, scientifique, peuvent souvent s'avérer être de très bonne qualité, parce qu'il est facile dans ce cadre de rassembler une communauté qui se regroupera autour d'un consensus. En revanche, pour les articles concernant des sujets qui font débat,

comme la politique, le religieux, voire même des théories scientifiques qui sont encore en discussion, ce mode d'écriture peu poser de réels problèmes. En effet, différents groupes d'opinions vont être tentés de s'affronter autour d'un article pour essayer d'imposer leur point de vue. Dans ces conditions l'article n'est plus un objet de savoir, mais un territoire à conquérir. Ce qui fait que « normalement » Wikipédia fonctionne c'est que face à un vandale, une communauté va venir s'interposer pour protéger et réparer l'article vandalisé. Chaque fois que ce rapport de force peut s'inverser, il y a risque pour la pertinence et la qualité des articles écrits dans ces conditions. Une secte, un parti politique ou n'importe quel groupe organisé d'individus peut décider de s'engager dans l'écriture d'un article sur le ou les thèmes qui sont chers à ce groupe. Face à ce type de pression organisée, le seul moyen que Wikipédia aie de se protéger est le retour à un processus rédactionnel plus traditionnel, il est donc demandé à tous les lecteurs de signaler un article qui ne correspondrait pas à la charte d'écriture de Wikipédia. Un administrateur Wikipédia a ensuite la possibilité de verrouiller l'article sujet à caution, s'il s'avère qu'il est effectivement entâché de prises de position qui n'ont pas leur place dans une encyclopédie. L'article restera ainsi verrouillé tant que les lecteurs et les rédacteurs n'auront pas trouvé un compromis ou un terrain d'entente, ceci au travers d'un forum électronique qui accompagne chaque article.

Il semble qu'il faille considérer Wikipédia comme ce qu'elle prétend être : une grande source d'information, à l'échelle du Web, généralement pertinente mais que chaque internaute a le pouvoir de corriger quand il trouve une erreur. Ceci au contraire du Web où le meilleur côtoie le pire sans qu'il soit possible de l'améliorer. Un autre article peut être intéressant à suivre, il recense

les articles de la presse anglo-saxonne qui utilisent Wikipédia dans leurs sources (Wikipedia, 2005e).

## D. Les systèmes hypertextes

Par rapport aux nouveaux usages introduits par le livre électronique, que nous avons évoqués dans la section 3.4, page 73 du chapitre précédent, l'organisation hypertextuelle est sans doute la modalité qui permet le plus facilement de les appréhender. C'est sans doute celle qui se démarque radicalement de l'ouvrage papier en constituant une nouvelle matérialité à partir du texte électronique dématérialisé de son support papier originel. Dans la dernière partie nous reviendrons vers la construction de nouvelles pratiques, voire de cette nouvelle matérialité pour le livre introduite par les systèmes hypertextes.

A ce moment de notre étude il nous faut simplement rappeler les problèmes habituellement attribués à la lecture hypertextuelle, soit la désorientation et la surcharge cognitive. Nous travaillerons sur ces problèmes dans les chapitres suivants, en liaison avec la *Construction du sens* et la *Représentation des parcours de lectures*.

Il existe de nombreux dispositifs permettant l'écriture hypertextuelle, la plupart ont ceci de particulier qu'ils sont nécessaires à l'écriture mais aussi à la lecture, ce qui rend la diffusion des documents compliquée, car elle demande aux éventuels lecteurs de se munir du même dispositif. La liste ci-dessous est bien loin d'être exhaustive, mais elle essaie de mettre en évidence les logiciels ou dispositifs les plus marquants.

**HyperWave**<sup>92</sup> Hyperwave a été développé par IICM – *Institute for Information Systems and Computer Media* – à Graz. Son nom original était Hyper-G et

---

92. <http://www.hyperwave.com/>

il a pris le nom de Hyperwave lors de sa commercialisation en 1996. Il s'agit d'un système d'information similaire à un système de gestion de base de donnée qui permet de lier des documents entre eux et en les enrichissant au moyen de métadonnées.

**Microcosm** <sup>93</sup> Microcosm est né à l'Université de Southampton, au royaume Uni, en 1985. Le dispositif a évolué et s'appelle maintenant Active Navigation. Il s'agit d'un système de gestion de l'information qui permet d'accéder plus rapidement au document cherché.

**Storyspace** <sup>94</sup> Storyspace est un vrai logiciel permettant l'écriture d'hypertextes plutôt qu'un gestionnaire d'information comme les précédents. C'est un outil qui s'adresse à des auteurs de fictions, à des scénaristes, à des professeurs désireux de créer des textes dans lesquels le lecteur devra faire ses propres choix et construire son propre parcours.

**Xanadu** <sup>95</sup> Xanadu est le projet de système hypertexte et multimédia initié par Ted Nelson en 1960. Le Web, HyperWave, Microcosm et Lotus Notes ont tiré parti des idées développées dans le cadre de Xanadu pour implémenter leurs propres solutions. Nous l'avons déjà mentionné dans le paragraphe *Historique* de la section 4.1, page 67. Il n'existe pas encore, à l'heure actuelle, de dispositif qui mette en œuvre les idées exposées dans ce projet.

**World Wide Web** <sup>96</sup> Le World Wide Web a été créé par Tim Berners-Lee au CERN au début des années 1990. Il s'agit de logiciels, de standards et de protocoles qui permettent d'utiliser une interface hypertextuelle à travers Internet. L'utilisation du Web pour rendre des ressources

---

93. <http://www.multicosm.com/>

94. <http://www.eastgate.com/Storyspace.html>

95. <http://www.xanadu.com.au/xanadu/>

96. <http://www.w3.org/WWW/>

accessibles à un large public soulève des questions sur la qualité de l'information, l'utilité et les modalités de la livraison à l'utilisateur final. Le multimédia ne fait qu'amplifier ce phénomène puisqu'il va falloir par exemple vérifier la qualité et la taille des images pour assurer des temps de réponse acceptables avec une qualité et une finesse minimales.

Il y a plusieurs aspects techniques complexes à résoudre dans l'environnement Web qui vise à mettre à disposition des documents au plus grand nombre :

- Format de fichier et compression utilisés, qui devront être supporté par la station de travail de l'internaute.
- Capacités du navigateur Web, notamment au niveau de sa conformité aux standards.
- Connections aux réseaux, pour permettre un débit correct dans toutes les situations, et ceci assorti d'un niveau de sécurité satisfaisant.
- Configuration matériel et système d'affichage de l'utilisateur final pour adapter le contenu à celui-ci. Ceci est particulièrement complexe si nous considérons la possibilité d'avoir un affichage adapté aux al-voyants par exemple. Ce qui nécessitera l'utilisation de très gros caractères, voire du braille ou d'un navigateur vocal.

### **1. Critique du Web comme système hypertextuel**

Ce n'est pas vraiment Internet mais plutôt l'interface hypertextuelle : la toile ou le Web pour « World Wide Web » qui nous intéresse. En effet le Web met à disposition une immense quantité d'informations textuelles et multimédia, sans pour autant qu'aucune hiérarchie ou catégorisation ne permette de différencier le ludique du sérieux, le pertinent de l'approximatif, voire de la désinformation ou pire. La force de la toile constitue donc sa première

faiblesse : un océan d'informations auquel tout internaute peut participer, librement, pour autant qu'il en ait les moyens techniques et financiers mais échappant ainsi à tout contrôle, le pire côtoie le meilleur. Ainsi le Web qui était perçu et décrit par certains comme le premier pas vers une grande bibliothèque universelle pourrait montrer rapidement ses limites, si ce n'est que même les bibliothèques peuvent parfois receler des ouvrages qu'une culture accepte mais qu'une autre rejette. Il n'est donc pas du tout certain que, même dans l'absolu, cette bibliothèque universelle de tous les savoirs objectifs existe, parce qu'il aurait fallu auparavant que ce savoir ait pu être objectivé. Il est toutefois certain que le contexte Web étant ce qu'il est, l'absence de contrôle rend encore plus aléatoire la qualité et l'objectivité des informations mises en ligne. Pour l'instant nous restons dans un paradigme où la quantité remplace sans doute la qualité et de nombreuses recherches s'attachent à dimensionner le Web dans le but d'en prouver le « potentiel », ceci en rapport direct avec la *quantité de savoir* qu'il représente ou contient. Ces études ne se posent pas ou rarement la question de la pertinence de ces informations que nous pouvons trouver.

Pour donner un aperçu de l'étendue des informations disponibles en ligne, nous devons tout d'abord rappeler les différentes formes techniques qui peuvent constituer un document avant qu'il ne s'affiche dans la fenêtre d'un afficheur Internet. En effet, une « page Web » peut trouver sa source dans un document HTML pour sa forme la plus simple, ou d'une combinaison de différents documents : scripts, programmes, documents HTML, XML ou autre ainsi que résultats d'extraction de base de données pour la forme la plus complexe.

La plus grande partie des informations pertinentes de la toile ne provient donc pas de documents physiques aisément identifiables mais de requêtes

dans différentes bases de données, ce qui rend la combinatoire des possibilités de représentation de ces informations immense. Les informations qui font typiquement partie de ce « Web profond » sont celles liées à des annuaires, téléphoniques ou professionnels par exemple, les définitions de dictionnaires, d'encyclopédies, les informations légales, journaux officiels ou autres, mais nous y trouvons aussi des informations non-textuelles, c'est-à-dire multimédia ou graphiques qui ne peuvent pas être étiquetées, labelisées automatiquement par les moteurs de recherche. Dans un autre ordre d'idée nous trouvons toujours dans ce « Web profond » des informations volatiles, dont le contenu est sans cesse modifié comme le sont par exemple les informations météorologiques, celles concernant des emplois, des réservations de train, d'avion, de train, de théâtre ou autres, ou encore les informations financières, boursières. Dès 1994, le Dr. Jill Ellsworth a utilisé le terme « Web invisible » pour nommer les informations qui étaient invisibles aux moteurs de recherche conventionnels d'alors qui étaient uniquement capables d'indexer des pages html statiques. Ce terme n'est pas très exact, les pages sont uniquement invisibles de ces moteurs, de nombreux moyens existent maintenant pour permettre d'y accéder. Le vrai problème n'est pas d'identifier ce qui est visible et ce qui ne l'est pas, c'est surtout la technologie d'accès des moteurs de recherche qui est en cause. Pour cette raison, une société américaine, BrightPlanet<sup>9798</sup>, a présenté en juillet 2000 les résultats d'une étude approfondie sur la quantité de documents disponibles en ligne, en arrivant à la conclusion que leur nombre s'élève à plus de 500 milliards de documents uniques dans ce qu'ils ont appelé le « deep Web » dans un document paru pour la première fois dans *The Journal of Electronic* (DeepWeb, July 2001).

---

97. <http://www.brightplanet.com/>

98. <http://www.brightplanet.com/technology/deepweb.asp>

## 4.4 La Gestion électronique de Documents

### A. Définition

La gestion électronique de documents (GED) est une technologie issue des NTIC qui se propose d'assurer la lisibilité, la pérennité et l'universalité des documents, ceci tout en améliorant leur accessibilité et surtout les possibilités de recherche. La GED est souvent associée avec des outils de gestion des connaissances (*Knowledge Management* — *KM*) ou de *Data Mining*, l'ensemble formant un tout cohérent d'outils. Les besoins de lisibilité et de pérennité sont généralement aisément résolus au moyen de simple solutions techniques d'archivage, plus ou moins sophistiquées, mais généralement assez proches de ce qui existe pour n'importe quel type de document électronique. Là où la GED doit satisfaire des besoins beaucoup plus spécifiques est tout ce qui touche au catalogage et à l'indexation. De plus, le support électronique interdit *a priori* la pratique courante qui consiste à annoter dans la marge d'un ouvrage ses propres commentaires, en vue d'en faciliter l'étude ou la critique. Pour permettre à la fois le catalogage et les annotations, des langages et métalangages permettant de donner une description sémantique et hiérarchique des textes ont été mis au point.

### B. Le traitement du document

« *Le véritable enjeu deviens l'accès aux documents et leur référencement.*

*La volumétrie de BN Opale plus s'élève ainsi à environ 7 961 373 notices bibliographiques au début 2003, soit la quasi totalité des documents imprimés, livres, périodiques, documents imprimés numérisés de la BNF.*

*Les catalogues en ligne, moteurs de recherche, bases bibliographiques indexent ainsi en permanence les nouvelles références, notamment dans le domaine des revues scientifiques et techniques. Parmi les plus connues citons article@inist le catalogue en ligne des articles et monographies disponibles depuis 1990 du fonds de l'institut de l'information scientifique et technique, articlescience, qui recense 7 millions de références issues de près de 9000 revues françaises et internationales depuis 1990, BVDL Journals (texte intégral de 200 périodiques), Elsevier science qui avec son journal table of contents contient le sommaire d'environ 1100 de leur revues scientifiques, Ingenta, base bibliographique qui permet l'accès aux références d'articles de 27 000 périodiques et au texte intégral de 5400 titres essentiellement de langue anglaise depuis 1998.*

*Mais au delà des livres et périodiques, c'est la quasi totalité des ressources en information du web qui est répertoriée par des moteurs de recherche de plus en plus puissants, certains logiciels tels Copernic constituant même des méta-moteurs de recherche qui intègrent les résultats de plusieurs moteurs de recherche lancés simultanément. (Van Cuyck, 2003) »*

Le traitement documentaire passe par 3 opérations :

- L'indexation consiste à repérer les thèmes abordés par le document, afin de les traduire en langages documentaires. Ces langages permettent de s'entendre sur la façon de désigner les concepts (répertoire de mots-matière RAMEAU, classification Dewey).
- Le catalogage consiste à établir une fiche signalétique du document. Il rassemble les renseignements nécessaires à son identification (nom de l'auteur, titre, nom de l'éditeur, année d'édition...).

- L'insertion dans un catalogue, c'est-à-dire la mise en mémoire et la création d'une série d'accès différenciés à l'information : accès auteurs, accès titres, accès sujets...

Le bibliothécaire indexe le document, alors que pour l'utilisateur ce sont ses questions qu'il tente de traduire en langage documentaire. D'où l'avantage pour les bibliothécaires d'utiliser un même langage documentaire.

Nous pouvons mettre cette description en relation avec les tâches clés de la recherche d'information, telles qu'elles sont décrites dans (Aussenac-Gilles et Condamines, 2004) :

- « 1. Construction de thésaurus ou de langages documentaires pertinents pour un fonds documentaire ;
2. Indexation d'ouvrages à partir de thésaurus existants ;
3. Recherche de textes pertinents en fonction d'une demande d'information particulière.

Ces trois moments font intervenir différents types d'acteurs. En 1, un documentaliste construit un thésaurus en direction d'un autre documentaliste (les deux rôles pouvant être tenus par la même personne). Cette tâche est surtout centrée sur la couverture du fonds, indépendamment d'utilisateurs particuliers. Il s'agit donc d'assurer que l'ensemble des usages langagiers (considérés comme autant de traces de connaissances) est pris en compte. En 2, un documentaliste utilise le thésaurus construit en 1 pour identifier des mots-clés pertinents à la fois du point de vue de l'ouvrage à indexer (usages langagiers) et de l'utilisateur pressenti (le documentaliste connaît les utilisateurs de son fonds et leurs besoins). En 3, la tâche est centrée sur un utilisateur particulier, auquel il faut fournir des textes pertinents et ce, même si, apparemment en tout cas, ces textes

*ne font pas partie d'un classement préétabli, i.e. n'obéissent pas forcément à une homogénéité langagière. »*

Sur la figure 4.9 suivante tirée du même ouvrage, nous avons une vue synthétique de ces étapes qui viennent d'être citées. Nous retrouvons les tâches du documentaliste, notées de 1 à 3 autour des personnages D (documentaliste) et U (utilisateur), qui ont pour but de préparer la recherche d'information en indexant les collections en sa possession sous la forme de différentes collections et mises à la disposition du public. L'intérêt étant que l'utilisateur puisse accéder plus rapidement aux informations qu'il cherche.

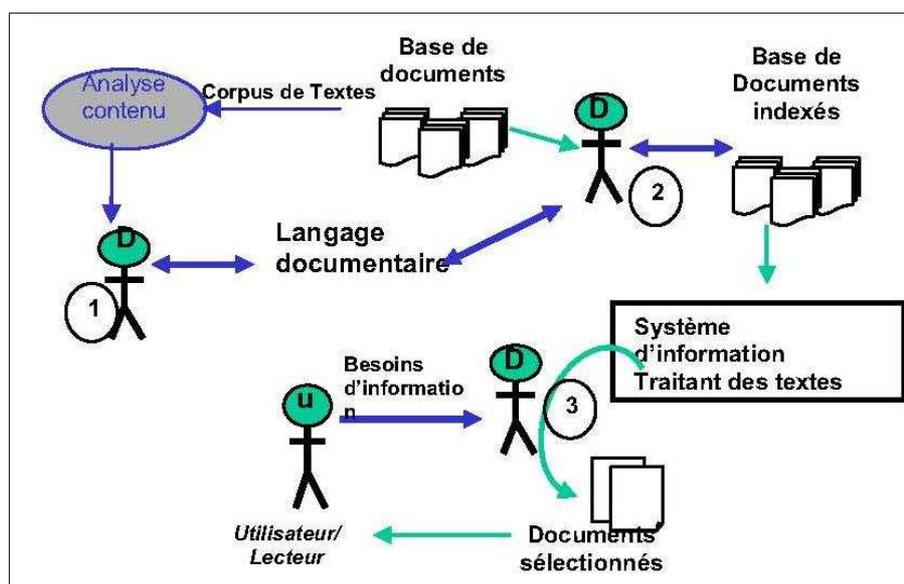


FIG. 4.9: Les tâches du documentaliste

## 1. L'indexation

L'indexation consiste à analyser le document pour les informations qu'il contient.

Des outils d'analyse communs à toutes les bibliothèques permettent d'extraire les principaux concepts des documents.

Cette description repose sur deux langages :

- les répertoires de mots-matières rassemblent des listes de termes autorisés pour la description du contenu. L'indexation d'un document est l'opération qui consiste à trouver les mots-matière (ou vedettes-matière) qui correspondent le mieux au contenu du document. Les notices des fichiers papier de la BU de Lille 3 ont été indexées à partir du langage documentaire Blanc-Montmayeur. Les bibliothèques universitaires françaises ont opté pour le répertoire RAMEAU, les notices du fichier informatisé du SCD (L'OPAC) ont été indexées à partir du répertoire d'autorités RAMEAU. La fonction des répertoires de mots-matière est de proposer aux bibliothécaires des listes de termes qui permettent de traduire les concepts identifiés dans les documents de la même manière, d'une bibliothèque à une autre. Ce sont ces termes qui donnent accès aux notices et permettent au lecteur de trouver les documents.
- les langages classificatoires permettent d'attribuer un indice numérique qui situe les informations dans un domaine de la connaissance. Les Bibliothèques Universitaires ont recours à la classification Dewey. L'indice est la traduction d'un concept par un chiffre issu d'une classification. Cet indice peut théoriquement être très développé pour représenter le ou les thèmes abordés dans le document.

**Les langages alphabétiques d'indexation : RAMEAU** RAMEAU<sup>99</sup> (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) est le langage d'indexation élaboré et utilisé par la Bibliothèque nationale de France, les bibliothèques universitaires, ainsi que de nombreuses autres bibliothèques de lecture publique ou de recherche.

Il contient, dans des notices d'autorité reliées entre elles, le vocabulaire et les indications qui permettent de construire les vedettes-matière dans

---

99. <http://rameau.bnf.fr/>

un fichier bibliographique. Langage d'indexation précoordonné, RAMEAU est composé d'un vocabulaire de termes reliés entre eux et d'une syntaxe indiquant les règles de construction pour l'indexation.

A la différence d'un thésaurus, la liste d'autorité encyclopédique n'est pas constituée *a priori*, mais au fur et à mesure des besoins d'indexation et évolue sur la base des propositions faites par le réseau de ses utilisateurs. Le répertoire RAMEAU comporte plus de 300 000 termes.

RAMEAU est accessible sur cédérom (Répertoire d'autorités de la Bibliothèque Nationale de France). RAMEAU est un langage documentaire régulièrement mis à jour. Il permet l'indexation des ouvrages.

**La classification décimale de DEWEY** A la fin du XIX<sup>e</sup> siècle, l'américain Melvil DEWEY a inventé la Classification Décimale de Dewey (CDD). Il s'agit d'une organisation hiérarchisée des connaissances. Cette classification a fait l'objet de nombreuses mises à jour (21 éditions à ce jour). C'est une classification qui part du général pour aller vers le particulier.

La classification Dewey permet de situer les documents dans un domaine de la connaissance et de leur attribuer une place dans la bibliothèque. Pour que le lecteur puisse trouver des documents facilement par l'intermédiaire des catalogues, le classement doit tenir compte du contenu de l'ouvrage, et regrouper en un même espace tous les documents ayant trait à un même sujet. C'est le principe de l'accès libre.

L'ensemble du savoir est divisé en 10 classes, qui correspondent aux grands domaines de la connaissance. La notation de la CDD se compose d'un indice qui comprend au minimum 3 chiffres. La CDD obéit à un principe de hiérarchie : du général au particulier. Plus l'indice Dewey est long, plus le

domaine de connaissance auquel il se réfère est précis. Les 10 classes sont décrites dans la liste suivante :

- 000 : Généralités
- 100 : Philosophie
- 200 : Religions
- 300 : Sciences sociales
- 400 : Langage
- 500 : Sciences pures
- 600 : Techniques
- 700 : Arts et Loisirs
- 800 : Littérature
- 900 : Géographie, histoire

Chaque classe comporte 10 divisions. Ces divisions correspondent à la déclinaison de chacun des grands domaines de la connaissance. Par exemple si nous retenons la classe 300 des Sciences sociales nous retrouvons les 10 divisions suivantes :

- 300 = Sciences sociales
- 310 : Statistiques générales
- 320 : Sciences politiques
- 330 : Économie
- 340 : Droit
- 350 : Administration
- 360 : Services sociaux
- 370 : Éducation
- 380 : Commerce
- 390 : Coutumes.

Le principe de la classification de Dewey est mis en pratique et décrit sur un cas particulier de déclinaisons d'une classe principale en division et subdivision sur la figure 4.10<sup>100</sup>. Nous pouvons retrouver la hiérarchie entre les thèmes qui se précisent avec l'allongement de l'indice de Dewey.

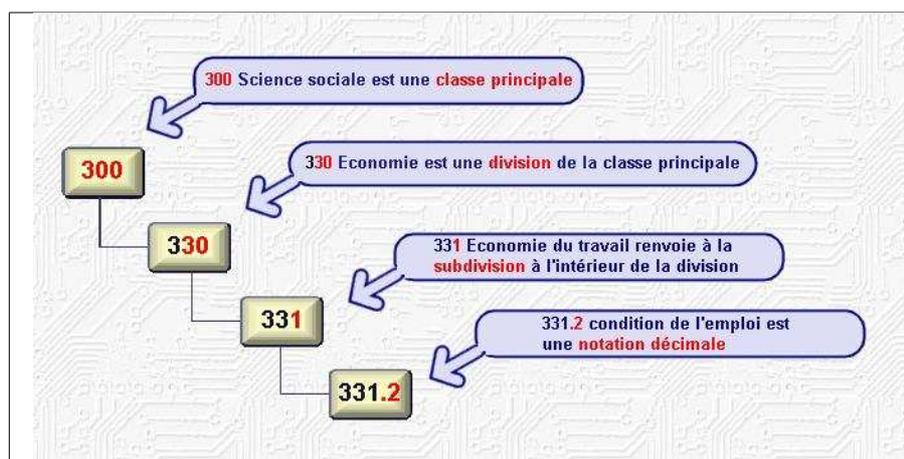


FIG. 4.10: Le principe de la classification de Dewey

Avant d'en terminer avec cette classification, il faut citer la CDU, mise au point par Paul Otlet que nous avons déjà cité dans la partie *Historique* de la section 4.1, page 67 alors que nous mentionnions son travail pour le développement d'une bibliothèque universelle qu'il avait baptisé Mundaneum. En effet, il avait mis au point dès 1905 et publié pour la première fois une proposition de *Classification Décimale Universelle* (CDU). Mais la CDU ne peut lutter contre la suprématie de la classification de Dewey, plus implantée dans le pays qui domine le marché de l'information, les États-Unis et elle est en train de disparaître lentement.

100. Ce schéma provient du site du Service Commun de la Documentation disponible en ligne à l'adresse <http://www.scd.univ-lille3.fr/>

## 2. Le catalogage

Le catalogage consiste à analyser le document en tant que support. Pour en réaliser la description, le personnel des bibliothèques et les documentalistes utilisent une norme internationale.

Cette norme identifie :

- les éléments à prélever (le titre, l'auteur, l'éditeur, etc)
- la partie du document où ils doivent être prélevés (la page titre par exemple)
- l'ordre dans lequel les éléments définis par la norme doivent être agencés et la forme qui doit être retenue

## 3. Le résumé

Les résumés sont écrits en langage naturel et reposent sur les opérations d'extraction, puis de condensation de l'information utile contenue dans le document d'origine. Des outils pour construire de manière automatique cette information existent, ils font appel à des techniques de traitement du langage naturel. Il est possible de distinguer plusieurs types de résumés, dont les deux principaux sont :

- les résumés indicatifs
- les résumés informatifs

L'écriture des résumés repose sur des normes (ordre des arguments, emploi de phrases nominales, vocabulaire conseillé...).

Certains logiciels documentaires modernes permettent l'interrogation sur les mots du résumé, ce qui vient compléter les possibilités de recherches par mots-clefs, il s'agit alors d'une interrogation « plein texte ».

## C. Les normes de balisage

### 1. DTD

Une Définition de Type de Document (DTD) est un document qui permet de décrire un modèle de document XML ou SGML.

Les DTD (ou Document Type Definition) sont les modèles de documents XML, là où les contraintes du format sont spécifiées.

### 2. SGML

SGML (Standard Generalized Markup Language) est une norme qui existe depuis 1986 et qui est utilisée pour baliser logiquement des textes et modéliser des documents structurés. Elle permet à tout utilisateur de définir, via l'écriture d'une DTD (Définition du Type de Document — *Document Type Definition*) un langage de balisage logique adapté à ses besoins. C'est une norme internationale (ISO 8879 :1986) (van Herwijnen, 1995), inventée par Charles F. Goldfarb. Un document décrit avec SGML ne comporte aucune référence à une présentation visuelle donnée, d'autre part il est indépendant de tout logiciel ou matériel. La technique des formes architecturales et des adresses de lieu permet de créer des liens entre plusieurs documents.

Par-delà ses nombreuses applications pratiques, SGML est d'un grand intérêt intellectuel en tant que métalangage et à travers le caractère inclassable des grammaires qu'il permet de construire.

### 3. XML

XML<sup>101</sup> (eXtensible Markup Language) est un standard de gestion de documents proposé par le W3C (World Wide Web Consortium)<sup>102</sup>, qui gère

---

101. <http://www.w3.org/XML/>

102. <http://www.w3.org>

aussi le standard HTML. C'est une forme simplifiée de SGML (Standard Generalized Markup Language). XML fait l'objet d'une riche littérature en tant que nouveau standard depuis 1998 et surtout en tant que nouveau passage obligé pour la publication sur le Web. Les choses semblent très simples au départ, un standard unique pour tout le Web, qui transcende les problématiques techniques de compatibilité logicielle et matérielle. A l'arrivée, comme cela a déjà été le cas pour HTML il y a quelques années, nous constatons que chaque acteur sur le Web, économique, académique ou institutionnel, essaie de tirer le standard vers ses propres besoins.

Beaucoup de systèmes électroniques de gestion de documents utilisent maintenant XML pour conserver les textes. La vocation première de XML est de conserver des données et nous pouvons identifier plusieurs initiatives dans ce sens, qui seront décrites avec les DTD de l'information-documentation dans la section D. suivante. Pour ne citer qu'elle, une initiative telle que Dublin Core travaille à mettre au point un format d'échange et, mot magique dans le milieu informatique, d'interopérabilité pour décrire les données ainsi que les métadonnées qui doivent les accompagner.

Malgré une apparente unanimité sur la pertinence de l'utilisation du standard XML, il est tout à fait légitime de se questionner sur ses réels apports et sur l'impact qu'il y aurait à l'utiliser ou à ne pas l'utiliser dans un système de bibliothèque informatisée par exemple. Arguments et contre-arguments peuvent être résumés de la façon suivante :

*« Oui, sans vouloir minimiser son impact, XML n'est rien d'autre que SGML. Le plus important, XML conserve les possibilités principales de SGML qui ont amené les groupes qui manipulent des données de haute importance, comme les utilisateurs de la TEI, à le retenir. Il permet de décrire des hiérarchies d'éléments personnalisables de manière arbitraire.*

*C'est un format ouvert, non-propriétaire, aussi les données sont plus à même de survivre aux changements technologiques. Cela reste un système plein-texte, de balisage inclus dans le corps du texte qui permet d'éviter un grand nombre de problèmes de portabilité et de pérennité qui surviennent avec les approches binaires. Cela autorise tout document XML valide à être aussi un document SGML valide.<sup>103</sup> (Derose, 1999, page 23) »*

C'est donc parce que XML est un standard à la fois ouvert et reconnu que nous pouvons avec confiance lui confier nos données, mais c'est aussi parce qu'il représente en fait un triple standard :

**Standard de stockage** C'est avant tout pour le stockage et la conservation pérenne des informations que nous l'avons présenté, il n'est donc sans doute pas nécessaire d'ajouter quoique ce soit.

**Standard d'échange** Pour les sites commerciaux XML n'est souvent uniquement qu'un format d'échange, un intermédiaire entre des applications hétérogènes.

Les données, stockées dans une base de données sous un format propriétaire sont traduites « à la volée » et transmises à une autre application qui, à son tour, va les traduire sous un autre format qui lui est propre pour les stocker dans une base de données.

**Standard de publication** Partant de là, étant donné que ces applications permettent de « produire du XML », un autre rôle souvent mis en avant, est l'utilisation de XML pour faciliter la transposition de différents types

---

103. Yet, despite minimizing its bulk, XML is still SGML. Most important, XML keeps the crucial capabilities of SGML that have led to groups with high-value data, such as TEI users, to choose it. It can describe arbitrary hierarchies of custom element type. It is an open, non-proprietary system so data is more likely to survive changes in technology. It remains a plain-text, embedded markup system, avoiding a large class of data portability and survivability problems that arise with any binary approach. It ensures that any valid XML document is also valid SGML.

de contenus vers le format Web. Cette fois le contenu traduit en XML est traduit une deuxième fois sous un format visualisable sur le Web, HTML généralement, par des procédés plus ou moins automatisés basés sur XSL, les feuilles de style ou d'autres technologies comparables.

Puisque ces applications permettent de « produire du XML », un autre rôle souvent mis en avant, est l'utilisation de XML pour faciliter la transposition de différents types de contenus vers le format Web.

Cette fois le contenu traduit en XML est traduit une deuxième fois sous un format visualisable sur le Web, HTML généralement, par des procédés plus ou moins automatisés basés sur XSL, les feuilles de style ou autres technologies comparables.

XML est aussi à la base des standards RSS, RDF qui doivent nous conduire vers le Web sémantique, dont nous reparlerons dans la section 2. de la page 101.

#### **D. Les DTD de l'information–documentation**

Nous avons déjà dressé une liste des formats utilisés pour le stockage des livres électroniques page 93 dans la section 4. *Normes, standards et formats* du chapitre 3. Ces formats sont basés sur différentes technologies, certaines étant propriétaires, d'autres ouvertes et publiées. À cette occasion, nous avons mentionné les DTD de l'information–documentation que nous allons voir maintenant, elles font partie des formats ouverts, leur finalité première est plutôt centrée sur le stockage des informations, à la fois textuelles et contextuelles. Dans ce cadre le livre électronique n'est qu'un simple dispositif de lecture.

## 1. TEI Text Encoding Initiative

Lancée en 1987, la TEI<sup>104</sup> est un standard international et interdisciplinaire qui permet aux bibliothèques, aux musées, aux éditeurs et aux chercheurs d'utiliser sous format numérique tous les types de textes littéraires ou linguistiques.

*« La TEI que l'on pourrait traduire par groupe d'initiative pour le balisage normalisé des textes est une norme de balisage, de notation et d'échange de corpus des documents électroniques fondée sur le SGML. Elle s'est élaborée de manière pragmatique à partir des besoins de structuration, de conceptualisation et de mise en réseau de textes. (Romary et Hudrisier, 2003) »*

La Text Encoding Initiative (TEI) est disponible sous deux formats une DTD SGML ou un schéma XML, elle est accompagnée par un volume de « recommandations » : les TEI « Guidelines » expliquant de quelle façon doivent être utilisés ces documents, DTD ou XML. Cette DTD est adaptée principalement aux besoins de la communauté des chercheurs en sciences humaines, mais elle peut en fait être étendue plus généralement à tout chercheur voulant explorer de vastes corpus textuels sous forme électronique. Les outils de balisage qu'elle met à disposition des chercheurs peuvent s'adapter à de nombreuses situations. Le linguiste aura la possibilité de baliser syntaxiquement des corpus, l'historien pourra marquer dans un texte des dates, des noms de lieu ou de personnage, le chercheur en littérature sera capable d'étudier la stylistique ou la genèse d'un texte, etc.

La richesse de ses possibilités pouvant au contraire rebuter certains projets dont le champ plus restreint ne nécessite pas de faire appel à un balisage

---

104. Cette initiative est représentée en ligne sur le site : <http://www.tei-c.org/>

complexe, une version épurée de la TEI a été développée sous la forme d'une DTD (ou d'un schéma XML) simplifiée appelée « TEI Lite ».

## 2. DocBook

DocBook<sup>105</sup> est un format ouvert né en 1992 à la suite d'un projet commun de HaL Computer Systems et l'éditeur O'Reilly visant à faciliter l'échange de documentation UNIX. Aujourd'hui, c'est le DocBook Technical Committee de l'OASIS<sup>106</sup> qui se charge de sa maintenance et de son évolution. Le standard DocBook a été développé initialement en SGML, il est désormais disponible en XML et se décline en une version simplifiée (Simplified DocBook).

A l'instar de la TEI, DocBook sert à baliser des textes narratifs, mais on l'utilise surtout pour produire de la documentation technique dans les domaines de l'informatique, des télécommunications et des technologies de l'information qu'il vise. Il est particulièrement adapté aux documentations techniques de matériels ou de logiciels informatiques, mais il n'est pas limité à ce type d'ouvrage. Comme il permet de séparer le contenu du document de sa présentation et qu'il s'agit d'un format libre et largement documenté, il garantit des documents pérennes. Il est particulièrement adapté au travail partagé de plusieurs auteurs ainsi qu'aux collections importantes de documents devant être mises à jour ou réorganisées fréquemment.

## 3. BiblioML

Le format de données BiblioML<sup>107</sup> est défini en fonction de la norme XML pour des références bibliographiques, basée sur le Format Bibliographique Unimarc, et pour des notices d'autorités, basée sur Unimarc / Autorités. Tous

---

105. <http://www.docbook.org/>

106. *Organization for the Advancement of Structured Information Standards*, en ligne à l'adresse <http://www.oasis-open.org/>

107. <http://www.biblioml.org/>

les outils XML peuvent être utilisés pour manipuler des documents BiblioML. Le format Unimarc a été utilisé comme référence pour l'analyse de données. En d'autres mots, BiblioML est une représentation XML du format Unimarc.

#### 4. Dublin Core

La *Dublin Core Metadata Initiative* (DCMI), ou *Initiative de métadonnées du Dublin Core*, ou plus simplement *Dublin Core*<sup>108</sup> est un forum ouvert dont l'objectif est de développer des standards d'interopérabilité de métadonnées en ligne qui supportent une grande variété d'utilisations et de modèles économiques. Les activités de la DCMI consistent en l'organisation de groupes de travaux, de conférences et d'ateliers, , et des efforts pour promouvoir l'acceptation générale des standards et pratiques des métadonnées. La documentation française donne en introduction les grandes lignes de cette initiative :

*« La norme de métadonnées du Dublin Core est un ensemble d'éléments simples mais efficaces pour décrire une grande variété de ressources en réseau. La norme du Dublin Core comprend 15 éléments dont la sémantique a été établie par un consensus international de professionnels provenant de diverses disciplines telles que la bibliothéconomie, l'informatique, le balisage de textes, la communauté muséologique et d'autres domaines connexes.*

*On trouvera une description de l'ensemble des éléments du Dublin Core dans la Section 4. Chaque élément est optionnel et peut être répété. Chaque élément possède également un ensemble limité de qualificatifs, des attributs qui peuvent être utilisés afin de raffiner davantage (et non pas étendre) la signification de l'élément. L'Initiative de métadonnées du Dublin Core (IMDC) a défini, en juillet 2000, des façons normalisées de*

---

108. Le site de référence se trouve à l'adresse : <http://dublincore.org/>

*« qualifier » les éléments au moyen de différents types de qualificatifs. Un registre de qualificatifs conformes aux « meilleures pratiques » de l'IMDC est en cours de construction.*

*Bien que le Dublin Core favorise la description d'objets ressemblant à des documents (car la description des ressources textuelles traditionnelles est une activité bien maîtrisée), son usage pour la description des ressources ne ressemblant pas à des documents traditionnels va dépendre, jusqu'à un certain point, des similitudes entre les métadonnées de ces nouveaux documents par rapport aux métadonnées habituelles d'un document. Il va aussi dépendre des objectifs visés par les métadonnées de ces nouveaux documents. (Hillmann, 2001) »*

Dublin Core a pour objectif de concilier les caractéristiques suivantes :

- Simplicité de création et de gestion
- Sémantique communément comprise
- Envergure internationale
- Extensibilité

**Deuxième partie**

**Hypertexte**

**et**

**construction du sens**



CONSTRUCTION DU SENS

---

*« Et quand la guerre sera finie, un jour, une année viendra où l'on pourra récrire les livres ; les gens seront convoqués, un par un, pour réciter ce qu'ils savent, et on composera tout ça pour le faire imprimer, jusqu'à ce que survienne un nouvel âge des ténèbres qui nous obligera peut-être à tout reprendre à zéro. Mais c'est ce que l'homme a de merveilleux ; il ne se laisse jamais gagner par le découragement ou le dégoût au point de renoncer à se remettre au travail, car il sait très bien que c'est important et que ça en vaut vraiment la peine. »*

Ray Bradbury

*Fahrenheit 451*

## Sommaire

---

<b>5.1</b>	<b>Introduction</b>	<b>225</b>
A.	Définition du sens	226
B.	Définition de la quantité d'information	230
<b>5.2</b>	<b>Mémoire de lecture</b>	<b>233</b>
A.	Réseau et apprentissage	234
B.	Restitution des parcours	237
C.	Modèle de représentation	238
<b>5.3</b>	<b>Le sens</b>	<b>239</b>
A.	Cadre de référence théorique	239
B.	Créer du sens, pour qui, par qui?	240
C.	Navigation et recherche d'information	244
D.	Représentation de la connaissance	245
<b>5.4</b>	<b>Sémantique interprétative</b>	<b>252</b>

---

## 5.1 Introduction

L'étude de la construction du sens dans un parcours hypertexte nous a amené tout d'abord dans le chapitre 3, page 53 à envisager les différentes modalités de la lecture, en particulier de l'influence du média numérique. La théorie de l'hypertextualité et sa mise en œuvre dans *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, au travers du modèle à trois modules développé par l'équipe « *Hypertextes et textualité électronique* », que nous avons vu au chapitre 4, page 163 nous ont permis de continuer notre évolution vers un écrit plus maléable, plus adaptable à son lecteur.

*« Qu'elle soit populaire ou érudite, ou lettrée, la lecture est toujours production de sens. (Goulemot, 1985, page 115) »*

Reste à étudier comment se fait cette adaptation, en particulier dans le cas où l'écrit réagit aux lectures. En même temps que le lecteur fait son chemin à travers *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, ce chemin se creuse un peu plus et devient petit à petit visible des autres lecteurs.

Ce que nous dit Henri-Jean Martin sur sa propre activité de lecteur, ou plutôt de *liseur* comme il a l'habitude de le dire est intéressant parce qu'il résume sans doute assez bien le problème de désorientation, en le situant par rapport à la construction de la pensée pendant l'activité de lecture :

*« J'utilise l'écran pour pénétrer certains textes difficiles, ceux de philosophes comme Husserl par exemple. Utilisant des systèmes d'agrandissement, j'ai ainsi l'impression de pouvoir dégager l'essentiel. Il y a donc là une nouvelle manière de lire et de comprendre qui me semble plus analytique. Par ailleurs il est bien évident que les techniques informatiques facilitent grandement la consultation et l'interrogation d'un texte. Elles instaurent une sorte de dialogue extrêmement fructueux. En*

*revanche, je reprends mon livre sur papier à partir du moment où je veux en dégager une vue d'ensemble, en saisir l'architecture. La construction globale de la pensée, la suite des idées se dégagent mieux. (Martin, 2004, page 279) »*

Bien entendu cette étude devra s'articuler avec notre dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, ce qui sera fait en partie dans la section présente et qui sera parachevé dans la partie III, page 367. Nous souscrivons en partie à ce qu'écrit Jean-Yves Prax :

*« Les nouvelles technologies sont un composant technologique essentiel de la capitalisation de la connaissance et de l'intelligence collective. L'outil ne fait pas le changement, mais bien utilisé il peut être un formidable catalyseur du changement. (Prax, 1998, page 3) »*

La remarque qu'il nous paraît importante d'ajouter à cette citation est que bien entendu l'outil ne fait pas le changement, l'automobile n'est *a priori* pas liée aux problèmes de santé publique, pourtant c'est l'utilisation qui est faite par les médecins qui leur a permis d'aller consulter plus loin, plus vite et donc de plus nombreux patients. Mais il s'agit d'utiliser le bon outil.

## **A. Définition du sens**

*« Sens se dit principalement en trois sens : comme sensibilité (le sens de l'odorat), comme direction (le sens d'un fleuve), comme signification (le sens d'une phrase). Un sens, c'est ce qui sent ou ressent, ce qui suit ou poursuit, enfin ce qu'on comprend. (Comte-Sponville, 2001, définition du sens, page 577) »*

Nous avons déjà abordé dans le chapitre 3 le nouveau rapport à l'écrit suscité par les documents électroniques. Le support électronique confère aux documents une nouvelle matérialité, ou plutôt une non-matérialité : un

document électronique n'est-il pas sensé être conservé dans une bibliothèque virtuelle ? D'une part nous trouvons un support plus flexible que l'impression sur papier et qui autorise une diffusion de manière beaucoup plus facile et d'autre part ce support est souvent accusé d'altérer la valeur des documents qu'il abrite, certains iraient même jusqu'à dire que le document électronique altère le sens même du document lui-même.

La versatilité, l'ubiquité autorisée par le support électronique laissent penser qu'il est si facile de modifier et de diffuser un nouveau document que les idées qu'ils contiennent sont tout aussi éphémères et inachevées. En fait, à y regarder de plus près, nous serions tentés de trouver dans le support électronique un renouveau de l'oralité, ou plutôt une oralité retrouvée à travers un nouveau média, qui n'est plus une langue parlée, mais une langue écrite qui aurait acquis le dynamisme et la spontanéité du discours oral. Si nous abordons la problématique du sens sous cet angle, c'est-à-dire en prenant le l'écrit électronique comme une nouvelle langue « parlée », nous pourrions mettre en évidence les caractéristiques particulières de construction du sens dans un corpus électronique, comparables à celles déjà décrites par Jean-Jacques Rousseau au sujet du discours :

*« L'écriture, qui semble devoir fixer la langue, est précisément ce qui l'altère ; elle n'en change pas les mots, mais le génie ; elle substitue l'exactitude à l'expression. L'on rend ses sentimens quand on parle, et ses idées quand on écrit. En écrivant, on est forcé de prendre tous les mots dans l'acception commune ; mais celui qui parle varie les acceptions par les tons, il les détermine comme il lui plaît ; moins gêné pour être clair, il donne plus à la force ; et il n'est pas possible qu'une langue qu'on écrit garde long-temps la vivacité de celle qui n'est que parlée. On écrit les voix et non pas les sons : or, dans une langue accentuée, ce sont les sons, les*

*accens, les inflexions de toute espèce, qui font la plus grande énergie du langage, et rendent une phrase, d'ailleurs commune, propre seulement au lieu où elle est. Les moyens qu'on prend pour suppléer à celui-là étendent, allongent la langue écrite, et, passant des livres dans le discours, énervent la parole même. En disant tout comme on l'écrirait, on ne fait plus que lire en parlant. (Rousseau, 1781) »*

Bien entendu, les documents étudiés dans <sup>C</sup>*o*<sup>L</sup>*i*<sup>S</sup>*ciences* provenant d'un corpus originellement « papier », il nous faudra naviguer entre écrit et électronique, entre le sens construit par l'auteur et le sens construit par les parcours du lecteur, voire par les parcours des lecteurs. C'est ce dernier cas qu'il sera intéressant d'étudier, la construction de sens au travers des dispositifs <sup>C</sup>*o*<sup>L</sup>*i*<sup>S</sup>*ciences* et la manière dont ils font émerger des « images » ou des « vues » du corpus, telles que les lectochromes, qui n'existent que pendant qu'un lecteur les visionne et ne sont jamais identiques à ce qu'ils ont déjà affiché ni à ce qu'ils afficheront. Par ce biais, le corpus écrit retrouve une dimension éphémère et insaisissable comme seul un discours écrit pouvait s'en prévaloir.

Comme nous l'avons présenté dans les premières pages Vanevar Bush avait pressenti la future naissance de l'hypertexte dans son article « *As We May Think* » qui plus que d'hypertexte traite de processus de pensée, comme le titre de l'article le montre. Ce que Bush nous explique c'est que pour qu'un document soit utile, il doit être consulté, voire annoté, complété, enrichi ; il doit donc être organisé et rangé de façon à permettre toute forme de consultation, en particulier celle qui procédera par associations.

*« La pensée humaine [...] opère par associations. On ne peut espérer reproduire artificiellement ce processus mental dans sa totalité mais on peut certainement en extraire des connaissances. On ne peut égaler la*

*vitesse et la souplesse avec lesquelles la pensée poursuit un cheminement associatif mais il doit être possible de surpasser la pensée en ce qui concerne la permanence et la clarté des items auxquels on fait appel. (Bush, 1945) »*

Un autre éclairage est apporté par Jean Clément qui résume bien la caractéristique essentielle du système hypertexte qui a rendu cette étude possible, c'est que

*« L'hypertexte peut s'envisager comme un système à la fois matériel et intellectuel dans lequel un acteur humain interagit avec des informations qu'il fait naître d'un parcours et qui modifient en retour ses représentations et ses demandes. (Clément, 1995) »*

Ce qui veut dire que les interactions du lecteur ne sont pas neutres. D'un lecteur à l'autre, on ne verra pas naître les mêmes informations et ils ne réagiront pas de la même manière aux informations qu'ils reçoivent.

Les questions que l'on est alors en droit de se poser sont : tous les hypertextes possèdent-ils ces qualités ? La réponse négative semble s'imposer, ou tout du moins, sans un minimum de préparation, l'hypertexte ne pourrait que donner à lire des trivialités qui finalement apparaîtraient plus ou moins de la même manière pour tous les lecteurs. Comment alors dans ces conditions, créer un système hypertextuel qui permette une mise en valeur des stratégies sémantiques. Le problème n'est pas nouveau et déjà évoqué par Jean-Pierre Balpe :

*« L'hypertextualisation se heurte donc à des besoins non seulement en sémantique générale - quelles sont les relations entre tel ou tel terme d'une langue ? - mais également en sémantique pragmatique. Dans la représentation de l'ensemble des connaissances humaines, comment telle*

*notion, représentée contextuellement par tel ou tel terme, intervient-elle ?*

*(Balpe et coll., 1996) »*

## **B. Définition de la quantité d'information**

Pour continuer notre travail, nous devons postuler que le sens se construit à partir des informations à notre disposition, à partir de là l'étude de la théorie de l'information devrait être capable de nous fournir les fondations sur lesquelles construire notre travail. En effet, si le sens dans un texte est produit à travers les informations qu'il porte, il peut être intéressant et surtout nécessaire de se doter des moyens pour analyser cette quantité d'information. Cela permettrait, le cas échéant, de lier sens et quantité d'information ou tout du moins de chercher à le faire.

La théorie de l'information, due à Shannon aux alentours de 1948, avec l'influence des grands théoriciens de l'informatique (Turing, Van Neumann, Wiener) présente le problème comme celui de la communication entre une source et un récepteur : la source émet un message que le récepteur lit. Nous ne pouvons parler d'information d'un message que par rapport à un ensemble donné de messages. En d'autres termes, déterminer la quantité d'information nécessite le choix d'un modèle. Les travaux de Shannon (Shannon, 1975) sont nés de l'étude au sein de la Compagnie Bell de problèmes particuliers aux télécommunications (télégraphe, téléphone, radio, télévision). Ils aboutissent à une théorie de la communication essentiellement mécaniste, c'est-à-dire ne tenant pas compte de la signification des messages transmis. Le problème à résoudre est purement technique : quel codage optimal peut-on appliquer à des messages choisis dans un ensemble connu afin de les transmettre le plus fidèlement et le plus rapidement possible en présence de parasites ? Shannon définit la quantité d'information contenue dans un message comme

une fonction de la fréquence d'utilisation des différents symboles composant le message. Pour des jeux de symboles suffisamment significatifs du point de vue statistique, il est d'usage d'assimiler la fréquence à la probabilité d'apparition des symboles. Ce qui est formulé par l'équation de la quantité d'information due à Shannon de la façon suivante :

$$H = - \sum_1^n p_i \log_2 p_i$$

Dans cette formule, nous supposons que le message est produit par une source émettant les symboles  $S_i$ , chaque symbole contenant la quantité d'information  $-\log p_i$  et  $p_i$  étant la probabilité d'occurrence du message transmis ( $i$ ) parmi les  $n$  formes qu'est susceptible de prendre ce message émis.

Cette équation étant très similaire à la forme de la célèbre formule du thermodynamicien Boltzman, qui permet d'évaluer l'entropie<sup>109</sup> d'un système, les scientifiques se sont interrogés sur la nature de la parenté information-entropie : similitude formelle ou traduction d'une réalité physique? Très similaire mais pas identique, une différence tout à fait notable étant le signe  $-$ , nous avons donc non pas ici l'entropie mais la néguentropie du système communicant.

*« La notion de quantité d'information est liée très naturellement à une notion classique dans la mécanique statistique : celle d'entropie. Tout comme la quantité d'information dans un système est une mesure de son degré d'organisation, l'entropie d'un système est une mesure de son degré de désorganisation, et l'un est simplement le négatif de l'autre. (Wiener, 1948, page 16) »*

109. C'est le potentiel énergétique disponible dans une enceinte gazeuse

Nous pouvons aussi voir dans cette reformulation mathématique de la thermodynamique vers la communication un glissement d'ensemble de la culture technique d'un domaine qui est maintenant totalement intégré et accepté dans la société vers un nouveau domaine à explorer.

*« Au niveau des ensembles techniques du XX<sup>e</sup> siècle, l'énergétisme thermodynamique est remplacé par la théorie de l'information, dont le contenu normatif est éminemment régulateur et stabilisateur [...] La machine, comme élément de l'ensemble technique, devient ce qui augmente la quantité d'information, ce qui accroît la néguentropie, ce qui s'oppose à la dégradation de l'énergie : la machine œuvre d'organisation, d'information, est, comme la vie et avec la vie, ce qui s'oppose au désordre, au nivellement de toute chose [...] La réalité technique, devenue régulatrice, pourra s'intégrer à la culture, régulatrice par essence [...] La technicité peut alors devenir un fondement de la culture. (Simondon, 1969, pages 15-16) »*

Il est *a priori* étrange que la même équation mathématique permette de modéliser deux phénomènes aussi distincts. Étrange mais pas impossible, certains modèles théoriques peuvent avoir des applications pratiques dans des domaines différents, comme par exemple l'étude des phénomènes turbulents qui peut être transposée pour modéliser et mieux comprendre aussi bien l'écoulement d'un fleuve, que d'un gaz ou d'une foule. Ce qui est plus intéressant dans cette constatation c'est que la théorie dominante du XIX<sup>e</sup> siècle est remplacée au XX<sup>e</sup> par une autre théorie, d'un domaine différent, pourtant basée sur une équation similaire. Mais la nuance qui rend la chose intéressante c'est que l'équation est utilisée à rebours. L'interprétation de Gilbert Simondon est intéressante sur ce point, elle met les deux théories en parallèle mais se dirigeant dans des directions différentes. Les technologies

du XIX<sup>e</sup> siècle ont été dominées par la thermodynamique et une équation, qui allait dans le sens de l'entropie et de la dégradation de l'univers. Dans cette phase, le progrès allait dans le sens de la conquête du monde. Le transfert vers une nouvelle réalité technique, celle de l'information et du XX<sup>e</sup> siècle a entraîné avec elle la même équation, mais utilisée à l'envers. La théorie de l'information, basée sur la négentropie va rendre les machines utiles au ralentissement de la dégradation de l'énergie et agir comme élément stabilisateur du monde.

## 5.2 Mémoire de lecture

Il est difficile d'aborder la mémoire de lecture sans aborder les processus de mémorisation eux-mêmes. Nous ne pouvons toutefois pas nous laisser entraîner dans un domaine trop vaste et qui dépasse de loin les ambitions de cette recherche. Rappelons simplement que notre contexte nous demande de nous concentrer sur la façon dont notre cerveau fait l'acquisition des idées qu'il trouve au cours de la lecture.

*« Le cerveau ne fait pas apparaître miraculeusement la conscience, l'intelligence, le comportement orienté, dans un univers stupide et aveugle, et dans un organisme aussi stupide que cet univers. Il fait déborder l'organisation, instinctive et intelligente, déjà à l'œuvre dans l'organisme, sur le monde extérieur. Il n'invente, en technique externe, que par la même invention organique par laquelle il s'est formé d'abord lui-même, et selon la même logique matricielle. (Ruyer, 1975, page 151) »*

A première vue, enregistrer les parcours de dizaines, voire de centaines de lecteurs n'est qu'un simple problème de base de données. Donc un problème

technique relativement simple à résoudre. Le cœur du problème se situerait dans la restitution de ces parcours.

Le simple affichage des paragraphes, dans l'ordre de leur appel, sera très difficilement interprétable.

Soit nous obtiendrons un texte d'une longueur rendant difficile l'analyse du parcours parce qu'il s'étalerait sur un document trop long.

Soit nous afficherions les paragraphes séquentiellement, comme dans un diaporama et là encore le parcours dans sa globalité resterait caché.

Pour afficher un seul parcours, on pourrait imaginer de n'afficher qu'un graphe représentant les paragraphes par leurs numéros, ce qui reste un simple problème technique à résoudre.

Le problème le plus intéressant à résoudre est de synthétiser dans une seule page, dans une seule représentation la somme de tous ces parcours. Le but n'étant pas de permettre de refaire tous les parcours, mais plutôt d'en comprendre les motivations, les attentes et surtout d'en restituer le sens.

## **A. Le réseau hypertextuel doté de fonctions d'apprentissage**

Avec un système hypertextuel dont l'architecture repose sur un agent logiciel que nous voudrions pouvoir qualifier d'intelligent, nous devons nous poser la question des moyens à mettre en œuvre pour que l'agent puisse être doté de fonctions d'apprentissage. L'idée qui est derrière cela est que l'agent puisse s'enrichir des parcours faits par les lecteurs en ligne. Il faut donc que ces parcours soient enregistrés et qu'au moment de la lecture ces informations soient accessibles à l'agent pour qu'il puisse en tenir compte. Le fait de devoir enregistrer des informations dans un système d'information se traduit généralement par l'ajout d'une base de données pour les stocker, mais ce modèle est loin d'être le plus efficace dans notre situation où les parcours

enregistrés devront être rendus accessibles à l'agent hypertextuel au cours de la lecture. En prenant un peu de recul par rapport au dispositif qui a besoin d'être mis en place, nous pouvons identifier que sa fonction est de renforcer les liens entre les paragraphes d'un même parcours, dans l'espoir de voir un « méta-parcours » émerger après un certain nombre d'expérimentations.

Cela ressemble à ce que font les fourmis pour construire un parcours entre la fourmilière et une destination où trouver de la nourriture par exemple. Chaque fourmi qui passe dépose des phéromones. Plus il y a de fourmis qui empruntent un parcours et plus ce parcours va être empreint de ces phéromones, ce qui va amener d'autres fourmis à emprunter ce même parcours. Est-ce qu'il est vraiment judicieux de mettre en place un dispositif basé sur le même principe ? C'est ce que nos expérimentation devraient pouvoir permettre d'identifier. De ce fait, les paragraphes les plus lus seront ceux qui vont être mis en avant et qui devraient être encore plus lus, dans ces conditions certains parcours peu fréquentés risquent de disparaître de la visibilité du lecteur non averti. En revanche, tous les lecteurs capables de mettre en place une lecture savante sur un thème particulier seront à même de construire de nouveaux parcours pour eux.

L'idée de base construite à partir du modèle des fourmis a trouvé les moyens de se réaliser à travers une approche simplifiée de réseau neuronal. Si nous assimilons chaque nœud du réseau hypertexte comme un neurone élémentaire, nous devons permettre à ce neurone de renforcer ses liens avec ses voisins les plus pertinents. En pratique, cela voudra dire que les métadonnées qui viennent s'ajouter au texte de chaque paragraphe vont devoir comporter des informations concernant les liens à renforcer. C'est donc un modèle où chaque nœud du réseau contient les informations nécessaires à l'agent hypertextuel pour construire une représentation du paragraphe,

à la fois concernant le texte mais aussi concernant les liens hypertextuels. Physiquement, les nœuds dans le système *Co<sup>L</sup>i<sup>S</sup>ciences* sont les paragraphes des ouvrages, codés suivant un balisage XML. À travers sa DTD, ce balisage a tout d'abord été mis au point pour assortir le texte de métadonnées concernant les notions travaillées dans le paragraphes. En étendant les fonctionnalités décrites dans la DTD, il est possible d'ajouter un balisage spécifique pour enregistrer les parcours de lecture. Ainsi, chaque paragraphe, c'est-à-dire chaque nœud du réseau hypertextuel porte en lui le texte mais aussi toutes les informations concernant l'ouvrage où apparaît ce texte, assorties des informations concernant les notions traités par l'ouvrage, et finalement des informations concernant les parcours. En fait, chaque nœud ne connaît que ses « voisins de parcours », c'est le moteur hypertextuel qui à partir de ces informations locales sera capable de reconstruire une vision globale des parcours empruntés par les lecteurs.

Il peut paraître étranger d'attendre un comportement global du système hypertextuel à partir d'informations locales à ses nœuds, c'est justement parce que ce comportement est basé sur une approche connexioniste, comme celle qui est mise en place pour la construction de systèmes à réseaux de neurones.

*« Dans le cadre de l'approche connexioniste[...] le sens n'est pas enfermé à l'intérieur de symboles : il est fonction de l'état global du système et reste lié à l'activité générale dans un domaine donné. (Varela, 1996, page 79) »*

C'est donc bien l'activité générale de lecture qui amène le système à altérer son état global, faisant émerger le sens construit par les parcours savants. Mais le sens n'existe ni n'est enregistré dans aucun de ses nœuds, c'est leur mise en relation à travers le réseau hypertextuel qui lui donne naissance.

*« Le sens, toutefois, ne réside pas dans ces constituants en soi, mais dans les schémas d'activité complexes émergeant d'une interaction entre plusieurs d'entre eux. (Varela, 1996, page 80) »*

## **B. Restitution des parcours**

Une autre difficulté qui doit être surmontée est celle de la restitution des parcours. En effet, cette restitution doit permettre de rendre compte à la fois de la spatialité (dans quel ouvrage je me trouve, quelles interrogations je mets en place et quelle représentation je choisis pour y répondre : fac-similé, plein texte ou notions) et aussi de la temporalité (combien de temps je reste sur un paragraphe, est-ce pour le lire ou pour sélectionner le saut vers le paragraphe suivant?). Il est donc relativement complexe d'imaginer un dispositif qui rende compte à la fois de l'espace et de la dynamique. Nous pourrions imaginer un dispositif qui fasse refaire le parcours en affichant les pages successivement comme elles ont été visionnées par le lecteur (ou la moyenne des lecteurs, puisqu'il faut rendre compte de cette valeur pour rendre visible son usage et à travers lui sa pertinence, pour autant que nos hypothèses soient justes). Mais de cette manière, nous retrouvons dans la restitution du parcours toutes les problématiques de désorientation et de perte de contrôle que nous avons déjà évoquées pour la lecture. La restitution du parcours devra s'appuyer sur un modèle de représentation qui va permettre au lecteur de prendre du recul pour que son œil puisse l'englober dans sa totalité et par là pour qu'il puisse percevoir le parcours dans sa globalité, à travers un ou plusieurs ouvrages.

## C. Modèle de représentation

La construction du parcours devrait donc passer par une phase de restitution graphique qui aurait pour but de « mettre à l'échelle » cette progression et de permettre sa représentation sur une page écran, ou au moins un nombre limité de celles-ci. Le terme de restitution graphique est sans doute vague, mais cela est à dessein puisque nous n'avons pas *a priori* de contrainte particulière si ce n'est celle technique sur lesquelles nous allons revenir rapidement. Cela peut donc être soit un affichage sous la forme de diagrammes et de schémas, soit sous une forme colorée pour rendre compte des durées par des différences de teinte par exemple, soit une combinaison de tout cela. Dans tous ces cas nous sommes dans un système qui sera difficilement abordable par des mal voyants, alors que la lecture électronique était au premier abord un moyen de leur rendre accessible des textes par l'utilisation de navigateurs adaptés à leur handicap et que cela doit faire partie des objectifs d'un projet comme *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*.

Parce qu'il faut aussi parler de contraintes techniques, il faut aussi considérer le fait que nous sommes sur le Web et que nous avons une volonté affichée de pérennisation et d'accessibilité. Il faut donc que cette représentation graphique puisse se faire sur n'importe quel navigateur et en respectant les standards officiels du Web, ceci dans des conditions de réalisation technique à notre portée, ce qui est une limitation sans doute plus triviale mais non moins incontournable.

*« Une représentation consiste à faire apparaître à l'esprit un contenu.*

*Les représentations sont produites et liées entre elles selon des modes réglés, dont je dirais qu'ils forment les modélisations primaires du savoir : le mode analytique, le mode symbolique et le mode pratique. [...] Ainsi trois modes fondamentaux : pratique, analytique, symbolique ; et trois*

---

*fonctions constitutives : cognition, substitution, délégation. (Reichler, 1989) »*

Nos réflexions ont pris corps dans un dispositif particulier du site *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, que nous avons baptisé *lectochromie* et que nous détaillerons un peu plus loin, dans la section 7.4, page 151 du chapitre 7 **Représentation des parcours de lecture**.

## 5.3 Le sens

### A. Cadre de référence théorique

Pour étudier la construction du sens, il faut bien entendu que nous ayons défini dans quel cadre théorique nous souhaitons nous placer pour tenter de préciser ce qu'est le concept de sens dans notre recherche. Nous pouvons trouver une mise au point entre les diverses définitions du mot « sens » dans (Sabah, 2000), qui présente l'avantage de faire ce travail dans l'optique de l'appliquer au traitement automatique des langues et donc de partager nos préoccupations. C'est en particulier l'étude des sémantiques et formalismes de représentation utilisés dans le cadre de la traduction automatique qui peut nous donner des clefs utilisables dans le contexte de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* et de la navigation notionnelle, plus précisément dans le cadre de la sémantique procédurale. Dans ce type de système le sens est la description des actions (dans un système hypertextuel ce seront les liens) à réaliser pour trouver un objet correspondant à un mot (ce qui correspond pour nous au document cible du lien). La sémantique lexicale qui cherche à caractériser le sens des mots sur la base des distributions effectivement rencontrées devrait aussi donner des résultats intéressants compte tenu de la taille de notre corpus.

Il est tout de même important de relever que nous sommes dans un contexte très fragile, puisque la question de la validité même de ce type de recherche est posée :

*« Mais même en admettant que l'on sache de quoi on parle, l'« automatisation » entraîne une autre question difficile (insoluble par principe, ont affirmé certains philosophes...) : une machine peut-elle traiter du sens ? (Sabah, 2000, page 79) »*

Il nous faudra donc commencer par accepter le postulat que nos machines étant capables de traiter des informations, sont aussi capables de traiter du sens que nous avons mentionné avant de rappeler la théorie de l'information de Shannon. Nous devons toutefois garder à l'esprit le contexte qui a révélé cette théorie et ne pas en attendre plus qu'un cadre de réflexion général.

## **B. Créer du sens, pour qui, par qui ?**

Parce qu'un ordinateur est devenu aujourd'hui un système d'information, il semble désormais admis que l'ordinateur est devenu une machine à produire de l'information. Nous n'irons peut-être pas jusqu'à dire qu'il la crée, mais il l'avale, la digère et la restitue. Ce faisant, grâce à ses capacités de stockage virtuellement infinies depuis que les réseaux se sont tous unis dans l'Internet planétaire, grâce à la computation et au croisement de ces informations, grâce à tous les traitements que les techniciens s'ingénient à mettre au point pour améliorer l'analyse et la prise de décision, grâce à tout cela l'ordinateur est devenu au XX<sup>e</sup> siècle l'outil de gestion du sens par excellence.

*« La machine [...] devient ce qui augmente la quantité d'information, ce qui accroît la néguentropie. (Simondon, 1969, page 5) »*

Mais il ne résout pas tous les problèmes, il n'est même pas capable d'en résoudre aucun, il ne fait qu'assister l'homme et le rendre plus efficace. Il met surtout à notre disposition une grande capacité de mémorisation et de création de liens, sorte de béquille cognitive qui nous permet d'aller plus loin. Pour l'instant, il n'est pas capable de juger de la pertinence du stockage de telle ou telle information, ni de la création d'un lien logique entre deux informations. Il commence à émettre des opinions, demain il pourra sans doute donner des avis pertinents sur la validité de telle ou telle information. Mais de toutes façons, il ne sera capable de la faire qu'avec les règles que nous lui aurons fournies au départ, il risque de lui manquer pendant longtemps la règle ultime qui est la capacité que nous avons de nous remettre en question.

*« Le sens ne naît pas de la mise en relation de tout avec tout. Par contre, le sens émerge de la mise en relation « d'objets cognitifs » dans un contexte. La possibilité de faire appel à des « informations connexes » peut perturber la construction du sens, distraire dans un parcours d'acquisition des connaissances. La difficulté de l'écriture délinéarisée réside, donc, dans le fait que l'on doit, par rapport à un contenu, premièrement prévoir des cheminements dans des unités globales de sens. Secondement, structurer les contenus en unités sémantiques homogènes et poser des relations pertinentes entre ces blocs pour construire le sens par « touches ». Le sens final, d'un parcours au sein d'un hyperdocument, naît de l'enchaînement des pages. Si une page, actionnée à la suite d'une autre, coupe le processus de compréhension, le lecteur désorienté dans la construction de sa lecture, va perdre le fil. Dans ce cas, il se produit une « interaction de tangentialisation » : la réponse faite ne prend pas en compte ce qui vient d'être dit. À l'image de la communication interpersonnelle, le lecteur est disqualifié par la réponse du dispositif qui*

*ne correspond pas à sa demande. Le système se met à parler d'autre chose.*

*(Jullia, 2001, page 6) »*

Quand un lecteur parcourt *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* et qu'il rencontre une notion sur un paragraphe, nous pourrions être tenté de penser qu'il se passe un phénomène de création de sens. La lecture du paragraphe associée à la mise en évidence de la notion, pourrait faire émerger une idée dans l'esprit de celui qui lit. Nous sommes en droit de nous demander si ceci participe de la création de sens, et si c'est bien le cas, par qui et comment se crée le sens? Les textes de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* sont mis en œuvre par des moyens informatiques, suivant une méthode que l'on dit formelle parce qu'elle ne traite que de la forme de ce qui lui est fourni. En fait, pour reprendre les termes de Jacques Arzac (Arsac, 2003, page 220), nous sommes en train de nous confronter au « mur du sens ». Il n'y a dans *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* aucune réelle signification en jeu, mais simplement la transcription des actions du lecteur, qui pointe au moyen de la souris des icônes ou des bouts de texte, ce qui se traduit par la sélection d'un fichier XML représentant le paragraphe à lire. Le fichier XML est décodé, c'est-à-dire que des suites de caractères sont alignées pour former des mots, certains de ces mots vont représenter des balises XML et un contenu « sémantique » va leur être attribué. En particulier, c'est au travers de ces balises que des notions vont être attribuées au paragraphe. Par convention dans *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, le même mot trouvé en dehors des balises appropriées ne représentera pas la notion. En particulier, il ne sera pas affiché de la même manière en temps que notion et en temps que mot. Pour le lecteur, la présentation doit l'amener à penser la notion de manière différente de la manière dont il va penser le mot. Le mot est inclus dans le flot du texte sans mise en valeur, la notion est affichée dans un cadre réservée et mise en évidence. Les notions lui sont présentées sur la page d'entrée dans le corpus et il pourra décider de suivre une notion

comme l'on suit un guide, alors que s'il décide de chercher un mot et de suivre plutôt ce type de cheminement, il restera celui qui choisi le chemin. En effet, dans le cadre de la recherche de mot, il va lui être présenter une liste d'occurrences du mot, sous la forme d'extraits de texte présentant le mot en situation, c'est toujours à lui de choisir quel fil il veut démarrer et pour poursuivre la recherche il devra toujours revenir à son point de départ, c'est-à-dire à la page de recherche plain texte.

Ce qui a créé du sens, c'est la convention de lecture qui est utilisée entre *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* et son lecteur. Dans un cadre c'est *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* qui propose au lecteur de le piloter, dans l'autre *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* ne fait qu'aider le lecteur à se construire un parcours personnel.

Il nous reste peut-être une dernière question à répondre, il s'agit de savoir pourquoi avoir attendu l'hypertexte pour cette démarche ? En fait il faudrait être sûr d'avoir attendu, ce qui n'est pas certain du fait que ce type de parcours a déjà été proposé dans les encyclopédies. Nous pouvons en revanche de manière plus certaine affirmer que, à notre connaissance, ceci n'a pas encore été fait en dehors du contexte des ouvrages encyclopédiques.

*« Le support du papier restreint et fige les possibilités d'agencement des éléments du texte. Même si une page peut être parcourue en divers sens, même si un livre peut être feuilleté dans le désordre, la matérialité du dispositif de lecture offert par le livre limite sérieusement les possibilités de vagabondage. Certains auteurs ont tenté d'imaginer des dispositifs plus souples, mais il a fallu attendre le texte numérique pour voir voler en éclat l'ordre immuable du texte et apparaître de nouvelles possibilités. (Clément, 1995) »*

Il semble qu'au départ, c'est la machinisation du livre qui a permis de le doter d'un dispositif qui rend possible cette navigation fluidifiée et automatisée.

## C. Navigation et recherche d'information

Avec l'hypertexte le texte n'est plus statique et prisonnier dans la page imprimée, toutes sortes d'accès croisés sont rendus possibles. Les liens peuvent porter à la fois sur les notes en bas de page, les notes en fin de volume, l'index et les illustrations, la confrontation des contextes d'utilisation, la prononciation... Au delà, il se pourrait même que certains modes de lecture ne se fassent qu'au travers de ces liens : ce serait, par exemple, le cas pour une documentation technique qui n'est consultée que lorsqu'une panne se présente pour accéder, au plus vite, à l'information requise, sans qu'aucune lecture linéaire ne s'en fasse jamais, dans ce contexte la lecture s'apparente à une navigation à travers un corpus d'informations. Cependant, les questions demeurent : quels liens hypertextuels sont pertinents pour telle ou telle application ? Comment faire pour que l'usage de ces liens ne distraie pas le lecteur de sa lecture en le condamnant à être un éternel papillon esclave de sa curiosité et de sa velléité ?

La navigation utilisée dans le contexte de la recherche d'information est un sujet d'étude qui a conduit à la réalisation de nombreux prototypes qui n'ont pas toujours été complètement validés, comme nous le fait remarquer Georges Vignaux dans un document interne qui a été repris et mis en ligne sur *CoLiSciences* :

*« Les études qui s'intéressent à la navigation comme moyen d'accès à l'information insèrent cette activité dans un grand volume de documents tels ces systèmes où la couche navigationnelle est soit un thésaurus soit un réseau de concepts. Les recherches qui s'attaquent à de telles collections grandeur nature sont essentiellement consacrées au développement de prototypes pour l'interrogation. La validation expérimentale de ces prototypes, l'appropriation qu'en réalisent les usagers, les démarches*

---

*d'interrogation qu'ils induisent ne sont que très rarement abordées.*

*(Vignaux, 2002), (Vignaux, 2001) »*

De plus, les recherches sur les types de navigations (« butinage » ou élaboration de requêtes) font souvent abstraction de la couche physique du réseau hypertextuel pour n'étudier que la couche logique.

*« En fait, un parcours des travaux consacrés à la navigation pour la recherche d'information conduit à cette remarque générale : la focalisation sur la navigation dans la couche conceptuelle ne doit pas faire oublier qu'il existe un second niveau de navigation, celle qui s'effectue entre les documents. (Vignaux, 2002), (Vignaux, 2001) »*

## **D. Représentation de la connaissance**

La représentation de la connaissance est une problématique partagée par plusieurs disciplines, chacune proposant un modèle de solution conforme au cadre dans lequel elle évolue.

### **1. L'intelligence artificielle**

La porte est laissée ouverte dans *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* pour implémenter un dispositif qui ferait usage de l'I.A., bien que cela n'a pu encore être mis en œuvre, ceci de deux manière distinctes mais complémentaires. D'une part, l'architecture technique basée sur un moteur hypertextuel qui gère l'affichage des pages telles que le lecteur les demande permet d'introduire à ce niveau des règles de décisions telles que celles que nous trouvons dans un système expert ou un moteur d'inférence. Le moteur hypertextuel devrait alors analyser le comportement du lecteur, le parcours qu'il suit et les appels qu'il fait aux paratextes, pour décider d'afficher une page avec plus ou moins de

commentaires, voire de proposer des compléments d'informations ou d'autres pistes à suivre. D'autre part, nous avons la possibilité d'ajouter au niveau du corpus des informations et de créer de nouveaux liens entre les nœuds de l'hypertexte. Le dispositif que nous avons baptisé lectochromie et qui est étudié page 151 est une première étape dans la constitution du corpus sous la forme d'un réseau de type neuronal, qui s'auto-enrichit des lectures que les internautes font.

## 2. Le Web Sémantique

Le Web Sémantique est un des futurs possibles pour l'organisation des connaissances sur la toile, son acceptation par les industriels n'est pas acquise et même si des prototypes sont déjà opérationnels dans les laboratoires du W3C, rien n'est encore réellement visible du public. Nous verrons que les recherches menées à l'occasion de la mise au point du Web Sémantique ont toutefois abouti à la mise au point de nouveaux formats de données et de quelques dispositifs qui préparent la nouvelle voie.

**Les limites du Web actuel** Nous avons déjà présenté dans la section 1. *Critique du Web comme système hypertextuel*, page 83 du chapitre 4, comment le système hypertextuel du Web fonctionne. Comme nous l'avons dit au même endroit, le Web tel que nous le connaissons aujourd'hui permet de mettre en relation un très grand nombre de documents, sans doute plus de 500 milliards<sup>110</sup>. En revanche la technologie Web actuelle, c'est-à-dire la technologie de mise en œuvre du système hypertextuel du Web, ne permet pas de contrôler ni de gérer ces relations.

---

110. Voir pour cela la définition du « Deep Web » page 84 et comment est fait ce recensement.

« Associés [...] à l'Internet, les hypermédias rendent disponibles des quantités impressionnantes d'informations sur à peu près tous les sujets possibles et imaginables. Mais l'information n'est pas connaissance, et disponible ne veut pas dire accessible. Le chaos informationnel ne génère rien d'autre que du bruit cognitif (Rouet, 2000, page 16) »

Le Web sémantique est une solution proposée par l'inventeur du Web, Tim Berners-Lee, dans un article paru dans le *Scientific American* pour répondre à certaines de ses limites comme nous avons commencé à les évoquer dans la section 4.1 consacrée à l'hypertexte. En abordant la toile comme un des formats de lecture électronique, nous avons déjà identifié la quantité d'informations « brutes », non qualifiées comme une limite à son utilisation. Cette masse d'information participe en effet largement à la désorientation et à la surcharge cognitive qu'induit le besoin de se situer dans son parcours en même temps que de remettre en question la pertinence de ce qui nous est offert. Mais la solution de Tim Berners-Lee va sans doute au delà car elle permet surtout de faire intervenir des agents électroniques pour assurer la médiation entre l'internaute et la toile. C'est donc à l'agent que reviendrait la tâche de l'orientation et de l'attribution du degré de pertinence.

**Présentation** Nous ne devrions pas pouvoir trouver de meilleure présentation du Web Sémantique que celle faite par Tim Berners-Lee, qui en est l'inventeur.

« Le Web sémantique est ce que nous obtiendrons si nous réalisons le même processus de globalisation sur la représentation des connaissances que celui que le Web fit initialement sur l'hypertexte. (Berners et coll., 17 mai 2001) »

C'est modestement l'approche que nous avons essayé de mettre en place dans *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, pour cette raison l'étude du modèle du Web Sémantique peut nous apporter des idées, voire des modes d'implémentation qui suivraient les recommandations du Web Sémantique et prépareraient *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* à cette nouvelle médiatisation du texte et des idées.

Le défi du web sémantique est de fournir un langage qui exprime à la fois des données et des règles pour raisonner sur les données et pour que les règles de n'importe quel système de représentation de la connaissance puisse être exportées sur le web. La logique doit être assez puissante pour décrire les propriétés complexes des objets, mais ne doit pas empêcher les agents de résoudre un paradoxe. Deux importantes technologies de développement du web sémantique existent déjà : eXtensible Markup Language (XML) et le Resource Description Framework (RDF). XML permet à chacun de créer ses propres balises – marques cachées qui mettent des annotations sur les pages web ou les sections de texte d'une page. XML permet aux utilisateurs d'ajouter une structure arbitraire à leurs documents sans rien dire de la signification des structures (Bosak, Bray, 1999). Le sens est donné par RDF, qui le code dans des ensembles de triples, chaque triple jouant le rôle du sujet, du verbe et de l'objet dans une phrase élémentaire. Les triples de RDF constituent des webs d'information sur des choses qui sont en relation avec le document.

**Ontologies** Deux bases de données peuvent utiliser des identifiants différents pour ce qui est en fait le même concept, le code postal par exemple. La solution à ce problème se trouve dans le troisième composant basique du web sémantique : l'ensemble d'informations appelé ontologies. En philosophie, une ontologie est une théorie à propos de la nature de l'existence, des types de choses qui existent. L'intelligence artificielle et les chercheurs du web ont

---

adopté ce terme, et pour eux une ontologie est un document ou fichier qui définit de façon formelle les relations entre les termes. Un type d'ontologie caractéristique du web possédera une taxonomie et un ensemble de règles d'inférence. La taxonomie définit des classes d'objets et les relations entre eux. Par exemple, une adresse peut être définie comme type de lieu, et les codes postaux de ville peuvent être définis pour s'appliquer seulement à des lieux, et ainsi de suite. Les classes, les sous-classes et les relations entre les entités sont un puissant outil pour utiliser le web. On peut exprimer un grand nombre de relations entre les entités en attribuant des propriétés aux classes et en permettant à des sous-classes d'hériter de leurs propriétés. Si les codes postaux sont de type ville et sachant que les villes ont généralement des sites web, on peut associer un site web au code postal d'une ville même si aucune base de données ne relie directement un code à un site.

Cela revient donc à définir un vocabulaire et des hiérarchies dans ce vocabulaire. Il faut cependant noter la différence avec le fait de créer uniquement un vocabulaire XML comme c'était déjà possible de le faire sans pour autant faire appel au langage des ontologies. Le guide dit :

*« Une ontologie diffère du schéma XML en ce qu'elle est la représentation d'une connaissance, et non pas un format de message. Beaucoup de standards Web basés sur l'industrie consistent en une combinaison de formats de message et de spécifications de protocole. On a donné une sémantique opérationnelle à ces formats, comme par exemple, A la réception de de message BonDeCommande, transférer Montant euros de CompteDebit vers CompteCredit et livrez Produit. Mais la spécification ne permet pas de construire un raisonnement en dehors du contexte de cette transaction. Par exemple, nous n'aurons généralement pas de*

*mécanisme pour conclure que parce que le produit est un type de Chardonnay il devrait aussi être un vin blanc.*<sup>111</sup> »

Les règles d'inférence dans les ontologies sont encore plus puissantes. Une ontologie peut exprimer la règle suivante : « Si un code postal de ville est associé à un code d'état et qu'une adresse utilise ce code de ville, alors cette adresse est associée au code de l'état. » Un programme pourrait en déduire, par exemple, que l'adresse de l'université Cornell, à Ithaca, doit se trouver dans l'état de New-York, qui est aux États-Unis, et devrait donc être formatée selon les standards américains. L'ordinateur ne comprend pas vraiment cette information, mais il peut maintenant manipuler les termes de manière beaucoup plus efficace et significative pour un usager humain. Avec les pages d'ontologie sur le web, les solutions aux problèmes de terminologie (et autres) commencent à voir le jour. La signification des termes ou codes XML peut être définie par des pointeurs de la page vers une ontologie. Des applications plus avancées utiliseront des ontologies pour associer l'information d'une page à des structures de connaissance et à des règles d'inférence. Ainsi avons-nous pour objectif de modéliser des règles locales d'inférence conduisant à des ontologies partielles constitutives d'ontologies générales. L'observation et l'analyse des stratégies de lecture est essentielle pour affiner les processus de construction hypertextuelle. Notre projet *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* entre dans une nouvelle grande étape : ajuster la construction hypertextuelle à un modèle des lectorats et des navigation selon les publics et les destinations.

---

111. An ontology differs from an XML schema in that it is a knowledge representation, not a message format. Most industry based Web standards consist of a combination of message formats and protocol specifications. These formats have been given an operational semantics, such as, Upon receipt of this *PurchaseOrder* message, transfer *Amount* dollars from *AccountFrom* to *AccountTO* and ship *Product*. But the specification is not designed to support reasoning outside the transaction context. For example, we won't in general have a mechanism to conclude that because the *Product* is a type of *Chardonnay* it must also be a white wine.

Le langage OWL *Web Ontology Language*<sup>112</sup> est en cours de mise au point dans les laboratoires du W3C. Le but est de définir un langage qui puisse être utilisé pour décrire les classes et les relations qui les lient et qui sont inhérentes aux documents et applications du Web.

### Les langages

- Langages d’assertion et d’annotation
- Langages de définition d’ontologies
- Langages de description et de composition de services

Il faudrait sans doute plutôt aller vers la construction d’un *Web Cognitivement sémantique*.

### 3. Métadonnées et annotations

L’enjeu des métadonnées est double. En premier lieu, il s’agit de tirer pleinement parti des possibilités qui sont offertes, pour faire participer *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* dans le réseau planétaire. Il est par exemple possible d’utiliser le Dublin Core<sup>113</sup> pour décrire les ouvrages en ligne, leurs auteurs, les éditions, etc. Ces informations étant mises à disposition des autres sites intéressés par le corpus de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, cela va leur permettre de citer les ouvrages bien entendu, mais aussi de présenter les textes de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* dans un autre contexte qui leur serait particulier et de permettre ainsi aux internautes de naviguer d’un site à l’autre à travers le Web avec plus de souplesse. En second lieu, il s’agit aussi de créer de nouvelles métadonnées qui seront spécifiques à *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* et nous permettrons d’innover pour créer une bibliothèque électronique très différente de celles qui existent déjà. *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* n’est pas

112. <http://www.w3.org/TR/owl-guide/>

113. Que nous avons présenté avec les DTD de l’information–documentation dans le paragraphe D., page 89, de la section 4.4 *Gestion électronique des documents*

seulement la bibliothèque, c'est aussi le bibliothécaire et ses conseils, armé des index et des catalogues. Pour que cela soit possible, il faut bien entendu faire plus que mettre à disposition du texte, il faut créer des liens, des liens avec du sens, des liens qui construisent des parcours érudits et pas simplement qui fassent faire du « lèche-vitrine ». Toutes ces informations seront codées sous la forme de métadonnées, qui pourront être interprétées à la volée par le moteur hypertextuel pour enrichir l'expérience qu'il peut offrir au lecteur en reconstruisant des liens pertinents.

## 5.4 Sémantique interprétative

*« Le web ne peut être traité comme un espace neutre, offrant simplement la possibilité technique de mieux communiquer. Ainsi que l'ont montré F. Ghitalla et D. Boullier dans un ouvrage récent, les usages du web passent par la construction d'un « espace organisé d'activités » et cette organisation prend appui sur des formes, des catégories et des normes qui pré-existent au web. En s'appuyant sur le concept d'« isotopie » emprunté à François Rastier (Rastier, 1997), ces auteurs montrent que l'expérience antérieure permet aux internautes de développer des représentations du web indispensables à l'action. C'est à partir de ces représentations que naissent non seulement le désir d'engager le dialogue avec la machine, mais aussi l'aptitude à anticiper et à interpréter les événements qui se produisent au cours de l'interaction. (Camus-Vigué et Gaudet, 2004, page 25) »*

La théorie de François Rastier se veut avant tout une tentative de rationalisation des phénomènes complexes de sens. Il propose d'utiliser des sèmes, ou traits sémantiques élémentaires pour la description des signifiés linguistiques.

Ces signifiés, ou sémèmes sont attribués à des signifiants définis de façon souple, du morphème à la lexie (groupement stable de morphèmes). Au niveau local, un sémème sera une collection de sèmes, de statuts différents. Au niveau global, ces sèmes correspondent à des classes de sémèmes ou à des formes de structuration de ces classes. Le principe est donc de justifier toute attribution d'un sème à un sémème, et cette justification ne peut se faire que par la mise en relation du sémème à décrire avec un ou plusieurs autres sémèmes ; cette justification se fait donc de façon globale. Le sens découle alors d'un système, suivant le paradigme structuraliste.

*« Dans le domaine de la linguistique, les théories du texte les plus en vue restent rattachées à la problématique logico-grammaticale par deux biais principaux. Le premier résume la textualité à des phénomènes phrastiques qui s'étendent sur des phrases adjacentes (concordances de temps, anaphores) qui sont autant d'isotopies locales. Malgré leur intérêt, les recherches sur la macrosyntaxe et la sémantique de la période ou du paragraphe restent en-deçà du texte et de la textualité. À cette extension de la syntaxe répond une autre voie : celle de la réduction propositionnelle du texte. On en connaît le principe, illustré notamment par Van Dijk : après un codage des phrases en propositions, on supprime les propositions jugées secondaires, pour ne garder enfin qu'une proposition, dite macroproposition, censée représenter le texte . Ce format propositionnel permet la réduction du texte à ce que peut concevoir la problématique logico-grammaticale ; aussi a-t-il connu, en psychologie cognitive et en psycholinguistique, un immense succès. (Rastier, 2001) »*



CHAPITRE 6

**LE DISPOSITIF *CoLiSciences***

---

## Sommaire

---

<b>6.1</b>	<b>Introduction</b>	<b>257</b>
A.	Présentation et objectifs	258
<b>6.2</b>	<b>Description du système <i>C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences</i></b>	<b>261</b>
A.	Une base de connaissances (multimédia)	264
B.	Un espace de travail collaboratif	266
C.	Une bibliothèque virtuelle	267
<b>6.3</b>	<b>Le dispositif technique</b>	<b>268</b>
A.	Modèle de publication	273
B.	Le stockage de données XML	275
C.	Délinéarisation et découpage du corpus	281
D.	L'influence des réseaux de Petri	284
E.	Structuration	289
F.	Appropriation et apprentissage	290
G.	Les médiations autour de <i>C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences</i>	294
<b>6.4</b>	<b>Notre architecture</b>	<b>302</b>
A.	Numérisation	304
B.	Architecture serveur	306
C.	Le corpus disponible	307
D.	Construction de la DTD	308

---

## 6.1 Introduction

Le projet *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* a engendré un site Web sur lequel sont mis en ligne les ouvrages d'un corpus de naturalistes, comme nous l'avons déjà dit en introduction générale. Mais le site *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* est plus qu'un simple site Web, d'ailleurs ce terme imprécis ne représente pas précisément un objet mais plutôt une classe d'objets. Sous cette même terminologie, nous pouvons trouver une simple collection de pages statiques, c'est-à-dire écrites une fois pour toutes comme dans un ouvrage papier, ou alors un portail d'information qui va puiser ses articles dans différentes bases de données, voire sur d'autres sites. Nous l'avons appelé dans la section précédente *dispositif C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, il pourrait répondre de manière plus précise à la terminologie de *bibliothèque numérique*. Encore que cette terminologie plus précise soit aussi plus restrictive. Qu'est-ce qu'une bibliothèque ? Un bâtiment, des rayonnages, des ouvrages, mais aussi des catalogues, un personnel pour enrichir ces catalogues et en faciliter l'accès. Dans le cas de la bibliothèque physique il peut exister une certaine imprécision, que nous pourrions retrouver sous le terme de bibliothèque numérique. Or, dans le cas de la bibliothèque numérique *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, c'est bien la bibliothèque au sens élargi qui a été construite. Nous y trouvons bien entendu les ouvrages, mais aussi les glossaires et encyclopédies. Cette représentation de la bibliothèque comme institution par un objet technique a besoin d'être précisée.

« *L'objet technique est au point de rencontre de deux milieux et il doit être intégré aux deux milieux à la fois. (Simondon, 1969, page 52) »*

Les services que le site, ou plutôt que la bibliothèque *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* fournit à ses lecteurs sont présentés sur la figure 6.1, réalisée par Georges Vignaux et son équipe « *Hypertextes et textualité électronique* ». Nous voyons effectivement

sur ce schéma que ce n'est pas une accumulation de dispositifs mais bien des services, comme ceux que pourrait rendre un bibliothèque traditionnelle.

## A. Présentation et objectifs

Parmi les paratextes disponibles en ligne pour présenter *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* nous pouvons trouver ces explications sur cette architecture enrichie d'outils et qui est le centre de plusieurs préoccupations de recherche, ceci dans des directions quelquefois parallèles mais souvent plutôt complémentaires.

*« Les ambitions générales de ce programme sont au nombre de quatre :*

- 1. Culturelles et patrimoniales : Il s'agit de collecter, numériser et mettre à disposition un grand corpus des ouvrages et articles de biologistes et naturalistes du 19<sup>e</sup> siècle, en langue française (près de 6.000 pages sont déjà offertes en mode texte et fac-similé) : Claude Bernard, Armand de Quatrefages, Isidore et Etienne Geoffroy-Saint-Hilaire, George Romanes etc.*
- 2. Intellectuelles : Le choix de ces textes permet de retracer, en partie, une « histoire des idées » de ces textes, à savoir le développement durant cette période, d'une science moderne du vivant.*
- 3. Scientifiques au sens de la modélisation sémantique : L'architecture de ce site traduit les réflexions de l'équipe centrées sur la problématique de l'hypertexte. Par des liens hypertextuels, le lecteur peut, à partir du texte, accéder à : 1) un glossaire des termes scientifiques et techniques, 2) un répertoire des notions, 3) un index des noms de personnes et des ouvrages cités dans chaque texte. Des parcours de lecture, surtout, lui sont proposés grâce à l'établissement de liens*

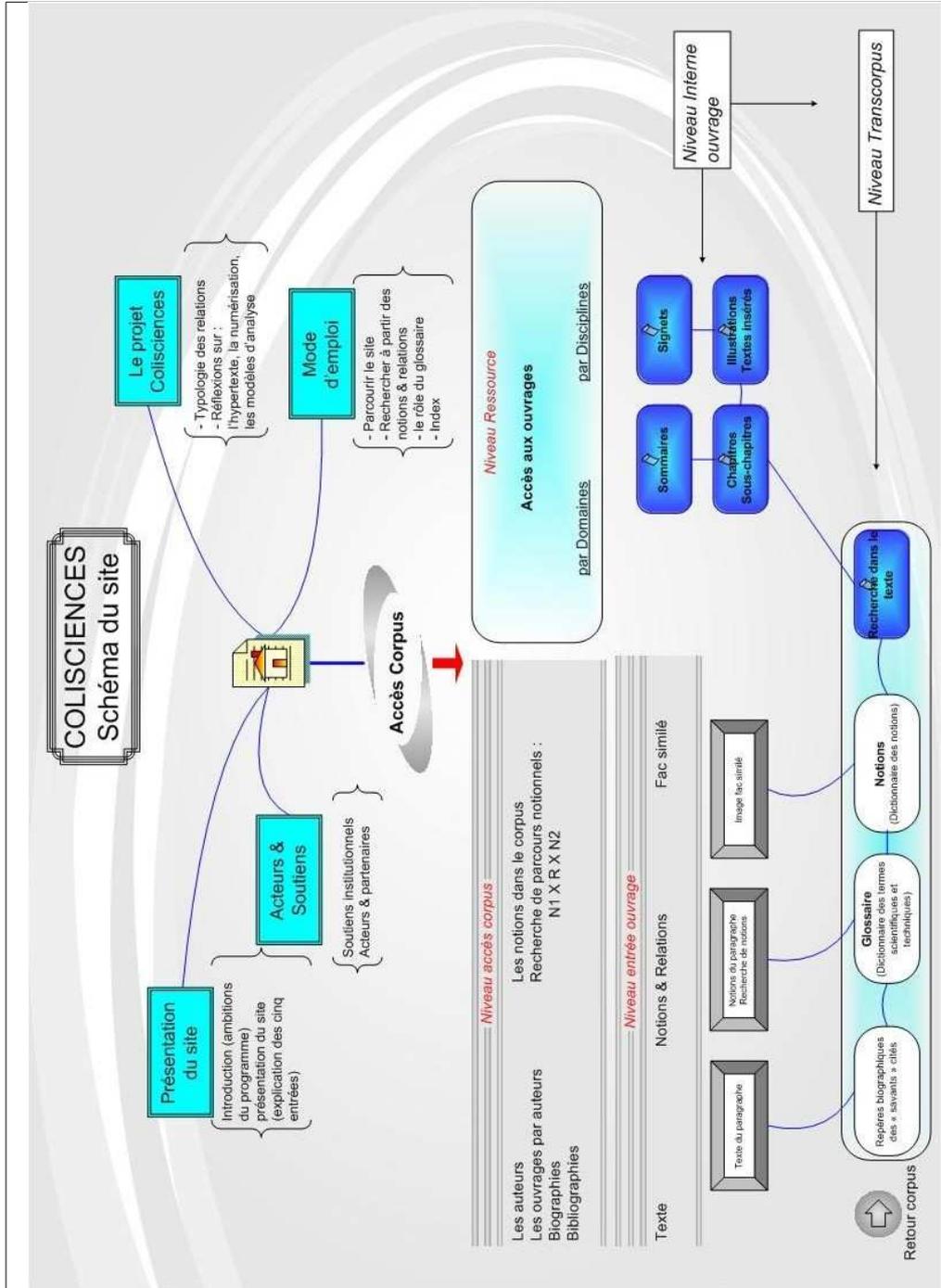


FIG. 6.1: Le plan du site  $Colisciences$

*hypertextuels exprimant les relations sémantiques et conceptuelles que les notions entretiennent entre elles au travers des textes.*

4. *Cognitives et pédagogiques : Une de nos problématiques centrales est celle de la lecture et de la navigation dans une double perspective :*
- 1) *Les modalités de la lecture vont-elles radicalement changer avec le support électronique ?*
  - 2) *Réciproquement, comment spécifier ces nouvelles conditions de l'offre de lecture pour l'apprentissage ?*
- Comment se construit le sens dans un hypertexte électronique (dans et à partir de) ?*

*(Vignaux, 2001) »*

Comme pour le livre électronique, notre dispositif hypertextuel propose une nouvelle façon d'aborder un ou plusieurs ouvrages. Les parcours dans *C<sup>o</sup>L<sup>i</sup>Sciences* peuvent se dégager complètement des contraintes physiques et structurelles du texte, c'est-à-dire qu'il est possible de *naviguer* à travers le développement d'une notion, ce qui demande d'étudier comment se fait l'appropriation de ces textes.

Le dispositif électronique peut construire une nouvelle matérialité qui est sans doute en grande partie à inventer ou peut-être plus simplement à réinventer. Le lecteur ne se repère plus par rapport aux pages écornées, voire souillées par de nombreuses lectures et relectures. La taille de l'ouvrage et la place du paragraphe en cours d'étude dans la globalité de l'ouvrage ne sont plus accessibles par le poids du papier ni son épaisseur. Elles sont quelquefois reconstruites par des artefacts, icônes ou autres graphismes.

En revanche, une nouvelle lecture est possible et un nouveau type de relation, d'interaction a-t-on l'habitude de dire pour les applications informatiques, se crée entre le lecteur et son texte. Comme le livre papier enregistre dans sa « chair » les passages du lecteur, il faut pour faciliter l'appropriation,

créer des outils ou des dispositifs pour enrichir automatiquement les usages qui ont été faits d'un ouvrage électronique. C'était notre volonté avec la *lectochromie* que nous avons présentée dans la section 7.4 du chapitre 7, page 151.

Une fois muni de ce dispositif, nous avons pu étudier jusqu'à quel point le système est autosuffisant et sur quelles parties l'intervention humaine devra en assurer l'adaptation. Pour l'instant il n'a pas été retenu de développer de fonctions d'analyse du texte pour proposer ou assister un chercheur dans la définition des notions. En revanche la *lectochromie* fonctionne de manière autonome et ne réclame que peu de réglages.

## 6.2 Description du système *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*

Le système *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* a été construit sur la base d'expériences précédentes, à partir d'un cahier des charges dont nous avons essayé de retirer toute contrainte *a priori*, nous pouvons dire que c'est un système *sur mesure*, un travail finalement plus proche de celui de l'artisan que de celui de l'ingénieur. Cela peut avoir un impact sur la manière de fonctionner du système, sur la qualité de finition de ses fonctions et cela en a aussi sur son intégration dans un paysage technique qui tend à se structurer autour de formes, de standards reconnus par l'usage ou par leur poids économique.

« Dans l'affrontement de la cohérence du travail technique et de la cohérence du système des besoins de l'utilisation, c'est la cohérence de l'utilisation qui l'emporte parce que l'objet technique sur mesures est en fait un objet sans mesure intrinsèque ; ses normes lui viennent de l'extérieur : il n'a pas encore réalisé sa cohérence interne ; il n'est pas

*un système du nécessaire ; il correspond à un système ouvert d'exigences.*

*(Simondon, 1969, page 24) »*

Nous avons déjà commencé à présenter le dispositif <sup>C</sup><sub>o</sub><sup>L</sup><sup>i</sup><sup>S</sup>ciences dans le cadre de l'introduction générale (page 1), il s'agissait alors d'introduire le dispositif hypertextuel tel qu'il a été élaboré et décrit dans notre mémoire de DEA (Augier, 2002). Plus concrètement, nous pouvons décrire <sup>C</sup><sub>o</sub><sup>L</sup><sup>i</sup><sup>S</sup>ciences par son contenu : ce sont près de 6 000 pages qui sont mises en ligne, empruntées à des éditions originales des textes de Claude Bernard, Ludwig Büchner, Étienne et Isidore Geoffroy Saint-Hilaire, Armand de Quatrefages, G.G. Romanes, Oskar Schmidt. Mais ce n'est qu'une première étape dans la numérisation. Par la suite, <sup>C</sup><sub>o</sub><sup>L</sup><sup>i</sup><sup>S</sup>ciences devrait pouvoir proposer une édition complète des livres de Claude Bernard (une quinzaine de titres), plus une sélection de ses articles, soit environ 5 000 pages supplémentaires. Tous ces ouvrages sont et continueront à être proposés en accès libre et gratuit. Pour ne pas être simplement une bibliothèque électronique, ces ouvrages sont assortis de tout ce qui peut en faciliter une lecture savante. Chaque auteur est présenté par : une courte notice biographique, une bibliographie de et sur l'auteur, une présentation ou réflexion sur ses travaux et leur portée. C'est ce que nous retrouvons dans la notice explicative en ligne sur <sup>C</sup><sub>o</sub><sup>L</sup><sup>i</sup><sup>S</sup>ciences :

*« Les textes sont dotés d'une « aide à la lecture », classique dans sa forme et dans son usage, mais impérative quand il s'agit de permettre au lecteur non-expert un accès « cognitif » qui ne soit pas rebutant, risque constant en raison de la difficulté même des lexiques engagés par de tels domaines de connaissance. Ainsi, nous réalisons un glossaire des termes scientifiques et techniques, glossaire auquel l'internaute peut recourir en cours de lecture. Le lecteur peut, grâce à de courtes définitions, saisir le sens usuel et/ou circonstanciel de tel ou tel terme exigeant un éclairage*

*spécifique. Ce qui est privilégié ici, c'est la possibilité d'une lecture le plus possible circonscrite à l'intérieur du site, sans la nécessité fastidieuse de recourir à des dictionnaires externes. (Vignaux, 2001) »*

Nous l'avons déjà évoqué plusieurs fois, *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* ne doit pas être réduit à un bibliothèque électronique, encore moins à un livre électronique en ligne. D'une part *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* est un système hypertextuel et d'autre part c'est un système communicationnel au cœur d'un réseau : Internet.

*« Ce corpus est entièrement inter-relié, via les notions-clés, en tant qu'elles sont des « moteurs sémantiques » permettant de suivre les usages et les transformations du vocabulaire scientifique et philosophique (à vocation cognitive), au sein des différentes représentations du vivant (telles que, par exemple : vie, méthode expérimentale, physiologie, raisonnement, etc.). Ce ne sont donc pas seulement des textes « bruts » qui sont ainsi rendus disponibles, mais des parcours d'exploration et de lecture qui sont proposés au travers de dispositifs de navigation permettant idéalement : (i) de trouver de la façon la plus économique en termes de temps et de « charge mentale » les informations souhaitées ou recherchées ; (ii) de « saisir » les idées contenues dans ces textes par d'autres moyens que les lectures linéaires impliquées par le dispositif « livre » habituel. (Vignaux, 2001) »*

La mise en place d'un tel dispositif peut quelquefois mener à un empilage de modules techniques qui, chacun pris isolément semblent très intéressants mais manquent de cohésion globale. Il faut généralement veiller à garder la vision globale de ce que le dispositif est sensé apporter et ne pas se laisser entraîner dans une course technologique qui n'a pas de fin, parce que la technologie évolue au moins aussi rapidement que le dispositif. Cette réflexion doit bien sur prendre en compte les motivations du lecteur, et plus précisément

comment va-t-il pouvoir articuler la navigation à travers les trois types de parcours, que nous avons déjà étudiés dans notre mémoire de DEA (Augier, 2002) :

- Parcours fac-similé
- Parcours texte plein
- Parcours notionnels

Penser le dispositif dans ses capacités éditoriales, mettre en ligne un corpus de manière pérenne et dans lequel il sera possible de réaliser des parcours notionnels va plus loin que l'automatisation de la construction des index et autres tables.

« [Les] tâches tout à la fois traditionnelles et nouvelles qui incombent aux bibliothécaires-médiateurs : recréer des contextes, reconstruire un environnement culturel et pédagogique pour tout ce qui est présenté, sélectionner des ensembles signifiants, créer tous les liens utiles à une bonne compréhension et propres à susciter des curiosités inédites. S'en tenir aux seules satisfactions immédiates de la technologie et à des productions sans envergure et sans lendemain reviendrait à rater une opportunité fantastique d'une diffusion large et intelligente des ressources patrimoniales. (Arot, 1998, page 125) »

## **A. Une base de connaissances (multimédia)**

Le site <sup>C</sup><sup>o</sup><sup>L</sup><sup>i</sup><sup>S</sup>ciences met à disposition de ses lecteurs un ensemble structuré de textes-sources et de textes-commentaires.

Les textes sont dotés d'une « aide à la lecture », classique dans sa forme et dans son usage, mais impérative quand il s'agit de permettre au lecteur non-expert un accès « cognitif » qui ne soit pas rebutant, risque constant en raison de la difficulté même des lexiques engagés par de tels domaines de

connaissance. Ainsi, nous réalisons un glossaire des termes scientifiques et techniques, glossaire auquel l'internaute peut recourir en cours de lecture, ce qui ajoute à la richesse hypertextuelle et permet une navigation plus étendue mais qui risque d'ajouter au sentiment naturel de désorientation que ressent généralement tout lecteur sur le Web. Le lecteur peut, grâce à de courtes définitions, saisir le sens usuel et/ou circonstanciel de tel ou tel terme exigeant un éclairage spécifique. Ce qui est privilégié ici, c'est la possibilité d'une lecture le plus possible circonscrite à l'intérieur du site, sans la nécessité fastidieuse de recourir à des dictionnaires externes.

Les « notions », quant à elles, sont des termes clés qui condensent la nature problématique des différentes parties d'un texte-source. Outre leur rôle particulier dans l'optique d'une hypertextualisation du corpus qui s'appuie sur elles pour en appréhender la trame conceptuelle, ces notions font également l'objet d'articles rédigés par l'équipe « *Hypertextes et textualité électronique* », articles dont la vocation est d'explicitier ces notions, en tant qu'elles signalent des moments spécifiques ou permanents de l'institution d'un domaine, d'un questionnement théorique, ou de controverses, etc., à l'intérieur des sciences de la vie. Dans la perspective d'une lecture de la portée sémantique du corpus en question, on distingue donc : (i) au niveau lexical, le glossaire et (ii) au niveau des idées, les « notions » constituées en noyaux générateurs de significations plurivoques. Une des premières « leçons » que l'on peut tirer de l'examen d'un corpus scientifique portant sur des auteurs et des sous-domaines variés –, mais concourant à l'élaboration d'un domaine scientifique de grande ampleur et tout de même unifié – est bien de montrer que l'abord d'une science, surtout dans une perspective historique, passe par l'exploration des champs sémantiques qu'exhibe le corpus.

## B. Un espace de travail collaboratif

Au départ les ambitions de développement d'un espace de travail collaboratif à travers *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* étaient assez nombreuses, les contraintes de la vie du projet n'ont pas encore permis de toutes les faire aboutir mais cela fait toujours partie de nos objectifs. L'idée sous-jacente derrière ce terme est de construire un système communicant qui mette en relation les lecteurs à travers lui, c'est-à-dire qu'il puisse susciter des mises en commun et des partages de points de vue ou d'idées au delà de ce qui pourrait être fait explicitement par les lecteurs, mais favoriser des échanges implicites. La différence entre les deux est un petit peu la même qu'entre un lettre et une bouteille à la mer. Dans le premier cas, l'émetteur de la lettre connaît son destinataire et lui rédige un message en fonction de cela, en étant presque assuré d'obtenir une réponse, dans le second le message s'adresse à un inconnu et il faut le rendre suffisamment explicite pour provoquer une réponse. D'ailleurs, pour commencer il faut s'être persuadé soit même de l'utilité d'envoyer une bouteille à la mer.

*« Avec l'usage des réseaux (par exemple Internet), nous nous trouvons dans un ensemble d'activités et de procédures par lesquelles les utilisateurs s'échangent des mots, des signes, des représentations... et nous définissons la communication comme le processus qui ordonne la production, la diffusion et l'appropriation des informations, des connaissances au sein d'un espace collectif. (Agostinelli, 1999, page 22) »*

Cet espace collaboratif correspond aussi à la nouvelle idée que l'internaute est en train de se faire de l'interactivité et de l'usage d'Internet plutôt comme espace de communication que comme espace de contenu. C'est ce que Pierre Lévy appelle *l'intelligence collective*.

*« L'interactivité qui renvoyait, il y a quelques années, à des situations de dialogue homme-machine évolue maintenant vers une « intelligence collective » (Lévy, 1994) »*

## C. Une bibliothèque virtuelle

La discussion autour de la bibliothèque virtuelle est très ambiguë, parce que *C<sup>o</sup>L<sup>i</sup>S*ciences est une bibliothèque virtuelle au sens d'une institution virtuelle qui propose des ouvrages numériques, donc dépourvus de matérialité<sup>114</sup>. Mais en fait nous pourrions aussi noter que *C<sup>o</sup>L<sup>i</sup>S*ciences prend forme dans le cyberspace et sous cet aspect se trouve en être un des éléments qui participe à sa construction. C'est donc finalement un élément bien réel dans un espace virtuel.

De part cette participation, *C<sup>o</sup>L<sup>i</sup>S*ciences contribue à donner, avec d'autres, une direction particulière à Internet. Nous reviendrons dans le dernier chapitre sur le projet Google qui vise à numériser en masse des ouvrages papiers pour les mettre en ligne. Ce projet paraît très ambitieux et innovateur, s'il est facile d'être d'accord sur le premier point, le second est discutable. Pour cela nous pouvons nous appuyer sur cette vision de la bibliothèque virtuelle :

*« Il est difficile de dire ce qu'il en sera dans quelques dizaines d'années. Mais si on admet que la « matière imprimée » a influencé, au cours de ces quatre derniers siècles, nos schémas mentaux, modifié notre rapport aux pouvoirs, par analogie on peut penser qu'il en sera de même de la « matière numérique », mais avec un coefficient accélérateur sans commune mesure. La bibliothèque du XXI<sup>e</sup> siècle ne peut plus être une*

---

114. Il s'agit en fait d'être dépourvu d'une certaine matérialité, le chapitre sur l'écrit numérique nous a déjà permis d'aborder ce sujet.

*accumulation de documents, elle sera un réseau organisé de connaissances*

*(Dupoirier, 1999, page 8) »*

Le projet Google n'est au départ qu'une accumulation de documents, il faut souhaiter que l'ajout des possibilités du moteur de recherche permettra d'atteindre le but du réseau organisé de connaissances, mais à ce stade nous ne pouvons pas en dire plus. En comparaison, *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* nous semble plus innovateur parce qu'au delà des documents qui ont été mis en ligne, le dispositif propose et crée des liens et du sens entre les textes.

### 6.3 Le dispositif technique

La construction du dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* est liée à l'aboutissement d'un certain nombre de projets, pour la plupart indépendants et œuvrant chacun à la réalisation de ses propres objectifs sans connaissance ni coordination entre eux. Notre travail de recherche utilise le résultat de ces projets, assorti de nouveaux outils pour mettre en forme le dispositif final. Pour essayer de présenter complètement la genèse du dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* nous devons prendre un peu de recul pour englober l'ensemble des dispositifs hypertextuels développés par l'équipe « *Hypertextes et textualité électronique* », ils font tous partie de la même lignée qui a contribué à faire de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* ce qu'il est aujourd'hui.

*« L'individualité des objets techniques se modifie au cours de la genèse. (Simondon, 1969, page 19) »*

Le dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* est l'individualisation présente d'un dispositif technique qui a déjà existé sous la forme de plusieurs individualités différentes. Pour revenir aux recherches de l'équipe « *Hypertextes et textualité électronique* », le premier dispositif qui a été mis au point consistait en un logiciel autonome, distribué sous forme de CD-ROM, contenant uniquement

*l'Introduction à la Médecine Expérimentale* de Claude Bernard et mettant en œuvre des parcours notionnels à travers cet ouvrage. Ce dispositif n'avait donc pas de possibilité de collaboration ou de partage d'information entre les lecteurs, en revanche l'ergonomie en avait été profondément étudiée et un accent tout particulier avait été mis sur la visualisation des parcours notionnels. Avec l'apparition d'Internet et du Web, c'est dans la direction de la mise en ligne vers un plus grand nombre de lecteurs que le nouveau dispositif a été pensé. Un site dédié au même ouvrage a été élaboré. Le Web apportait la diffusion, mais aussi des contraintes techniques qui rendaient le programme précédent inutilisable. Bien que son interface hypertextuelle soit beaucoup plus élaborée et qu'il aurait été très intéressant de la voir reproduite dans cette deuxième génération ce la n'a pas été possible. Ces deux générations successives de dispositifs étaient donc parfaitement complémentaires mais aucune ne donnait entièrement satisfaction, c'est ce qui a amené à une réflexion sur le montage d'une troisième génération qui ferait la synthèse des possibilités des deux dispositifs. C'est ainsi que le site *CoLiSciences* actuel peut être présenté comme une évolution qui bénéficie de cette double parenté : plus d'ouvrages, plus d'interactivité que dans l'ancien site, mais cette fois disposant d'une interface hypertextuelle plus élaborée, bien qu'elle n'atteigne toujours pas les qualités ergonomiques et surtout les possibilités navigationnelles du logiciel de départ. Ce site n'est pas encore totalement abouti que déjà se profile un autre projet, recentré sur Claude Bernard et allégé des parcours notionnels. Nous sommes passés d'un objet individuel à un objet partagé, d'un objet muni d'une interface complexe et puissante à un objet où il est difficile de la reproduire, mais où il devient essentiel de se recentrer sur ce qui est utile et fait sens pour le lecteur. C'est ainsi que nous pouvons constater que :

« Cette genèse s'accomplit par des perfectionnements essentiels, discontinus, qui font que le schème interne de l'objet technique se modifie par bonds et non selon une ligne continue. (Simondon, 1969, page 40) »

Pour représenter ces bonds résultants des modifications de l'environnement dans lequel a été construit le dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, nous pouvons présenter cette genèse de manière plus simple sous la forme d'une généalogie faisant intervenir les « briques » principales de notre architecture technique. Cette généalogie se construit principalement à travers deux lignées :

- Celle des dispositifs hypertextuels que nous venons de discuter. Ce sont essentiellement des évolutions endogènes de notre objet technique.
- Celle des différentes briques logicielles qui ont servi à bâtir le dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* actuel. Ce sont de nombreuses évolutions exogènes qui contribuent à modeler le dispositif de manière plus ou moins directe et visible.

Nous pouvons les identifier sur la figure 6.2 puisque nous pouvons trouver d'une part la famille des sites et outils engendrés par l'équipe « *Hypertextes et textualité électronique* » et d'autre part nous trouvons aussi la grande famille des Logiciels Libres qui supporte notre infrastructure technique.

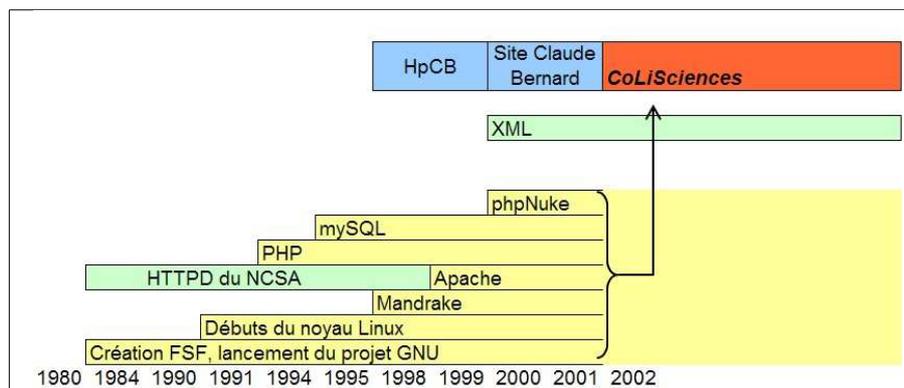


FIG. 6.2: Généalogie du dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*

L'utilisation de Logiciels Libres permet de ne pas avoir à développer tout le dispositif mais uniquement ce qui est spécifique à nos objectifs. En revanche il est difficile de mesurer si nous sommes plutôt dans la position de bénéficiaire d'une fondation solide pour *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, ou si nous sommes plutôt dans une position fragilisée par les mises à jour qu'il pourrait se produire dans les autres projets, pour lesquelles nous n'avons aucun contrôle.

*« Chaque pièce importante est tellement rattachée aux autres par des échanges réciproques d'énergie qu'elle ne peut pas être autre qu'elle n'est.*

*(Simondon, 1969, page 21) »*

Une autre façon de construire cette généalogie est de reprendre les empilements de couches techniques pour essayer de recenser les branches possibles, plutôt que d'utiliser l'échelle temporelle plus traditionnelle. Pour illustrer directement notre propos, il nous faut décrire le contexte de développement du dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*. Comme pour tout système informatique un peu complexe et qui est utilisé 24h sur 24, les modifications ne peuvent pas être faites directement sur le site « de production », le site officiel qui abrite les ouvrages et les met à disposition du public. Les nouveaux développements, les essais et les tests doivent être faits sur des machines « privées ». Il existe donc plusieurs instances du dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, ces instances sont quelquefois parfaitement synchronisées, puis une nouvelle étape de développement est lancée et la machine qui l'héberge est désynchronisée, puis c'est la machine de test qui est modifiée, puis à nouveau synchronisation avec la mise à jour de la machine de production. Nous avons là une espèce de mouvement de flux qui pousse *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* vers l'avant, un peu à l'image de la chenille qui se replie sur elle-même pour avancer.

Cela veut aussi dire que le dispositif n'est pas unique ; non seulement il n'est pas unique, mais l'architecture technique « profonde » qui l'anime est

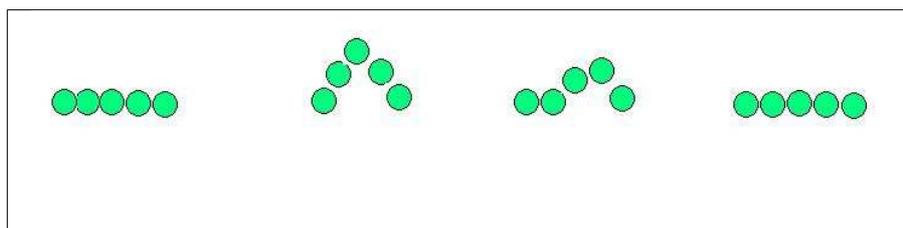


FIG. 6.3: Le dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* avance comme une chenille

presque à chaque fois différente, ceci a été d'abord dû aux aléas de maintenance d'une machine par rapport aux autres, puis cela a permis de vérifier la portabilité et la transparence de notre code par rapport à l'architecture du serveur qui le supporte. Contrainte auto-imposée qui sert de garde fou pour valider les options que nous avons prises pour atteindre la pérennité du dispositif.

C'est ainsi que l'on trouve supportant le dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, les architectures techniques suivantes :

- Windows 2000, easyPHP
- Windows XP, Uniform Server 3.1.1 (Apache 2.0.50, PHP 5.0.0, MySQL 4.0.20d)
- Red Hat
- Mandrake 8.1 (Apache 1.3.22, PHP 4.3.9, MySQL 3.23.41)
- Sun OS (Apache 2.0.53, PHP 4.3.11, MySQL 3.23.41)

Plus qu'une liste, il faut voir là la maléabilité du dispositif, le fait qu'il soit capable de se fondre dans un paysage technique complexe et mouvant. Plus précisément, qu'il soit capable d'ubiquité et de reconfiguration, qui l'amène à la convergence du dispositif comme l'énonçait Gilles Simondon :

*« Le problème technique est donc plutôt celui de la convergence de fonctions dans une unité structurale que celui d'une recherche de compromis entre des exigences en conflit. (Simondon, 1969, page 22) »*

## A. Modèle de publication

Notre modèle de publication est assez spécifique sur deux aspects : d'une part il dépasse les problématiques des Bases de Données traditionnelles pour utiliser un modèle basé presque exclusivement sur l'utilisation de documents XML. Et d'autre part, les documents XML sont utilisés directement en ligne, sans passer par une phase de traduction, par exemple en HTML au moyen de XSLT. En effet, le contenu qui fait l'objet de la publication étant majoritairement stocké sous format XML, est directement affiché, modifié en ligne ce qui permet d'éviter de multiplier les interfaces et les traductions de formats qui sont souvent sources d'erreurs ou de perte d'information. Une Base de Données est tout de même utilisée pour la gestion administrative du site.

*« Ce à quoi on donne un nom unique, comme, par exemple, celui de moteur, peut-être multiple dans l'instant et peut varier dans le temps en changeant d'individualité. (Simondon, 1969, page 19) »*

Ceci est toujours d'actualité dans notre contexte, un site Web, un parcours hypertextuel, une base de données sont des termes génériques et très vagues, qui nécessitent d'être définis et précisés. L'utilisation de XML nous permet d'utiliser un format de stockage ouvert et donc d'être indépendant de la technologie de stockage. En particulier si dans l'avenir une nouvelle technologie de publication faisait jour, ou si nous voulions construire un autre objet qu'un site Web, par exemple un CD ou un DVD multimédia, une transformation basée sur la technologie XSLT devrait nous permettre de passer facilement du format XML à un autre, XML ou quoi que ce soit.

**Modèle de publication traditionnel** Dans le modèle de publication traditionnel, tel qu'il est représenté figure 6.4, les données qui font l'objet de la mise en ligne sont principalement stockées dans une Base de Données, facilement

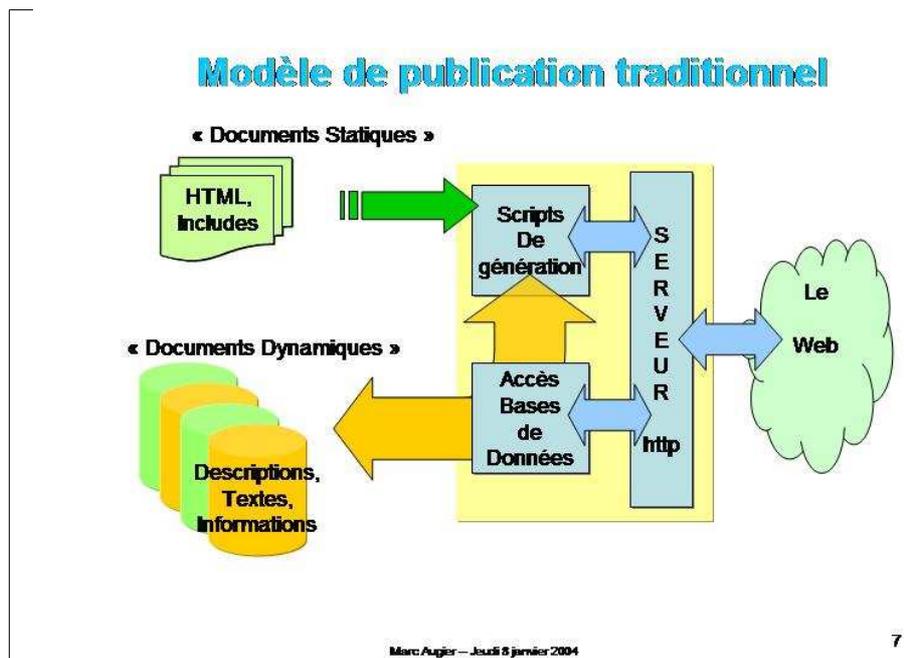


FIG. 6.4: Modèle de publication traditionnel

accessible depuis différentes applications, d'ailleurs dans de nombreux cas elle aura été produite, ou continue même d'être produite par une application extérieure au site Web.

Les avantages de l'utilisation d'une Base de Donnée sont liés à son caractère dynamique, elle peut être mise à jour pendant le fonctionnement du site de manière transparente et automatique et elle permet de faire des requêtes de sélection très précises, par exemple en croisant plusieurs informations.

En revanche, une Base de Données manipule difficilement les « gros » fichiers, textes ou images. Suivant la technologie du Système de Gestion de Bases de Données (SGBD), ce terme pourra s'appliquer à des fichiers de plus de 255 caractères seulement.

Pour améliorer ce type de performance, deux solutions sont envisageables : par les améliorations du SGBD ou par l'utilisation d'une autre technologie de stockage.

## B. Le stockage de données XML

Nous avons déjà présenté le cadre général d'utilisation de XML dans le contexte de la numérisation au cours du chapitre 4. Nous avons vu comment le standard utilisé généralement dans ce contexte est passé de SGML à XML, quels étaient les différents standards disponibles pour numériser et baliser du texte. Nous verrons plus loin comment s'est fait la construction de la DTD que nous avons retenu pour le modèle de publication *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*. Dans un premier temps nous allons commencer par détailler comment le stockage des ouvrages, basés sur des fichiers XML a été architecturé, ceci en cherchant à combiner les avantages des différents outils à notre disposition, avec encore une fois en arrière plan les expériences précédentes et leurs résultats. Le stockage de données numériques peut se comprendre à plusieurs niveaux : au niveau physique nous dirons que les données sont soit des fichiers, soit des enregistrements dans une base de données. Au niveau logique, nous dirons que les données sont soit brutes, soit balisées. Il est ensuite possible de combiner ces niveaux, dans notre cas nous avons des données balisées, reste à décider si elles devront être codées dans un fichier ou dans une base de données. Si elles sont codées dans un fichier, le problème s'arrête là, en effet la structure de balisage est reportée telle quelle dans le fichier qui n'apporte pas une structuration supplémentaire, si ce n'est une granularisation du contenu par fichier. En revanche, si elles sont codées dans une base de données, la structuration de la base de données doit venir en complément ou en support du balisage et le danger est que cette structuration ne soit qu'un filtre supplémentaire qui ajoutera de la complexité et de la lourdeur.

**Utilisation de fichiers** Comment les données XML doivent-elles être conservées ? La réponse la plus simple est d'utiliser un simple fichier texte, puisque

c'est ainsi que la norme XML (XML, 2004) le propose : une chaîne de caractères Unicode ou ASCII, stockée sous la forme d'un fichier texte. Pourtant, quand il est fait référence à la conservation des données, ce sont d'abord aux bases de données relationnelles que l'on pense. Nous avons là plusieurs technologies, qui peuvent être utilisées de manière indépendante ou combinées.

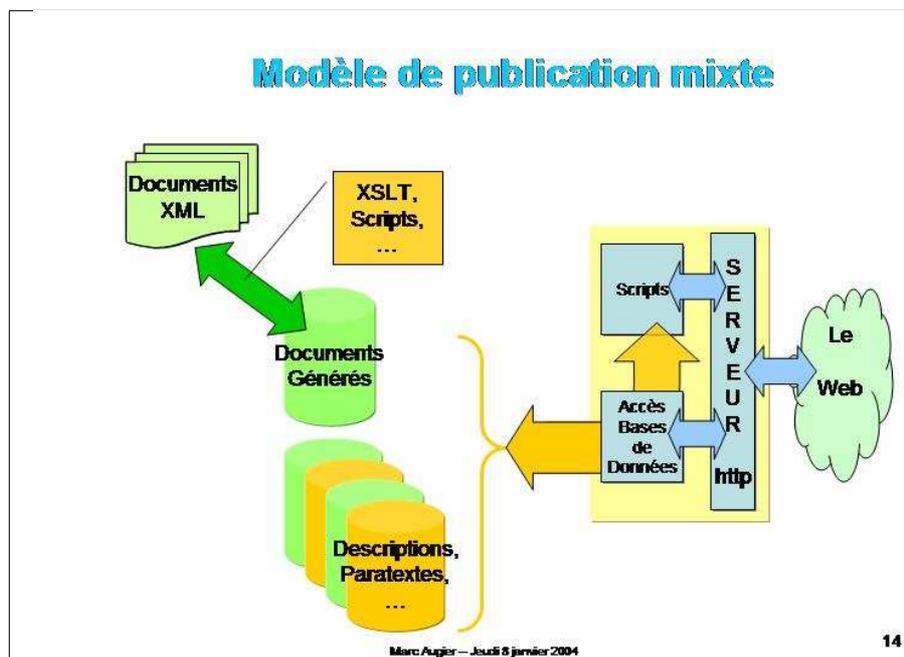


FIG. 6.5: Modèle de publication mixte

**Modèle de publication mixte** Afin de pouvoir bénéficier des avantages de XML, tout en gardant les possibilités d'animation dynamique d'une Base de Données, il paraît simple d'organiser une chaîne de transformation de XML vers un format importable dans la Base de Données au moyen de XSLT. C'est l'architecture que nous voyons mise en œuvre sur la figure 6.5.

La contrainte dans ce cas viendrait de la nécessité de mettre à jour les fichiers XML d'une part, et les informations dynamiques liées à leur contenu d'autre part. Cela nécessiterait de synchroniser les deux « fonds » d'infor-

mation, à travers la chaîne de transformation XSLT, ce qui peut rapidement mener à un casse-tête technique et au final, à des corruptions de données.

**Modèle de publication XML** Ce modèle est souvent retenu pour des sites d'importance moyenne, quand il devient trop difficile de modifier les pages une par une, mais que les besoins de modification dynamique ne nécessitent pas l'emploi d'une Base de Données, qui est coûteuse en ressources, à la fois système et humaines.

Les documents XML contiennent les informations à afficher sur le site, les scripts de génération ont la responsabilité de l'affichage de ces informations. La mise en forme peut être stockée dans des feuilles de style CSS, ou même décrite dans des fichiers XML spécifiques. Dans ce cas nous aurions en quelque sorte deux « Bases de Données » XML, une pour le fond, une pour la forme.

C'est l'architecture que nous voyons représentée sur la figure 6.6.

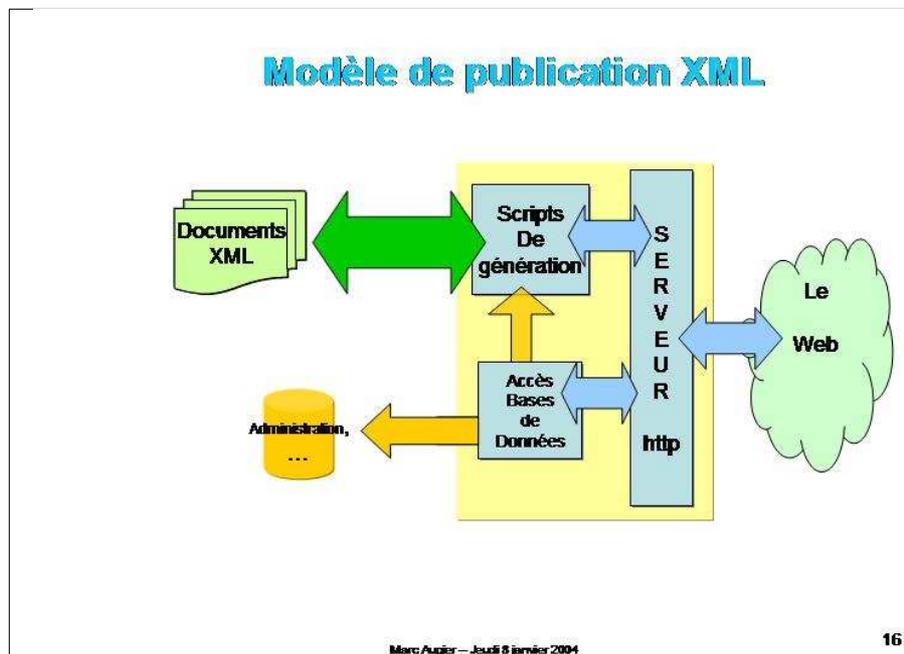


FIG. 6.6: Modèle de publication XML

**XML, solution de publication enrichie** *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* se trouve au carrefour du Web et du domaine de la gestion de documents électroniques, pour lequel SGML fournit depuis longtemps des moyens de classification. C'est donc presque naturellement que nous nous sommes tournés vers XML, qui se trouve aussi au carrefour du Web et des technologies SGML. Mais notre modèle de publication Web va bien au-delà de la simple gestion de document électroniques : *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* est en plus un système hypertextuel avancé pour lequel le rôle des méta-données est majeur. Le texte original est grandement enrichi de notations notionnelles, de manière à permettre une navigation non plus à travers les mots mais à travers les idées d'un corpus.

Notre idée, plutôt que d'utiliser XML uniquement comme intermédiaire, est de l'utiliser directement comme format de stockage. Bien entendu, un système à base de XML est *a priori* moins performant pour exécuter des requêtes de recherche qu'une base de données relationnelles, en revanche, il dispose par construction de fonctionnalités sémantiques qu'il pourrait être beaucoup plus complexe à implémenter dans une base de données.

Mais notre modèle de publication Web va bien au-delà de la simple gestion de document électroniques : *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* est en plus un système hypertextuel avancé pour lequel le rôle des métadonnées est majeur.

Le texte original est enrichi de notations notionnelles, pour cela chaque paragraphe a été annoté et certaines notions « clefs » lui ont été attribuées.

Il est possible de naviguer non plus en suivant les mots par une simple recherche textuelle, mais à travers les idées du corpus en faisant référence aux notions.

Grâce à XML, le balisage notionnel fait partie intégrante du document lui-même, il suit le document et il n'y a donc aucun risque de désynchronisation

ou de perte d'information comme cela pourrait être le cas avec une Base de Données.

**Les avantages de XML** C'est un système ouvert, sur lequel nous pouvons implémenter de nombreuses technologies, propriétaires ou ouvertes. Dans notre cas, c'est la possibilité d'utiliser le langage de scripts php au travers de la DOM, ceci pour interroger directement les documents XML et produire du HTML à la volée qui nous a renforcé dans notre volonté de faire « sans » Base de Données.

A terme, nous pouvons envisager l'ouverture de canaux RDF qui permettraient à d'autres sites de réutiliser le contenu *C<sup>o</sup>L<sup>i</sup>s<sup>c</sup>iences*, depuis notre site et en fonction des demandes de leurs propres internautes, sans besoin d'implémenter des interfaces coûteuses en temps de développement et en ressources, ni d'exporter notre contenu dans sa globalité pour le remettre en forme sur un autre site.

La différence avec l'utilisation d'une Base de Données se fait à deux niveaux :

- Modèle hiérarchique contre modèle relationnel
- Fichier contre Système de Gestion de Base de Données (SGBD)

En effet, un système à base de XML est très similaire à une Base de Données hiérarchique, il a pratiquement les mêmes avantages et les mêmes inconvénients. Ce système est très rapide pour effectuer des requêtes qui « descendent » l'arborescence hiérarchique des données, en revanche il n'est pas du tout adapté pour « croiser » des informations de même niveau. La traditionnelle requête multi-critères qui combine des informations provenant de plusieurs tables ne pourra pas être mise en œuvre dans ce modèle.

**Les avantages des Bases de Données** C'est au niveau des performances brutes que nous pourrions être les plus inquiets par rapport à notre choix technologique, en particulier sur l'impossibilité d'établir des requêtes croisées ou multi-critères.

La mise à jour des enregistrements d'une base de données est elle aussi beaucoup plus simple, ou du moins les outils à mettre en œuvre sont beaucoup plus efficaces et standardisés, mais au cours de l'année écoulée, nous avons pu voir apparaître de nombreux outils et il semble que nous pourrions rapidement compter sur une vraie solution ouverte d'édition en ligne de documents XML sous peu.

L'aboutissement du projet Apache d'implémentation d'une base de données « native » XML serait en fait la communion des deux mondes et la preuve que notre choix reposait sur un vrai besoin que les simples bases de données ne satisfont pas.

**Le couplage XML-Base de Données** Comme nous l'avons dit précédemment, il semble toujours opportun de tenter de mettre en œuvre une solution qui additionnerait les avantages en éliminant les inconvénients de chaque technologie. C'est un peu la voie que nous avons suivie puisque nous avons à côté du contenu du site proprement dit, stocké dans des fichiers XML, toujours la présence d'une base de données pour assurer l'administration du site. Mais comme nous l'avons aussi vu, dans ce type de mise en œuvre il faut bien entendu veiller à ne pas dupliquer les informations à la fois dans un format et dans l'autre.

Dans notre cas, nous n'avons pas prévu de passerelle entre les deux formats, ce qui souligne bien le fait que les données présentes dans un format

sont réellement différentes et indépendantes de l'autre format comme cela est présenté sur la figure 3 précédente.

## C. Délinéarisation et découpage du corpus

### 1. L'unité d'information

La plupart du temps avec un livre papier traditionnel, l'adressage et l'accès à une unité d'information se fait à travers le numéro de page. C'est par exemple le cas avec le sommaire, l'index ou toutes les tables utilisées sur ce support. Pour faire référence à une information, que ce soit un chapitre dans le sommaire ou un mot dans l'index, c'est toujours la page qui est désignée par son numéro. Le lecteur doit feuilleter l'ouvrage jusqu'à la page désignée pour trouver l'information qu'il cherchait. Dans le cas du sommaire, ce sera simple puisque les titres sont mis en évidence, mais dans le cas de l'index, il lui faudra sans doute lire la page en entier pour repérer la citation qu'il cherche. Pour faciliter ce travail de recherche du lecteur, certains ouvrages (en particulier religieux) ont aussi leurs paragraphes ou versets numérotés. Il est ainsi possible de pointer le lecteur vers une unité d'information plus restreinte. Le paragraphe, le chapitre ou la partie sont aussi quelquefois utilisés, suivant le cas et il sera possible de trouver des références telles que : *le chapitre 1 traitant tel sujet* mais c'est alors l'effet inverse qui est produit, l'unité d'information devient un ensemble plus ou moins important de pages.

Pour le corpus *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, la granularité avait besoin d'être aussi fine que possible, de manière à pouvoir isoler les notions traitées par le texte les unes des autres. La phrase aurait pu être choisie, mais elle aurait conduit à hacher totalement la lecture, passer d'une phrase à une autre après une manipulation, aussi simple soit-elle, conduit à freiner et perturber le lecteur. La page n'avait

plus de réel sens, puisque nous sommes libérés du support physique. entre les deux, nous avons trouvé le paragraphe, qui répond au critère de finesse de grain, les notions développées au cours d'un même paragraphe sont d'un nombre réduit, la lecture n'est pas réellement interrompue, sur le papier, la fin d'un paragraphe conduit là aussi à une mini-rupture : aller à la ligne, saut de ligne, départ indenté qu'il est peut-être possible d'assimiler à la rupture engendrée par le *clic* qui permet d'appeler le paragraphe suivant.

## 2. Délinéariser mais pas désorienter

Nous l'avons dit, la lecture d'un ouvrage électronique induit un sentiment de désorientation du fait de l'abstraction du support ; le système hypertexte amplifie encore ce phénomène du fait de la non linéarité de la lecture et qu'il est en fait possible de passer plusieurs fois par le même nœud tout en ayant suivi des parcours différents.. Notre travail de granularisation des ouvrages a été couplé à une réflexion sur l'ergonomie de l'interface du système hypertextuel de manière à redonner au lecteur des points de repères, des points d'appui pour sa pensée. La « page » *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* représente en fait uniquement un paragraphe de l'ouvrage original. Par exemple le nom de l'ouvrage et de son auteur sont toujours présents, sur toutes les pages du corpus. De la même manière qu'avec un livre papier, il est très simple de le retourner pour prendre connaissance de sa couverture ; l'en-tête de chaque page du système hypertextuel reprend le même type d'informations, qu'il est possible de masquer en partie pour ne pas perturber la lecture. Un autre point de repère est de montrer systématiquement sur chaque paragraphe le fac-similé de la page à laquelle il appartient.

### 3. Génération automatique des liens

Traditionnellement, un site Web propose des documents (au format HTML) reliés entre eux par des liens. Ces liens sont écrits dans le document source et pointent sur le document cible, sans qu'il ne soit possible, ni dans l'un comme dans l'autre de vérifier l'existence de l'autre : le document source ne sait pas si le document cible existe, et le document cible ne sait pas qui pointe vers lui.

Dans ces conditions, et sans outils de vérification appropriés, un site Web peut aboutir à un résultat très éloigné de ce que la théorie hypertextuelle nous a appris et surtout complètement incohérent, voire inutilisable, pour rester en dehors de tout débat sur la pertinence d'appliquer la théorie de l'hypertexte sur la toile.

### 4. Problématique du changement de support

A travers les premiers points décrivant les opérations et les réflexions que nous avons menées pour mener à bien le découpage du corpus, nous avons déjà répondu à certains des questionnements liés au changement de support. Passer du papier au numérique, c'est aussi passer de la page au paragraphe, c'est passer du texte seul au texte annoté et enrichi, c'est aborder un nouveau médium avec toutes ses nouvelles possibilités et ses nouvelles contraintes comme nous l'avons vu dans la section **3.4 - Le livre, du papier à l'électronique** à la page 73. Le changement de support se fait ainsi non pas seulement en tant que technique de numérisation, mais aussi en tant que mutation.

## D. L'influence des réseaux de Petri

Nous avons dit en introduction que *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* implémentait un système hypertextuel à trois modules, tels que nous l'avons défini ensuite dans le chapitre consacré à une réflexion sur l'hypertexte. Toujours dans l'introduction, nous avons présenté notre implémentation physique de ce modèle, suivant la mise en œuvre d'un moteur hypertextuel qui remplit les fonctions de présentation et de consultation. Pour que soit bien compris la complexité des situations que ce moteur à en charge d'évaluer, nous pouvons nous référer au diagramme des parcours hypertextuels qui doivent être rendus possibles à travers le corpus, tel qu'il est présenté figure 6.7.

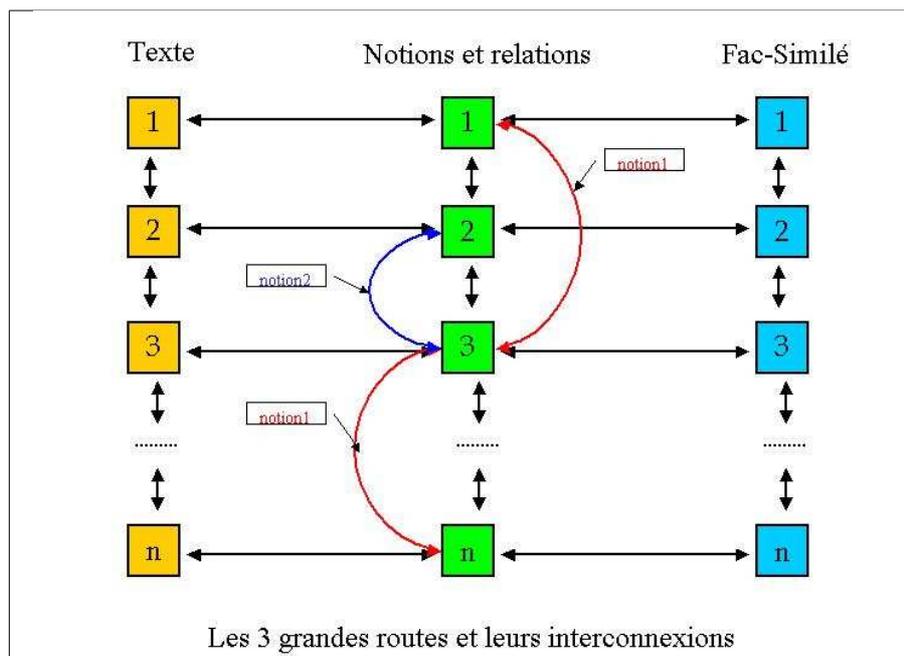


FIG. 6.7: Page d'accueil *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*

Cette figure montre tous les chemins que le moteur doit évaluer, ouvrir ou fermer suivant la disponibilité des nœuds du réseau et la navigation du lecteur. C'est à la fois la quantité des choix possibles, mais surtout la nécessité d'évaluer ces choix à un instant donné en fonction des étapes précédentes

qui nous a conduit vers la mise en place d'un moteur hypertextuel dont le fonctionnement serait semblable à celui d'un automate au sens industriel du terme. A un instant donné et en fonction de sa programmation, l'automate dispose d'un certain nombre de possibilités de fonctionnement, les événements extérieurs ou l'opérateur conduisent vers le choix de l'une de ces possibilités, ce qui fait avancer l'automate d'une étape dans sa programmation et ainsi de suite.

Afin de ne pas trop alourdir l'exposé, nous commencerons par proposer au lecteur un premier exemple d'utilisation fondé sur la théorie des graphes qui permet de visualiser la navigation d'un lecteur à travers le corpus. Cette navigation est modélisée comme les différents états d'un automate fini, ce qui permet de décrire l'évolution de la navigation au cours du temps en fonction de l'état précédent (le paragraphe et ses conditions de visualisation) et d'un signal d'entrée (la requête du lecteur). C'est ce qui est présenté sur la figure 6.8.

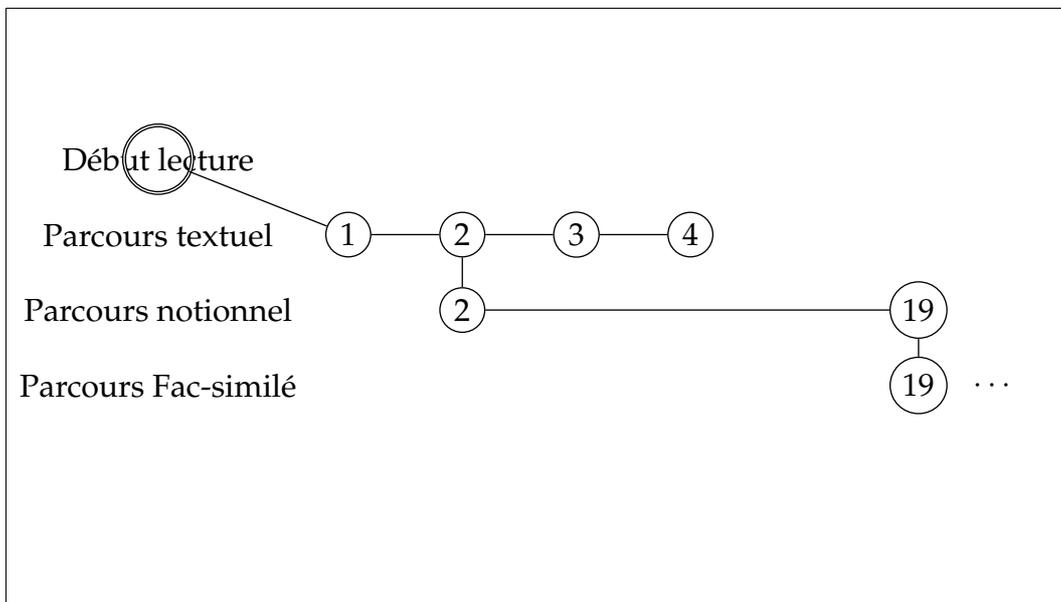


FIG. 6.8: Graphe d'un parcours géré par le moteur hypertextuel

Ces constatations sur le comportement que nous retrouvons dans le moteur qui anime le système *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* nous a conduit à modéliser son mode de fonctionnement sur celui d'un automate fini et plus précisément d'après le modèle théorique des réseaux de Petri.

Les réseaux de Petri ont été introduits par le mathématicien allemand Carl Adam Petri en 1962 (Petri, 1962; David et Alla, 1997) pour modéliser des systèmes à événements discrets. Les réseaux de Petri sont une technique de description formelle des systèmes à événements discrets. Un réseau de Petri modélise un système à événements discrets par un ensemble de variables d'état, les places (représentées par des ellipses) et par un ensemble d'opérateurs de changements d'état, les transitions (représentées par des rectangles). Les places et les transitions sont reliées par des arcs. Les arcs représentent les conditions nécessaires au changement d'état et les changements d'états.

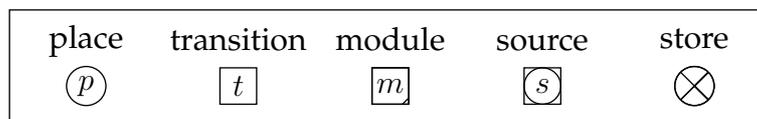


FIG. 6.9: Les nœuds d'un réseau de Petri

Cette théorie fonde de nombreux travaux dans le champ de l'automatisme, elle permet bien entendu de gérer des situations très complexes qui dépassent nos enjeux mais elle peut s'adapter à des situations beaucoup plus simples. Il est par exemple possible de modéliser le changement des saisons au cours d'une année. Dans ce système les quatre états possibles sont donc *printemps*, *été*, *automne* et *hivers*. Chaque état est représenté par une place, qui porte son nom. Le passage d'une saison à une autre se fait grâce à une transition, il existe en fait aussi quatre transitions pour répondre aux quatre situations de passage et il n'est possible de passer d'une saison à une autre que si la transition entre elles est franchissable. C'est ce réseau qui est représenté sur la figure 6.10.

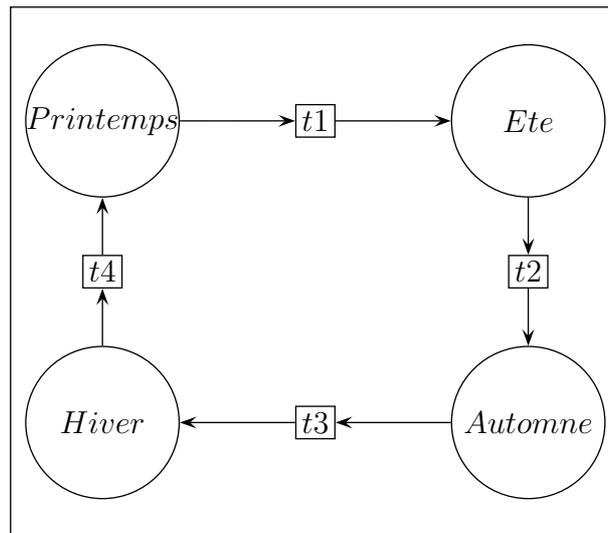


FIG. 6.10: Modélisation des quatre saisons (Petri)

Il pourrait être possible de modéliser ainsi tous les chemins existants à travers le corpus *C<sup>o</sup>L<sup>i</sup>s<sup>c</sup>iences*. Chaque paragraphe de texte devrait alors être représenté sous la forme d'une place, et nous devrions tracer tous les arcs qui existent entre ce paragraphe et ceux auxquels il serait susceptible de conduire, sous des conditions représentées par les transitions. Compte tenu de la taille du corpus, et des interactions qui sont tissées entre les paragraphes de texte un tel graphe ne serait sans doute ni lisible, ni utile. C'est donc un modèle simplifié que nous allons utiliser pour nous permettre de modéliser rigoureusement le moteur hypertextuel *C<sup>o</sup>L<sup>i</sup>s<sup>c</sup>iences*, en prenant comme parti pris de montrer uniquement ce qui se passe pour passer d'un paragraphe à un autre, quel que soit ce paragraphe.

La figure 6.11 schématise le comportement du moteur *C<sup>o</sup>L<sup>i</sup>s<sup>c</sup>iences* pour une situation donnée. Il est capable de se situer et de situer les destinations possibles à partir du paragraphe en cours d'affichage. Quand le lecteur déclenche une requête, celle-ci est interprétée par le moteur, qui sera capable dans tous les cas, de fournir une réponse aussi adaptée que possible à la

requête. Pour garder la terminologie des réseaux de Petri, disons qu'à tout instant le moteur s'attend à recevoir une **transition** lui indiquant le choix du parcours, celles-ci sont au nombre de trois :

- Parcours fac-similé.
- Parcours plein texte.
- Parcours notionnel.

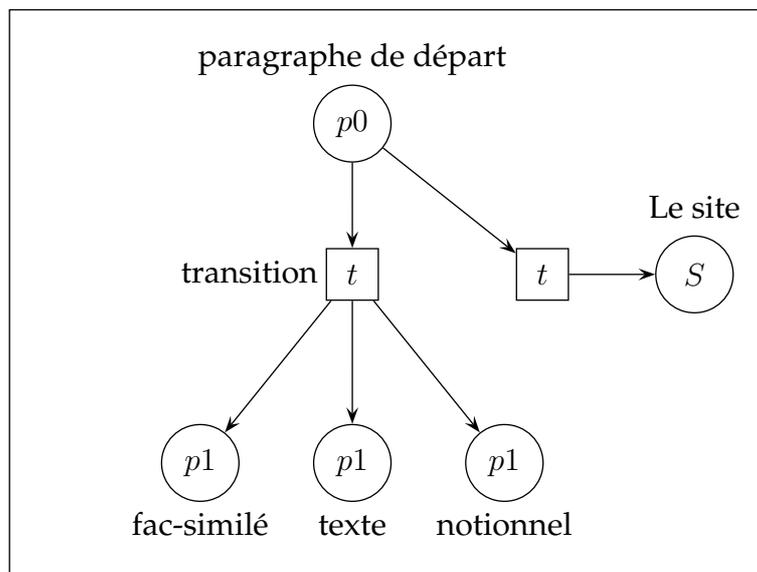


FIG. 6.11: Modélisation du moteur *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*(Petri)

Pour compléter cette vision tout de même simpliste du moteur hypertextuel, il faudrait au moins rajouter une transition particulière. Celle qui est utilisée pour stopper le moteur, sortir de la navigation hypertextuelle et retourner à une navigation Web classique.

Bien entendu, sur les transitions peuvent s'ajouter des attributs complémentaires permettant au moteur de se situer : numéros de pages et de paragraphes, mots-clefs de recherche ou notions, ...

Bien entendu, certaines requêtes peuvent ne pas avoir une réponse positive, en quelque sorte ce sont des requêtes non prévues par la programmation de l'automate qui doit être capable de les gérer. Cette non-réponse est aussi

une information, le moteur devra être capable de l'interpréter pour afficher à destination du lecteur une page qui apporte cette information issue de la pré-analyse du moteur de telle sorte que le lecteur devrait être en mesure de comprendre ce qui ne va pas dans sa navigation, ou dans le site, puisque comme tout objet créé par la main de l'homme il est sujet aux erreurs. L'enjeu du traitement de ce type de requête sans réponse est d'éviter les problèmes de désorientation dus aux erreurs de manipulation ou aux problèmes techniques.

## **E. Structuration**

La question de la structuration peut s'aborder de plusieurs manières. Par exemple la construction du système hypertextuel peut induire une structuration, (d'ailleurs une piste à explorer plus avant serait de qualifier la possibilité pour un tel système de ne pas induire de structuration), mais alors qu'en est-il de la liberté du lecteur ? Le lecteur n'est libre que d'aller où l'auteur l'a autorisé à aller et suivant le parcours qu'il a pré-établi. Cette liberté est plus ou moins contrainte : dans certains sites, les liens ont été spécifiquement codés, les uns après les autres ; sur d'autres sites, les liens peuvent être le résultat d'un calcul, ce calcul peut laisser plus ou moins de degrés de liberté au lecteur. L'étude de la structuration nous amène aussi à la voir comme imposée par la structure physique de l'ouvrage. Et dans ce cas, il nous paraît difficile d'étudier les problématiques de l'hypertextualité dans un corpus de littérature (donc sans doute intimement lié au support papier originel) sans nous poser de question sur le livre électronique. Le dispositif électronique reconstruit une nouvelle matérialité que le lecteur doit s'approprier. A la différence du livre papier, qui n'existe plus aujourd'hui pratiquement que sous la même instance, ou tout du moins des instances très semblables, le livre électronique se décline sous des formats, des dispositifs et donc des instances très différents : ordinateur ou

matériel dédié, portable ou fixe, d'après un contenu sur CD, sur disque dur ou sur « la toile », en utilisant un logiciel générique ou dédié, hypertexte ou linéaire.

## **F. Appropriation et apprentissage**

Une des questions qu'il paraît judicieux de se poser avant de construire un système hypertextuel consiste à essayer de comprendre comment se fait l'apprentissage de la navigation dans un tel système. C'est même en fait un enjeu majeur puisque la désorientation est régulièrement citée comme un des obstacles majeurs à la bonne compréhension des messages ou des textes publiés en ligne sur le Web.

Encore plus qu'avec un ouvrage papier, la collaboration active du lecteur est incontournable dans le cas du dispositif hypertextuel, sans cela il est voué à l'inutilité. C'est cette idée qui anime Umberto Eco et qu'il décrit dans son ouvrage *Lector in fabula* (Eco, 1985), d'ailleurs sous-titré dans l'édition française *Le rôle du lecteur* pour éviter toute ambiguïté sur le propos et sans doute pour souligner que le sujet est bien l'étude des processus cognitifs déclenchés chez un individu par la lecture.

Le dispositif hypertexte dans sa dimension technologique va être perçu par le lecteur comme un nouvel outil, pour lequel il va devoir passer par un apprentissage qui pourra le conduire à son adoption, après une première phase d'accommodation ou d'adaptation. Il entraîne aussi l'utilisation d'un nouveau médium, qui engendre de nouvelles relations entre l'homme et la machine, que nous souhaitons plus riches mais pour cela elles devront être mises en scène de manière appropriée.

Si on s'accorde sur l'idée que l'hypertexte ne peut se limiter à faciliter l'accès à des ressources, il faut essayer d'identifier ce qui permet d'aller

plus loin. Il faut penser en termes d'aide à la compréhension et à l'analyse autant qu'en termes d'aide à la navigation et multiplier les objectivations, les représentations, Pour cela, deux éléments sont à prendre en compte :

1. Faciliter toutes les possibilités de visualisation globale du corpus,
2. Aider à garder, représenter et organiser les parcours en essayant de traduire le dynamique (temporel) par du spatial.

*« L'appropriation résulte de la conjonction de deux processus adaptatifs que sont l'accommodation et l'assimilation. (Piaget, 1967) »*

Une problématique difficile à résoudre est celle de la mesure de la capacité des lecteurs à faire l'appropriation du dispositif, par exemple parce qu'il n'est pas facile d'identifier ce qui doit être considéré comme faisant partie de l'utilisation et ce qui peut être qualifié d'adoption.

*« Dans une approche complexe, nous entendons l'appropriation comme relevant à la fois de rétroactions négatives (accommodation), par l'intériorisation de nouvelles normes, mais également comme rétroactions positives (assimilation) par la création de sens apporté par les utilisateurs. (Puccia et coll., 1999, page 55-56) »*

### **1. *c*<sup>L</sup>*i*<sup>S</sup>ciences : un outil pour l'étude des processus d'appropriation**

Considérant la complexité intrinsèque de ce qui est proposé, à la fois en termes de contenu et d'exploitation de celui-ci, ainsi que l'absence d'une tradition d'exercitation dans ce domaine, il paraît impossible de trouver des formes d'utilisation éducative suffisamment proches de ce qui peut être fait dans les formations institutionnelles. Cela conduit soit à imaginer des activités simples centrées sur des points précis soit des scénarios plus sophistiqués. Une première idée est celle de la création de parcours. On peut d'abord regarder

der du côté des experts, essentiellement des enseignants de SVT (sciences de la vie et de la Terre), des historiens des sciences et des philosophes. Ces experts sont conduits à bâtir des cours, notamment autour de notions ou de problèmes abordés ou traités dans le corpus indexé dans le cadre du projet *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*. Comparer les chemins construits, correspondant en gros à des visites guidées, et décrire comment sont utilisés les outils proposés (par exemple les relations entre les notions aidant à construire une généalogie de ces notions en termes d'histoire des idées), est un objectif essentiel. Les observations des experts peuvent fournir des indications sur les usages possibles par des plus novices et les chemins collectés enrichissent le corpus par l'adjonction de parcours sémantiques finalisés. Si nous voulons aller plus loin dans la structuration, nous pouvons nous inspirer des travaux autour de l'argumentation, domaine qui a pris ces dernières années une place importante dans les programmes de français (Perelman, 1997), (J.B.Grize, 1990), (Vignaux, 1988). Fournir des structures argumentatives vides à remplir au cours d'une exploration, puis à réorganiser est une piste à creuser. Il faudrait également penser aux systèmes de visualisation et de structuration tel Aquanet (Marshall et Rogers, 1992); (Marshall C., 1994).

Revisiter le concept d'hypertexte apparaît donc fondamental, y compris dans le contexte éducatif actuel (secondaire et supérieur) actuel où la recherche d'informations, leur organisation et leur réorganisation, dans le cadre d'activités interdisciplinaires, occupent une place croissante.

Un certain nombre d'expérimentations ont été envisagées par l'équipe « *Hypertextes et textualité électronique* », bénéficiant des situations prototypiques d'usage des hypertextes qu'autorise *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*. Le premier volet comporte :

1. Butinage et transmission de connaissances

Quel est l'impact du support et du mode de présentation d'un cours traditionnel sur les gains d'apprentissage ? Pour répondre à cette question, Dessus et Hédon (1996) ont présenté un cours destiné à un public adulte dans quatre formats différents : sur support informatique, sur support papier, avec ou sans structure hypertextuelle. Les gains de mémorisation des concepts essentiels du cours se mesuraient à partir de la construction de cartes de concepts par les sujets après l'exploration. Les personnes travaillant sur support papier obtiennent les meilleurs gains. D'autres expériences ont montré que la structure hypertextuelle est vraiment intéressante si la lecture est finalisée. Si, au contraire, la lecture n'a aucun objectif précis (lecture d'agrément) (Rouet et Tricot, 1995), la rétention est meilleure avec une simple présentation linéaire du support électronique. C'est ce que nous entendons tester selon notamment trois types de dispositifs : (i) Texte brut sur support écrit papier ou (ii) sur support électronique, (iii) texte organisé sous forme hypertextuelle.

2. Jardinage et création d'hypertextes

On peut passer de la valorisation d'informations au jardinage. L'objectif est ici la création d'un hypertexte, mais l'accent n'est pas tant sur le produit final que sur le processus d'écriture. Il ne s'agit pas d'une écriture au sens classique (cf. les distinctions introduites par Mangenot (Mangenot, 1996) : les matériaux multimédias créés ou rassemblés par les auteurs doivent être associés et structurés. Le sujet rend explicites des connaissances implicitement portées, mais non apparentes, dans un ensemble de documents (Nanard, 1995).

De plus les systèmes d'ancrage des hypertextes permettent de construire un discours (incluant du texte, au sens linguistique) dont la lecture n'est pas

nécessairement séquentielle. Il devient alors plus facile de viser l'élaboration d'un produit moins ardu qu'une rédaction au sens traditionnel, de construire, par exemple, des histoires à tiroirs plutôt qu'un texte narratif. C'est la raison pour laquelle les situations de jardinage commencent à être exploitées avec des enfants (Beaufils, 9-11 Mai 1996) ou avec des adultes (Amélineau et Giovanni, 1996). C'est ce type d'expérimentation que nous entendons aussi développer selon différentes procédures : parcours libre, parcours assisté, parcours finalisé, parcours cohésif, parcours fragmentaire.

## **G. Les médiations autour de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>***

### **1. Pourquoi publier sur la toile**

La publication d'un site Web, en dehors des pratiques commerciales, est très certainement dans la continuité de l'œuvre des encyclopédistes<sup>115</sup>. Il s'agit d'une part de pérenniser un patrimoine et de partager des connaissances. D'autre part, l'apport de la toile, outre la puissance de la diffusion, est de permettre au texte de vivre, de s'enrichir par les relations que qu'il est possible de continuer à entretenir avec les lecteurs. La pérennisation d'un patrimoine est une démarche qui s'inscrit dans la durée, aspect qu'il est souvent difficile de préserver dans un environnement où les technologies se succèdent et deviennent obsolètes rapidement. Pour les mêmes raisons le partage des connaissances a besoin de reposer sur des technologies et surtout des standards à la fois éprouvés, matures et largement disponibles, sans devoir satisfaire une longue liste de pré-requis matériels ou logiciels. L'ouverture de ces standards doit permettre aux contenus publiés, et surtout à la forme de leur publication de rester disponibles tels quels ou de suivre les évolutions futures sans discontinuité.

---

115. Yannick Maignien. *La Bibliothèque virtuelle ou de l'Ars memoria à Xanadu*, 1995

*« L'acte de communiquer ne se traduit pas par un transfert d'information depuis l'expéditeur vers le destinataire, mais plutôt par le modelage mutuel d'un monde commun au moyen d'une action conjuguée : c'est notre réalisation sociale, par l'acte de langage, qui prête vie à notre monde. (Varela, 1996, page 114,115) »*

Les usages permis par *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* trouvent leur fondation dans ceux du livre électronique, enrichis par les possibilités apportées par la connectivité offerte par la mise en ligne sur le Web. Ce n'est pas la disponibilité en ligne 24h sur 24 et 7 jours sur 7 qui apporte une grande nouveauté, puisque nos bons vieux livres nous offrent déjà une telle disponibilité, d'ailleurs assortie d'une fiabilité que les ordinateurs sont encore bien loin d'atteindre. La nouveauté se traduit en fait de plusieurs manières et elle se situe dans la connectivité. C'est sans doute le travail collaboratif qui est le plus porteur d'un nouveau sens et de nouvelles ambitions pour les livres électroniques. Car en effet, ce n'est pas que les ouvrages soient tout le temps disponible mais surtout que le même ouvrage soit disponible à tous en même temps, qu'il puisse être commenté et que ces commentaires soient immédiatement visibles de tous. Nous retrouvons là une utilisation qui existait avec les manuscrits et qui a disparu avec l'imprimé. Les marges des manuscrits prévoyaient que le lecteur puisse annoter l'ouvrage, ainsi les lecteurs suivants pouvaient prendre connaissance de ces remarques et continuer à enrichir l'ouvrage avec les leurs. L'ouvrage manuscrit continuait à évoluer avec ses lecteurs.

Le livre électronique autorise ce genre d'annotation, qui plus est, il les facilite puisque le même ouvrage peut être lu et annoté par plusieurs personnes en même temps. Toutefois, ce type de dispositif doit être rigoureusement pensé, au risque de se retrouver noyé dans un flot de remarques inintéressantes, hors de propos ou pire. En effet, l'ouvrage étant accessible à tous en lecture,

il convient sans doute de mettre en place une procédure d'autorisation ou d'identification des internautes, voire de validation des annotations. Mais alors, *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* ne serait plus uniquement un lieu de publication mais se poserait un problème de censure. Qui autoriser ? Sur quels critères ? Comment susciter des commentaires ouverts et objectifs quand celui qui les écrits sait qu'ils seront relus, filtrés, validés ?

Pour revenir aux usages et essayer d'en résumer la portée, nous pouvons les classer en deux catégories :

**Les usages individuels** D'une part nous pouvons considérer l'usage le plus simple qui puisse être fait de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* et qui consiste pour un internaute à prendre connaissance du corpus et à en lire les ouvrages. En fait, ce travail individuel va plus loin, par exemple pour les chercheurs qui participent au montage de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, chacun intervenant selon ses compétences : ergonomie, architecture informatique, analyse des textes et rédaction des paratextes.

**Le collaboratif** D'autre part, pour les lecteurs, étudiants ou chercheurs qui étudient les textes. Leur passage au travers des ouvrages enrichit la base de connaissances de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* dans le but de faire émerger des parcours particulier et au-delà une interprétation particulière de ces ouvrages. Il existe aussi pour eux une manière de participer plus activement à la construction du site en proposant des listes de signets, c'est-à-dire de pages ou paragraphes regroupés autour d'un thème qu'ils ont choisi. Pour l'instant, aucun dispositif d'annotation en ligne n'a été utilisé.

Nous sommes donc dans une situation paradoxale où le travail collaboratif est fait en majorité par les individus isolés qui parcourent *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, alors que dans l'équipe « *Hypertextes et textualité électronique* » qui met en place le

site, chacun adopte plutôt une approche individuelle pour faire avancer les travaux dans son propre domaine de compétence.

## 2. Les dispositifs disponibles dans *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*

L'analyse des usages qui sont faits de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* peut être faite par deux types d'outils : d'une part ceux que nous avons trouvés déjà prêts à l'emploi dans phpNuke et que nous avons pu réutiliser tels quels, d'autre part

### Outils standard issus de phpNuke

**Referers** Un « *referer* » est un site qui fait référence à un autre. Il existe dans les serveurs Web Apache la possibilité d'obtenir l'adresse d'où est parti un internaute avant d'arriver sur la page qu'il est en train de lire. Cette possibilité est utilisée par phpNuke pour dresser la liste des adresses des sites d'où sont partis les internautes avant d'arriver à *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*. Pour un *webmaster*, cela permet de savoir comment les visiteurs ont eu connaissance de notre site : moteur de recherche, site partenaire, etc... L'annexe D donne un exemple de cette liste et les enseignements qui peuvent en être tirés.

**Statistiques de connexions** Les statistiques de connexion sont là aussi construites à l'aide d'informations fournies par le serveur Apache, assorties d'informations générales comme l'heure. Il est ainsi possible de connaître pour chaque visite sur *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* : l'heure de la visite, le matériel informatique utilisé par l'internaute, en particulier quel système d'exploitation est installé sur sa station de travail et quel *butineur* il utilise pour afficher les pages Web. Encore une fois ce sont surtout des informations techniques utiles avant tout pour le *webmaster*. L'annexe F donne un exemple de ces statistiques.

### Outils spécifiques

**Analyse des parcours** Le but de la lectochromie est l'analyse des parcours à travers le corpus, en utilisant une échelle de couleurs pour quantifier la fréquentation d'un paragraphe, ou d'un ouvrage. Ce dispositif est étudié en détails dans la section 7.4 qui se trouve à la page 151.

**Analyse lexicale** Il s'agit d'un dispositif relativement courant, que se charge de compter les occurrences des mots qui se trouvent dans le corpus. Cette analyse est filtrée automatiquement par une première étape de lemmatisation, qui peut être suivie d'un filtre manuel sur certains mots spécifiques.

**Signets** Les signets peuvent être créés par n'importe quel lecteur enregistré sur *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*. Il s'agit de regrouper sous une rubrique des paragraphes du corpus. Le nom de la rubrique ainsi que le choix des paragraphes sont laissés entièrement au libre arbitre du lecteur. Les signets peuvent être de deux types : privés et alors il n'y a que leur créateur qui puisse les voir, publics et tous les internautes peuvent les consulter. L'équipe « *Hypertextes et textualité électronique* » a établi quelques listes de signets publics pour montrer certaines pages remarquables, et faciliter l'accès du corpus aux lecteurs « simples curieux ».

**Parcours** La lectochromie donne une vision globalisée des parcours fait dans le corpus. Les mêmes informations peuvent être utilisées non plus accumulées mais discrétisées pour reconstruire les parcours des lecteur du corpus *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*. A terme, ce dispositif sera ouvert aux modifications et un lecteur pourra revoir ses parcours et les modifier, pour arriver à un véritable outil dynamique de construction de parcours.

### 3. Les interfaces

Nous arrivons maintenant à réfléchir sur le site *CoLiSciences* en tant qu'interface entre un corpus et des lecteurs. S'agissant de proposer un corpus assez vaste et riche, à travers le seul écran d'un ordinateur, il faut bien entendu prendre en compte les problèmes d'ergonomie et d'interface homme-machine qui ne manquent pas de se poser.

**Fonctionnalités de l'interface utilisateur** La diffusion d'une bibliothèque numérique sur Internet se trouve toujours confrontée aux mêmes problèmes. L'équipe en charge de la numérisation de la BNF a une position assez similaire à la nôtre :

*« L'ergonomie de la bibliothèque numérique sur Internet ne pouvait être conçue sur le même modèle que celle des postes des salles de lecture. A la différence des lecteurs situés dans la bibliothèque, le lecteur distant ne peut être aidé et conseillé. Il est seul et susceptible d'utiliser des matériels assez variés, des navigateurs différents etc. Aussi, nous sommes-nous efforcés de concevoir des interfaces homme/machine simplifiées tout en offrant une grande richesse des modes d'accès possibles. (Renoult, 1999, page 83) »*

Comme *Gallica*, notre site *CoLiSciences* permet d'accéder aux différents ouvrages en empruntant des chemins divers : recherche par auteur, par discipline, par domaine. De nombreux textes de présentation, des glossaires, des bibliographies enrichies de notes tissent leur toile hypertextuelle autour des ouvrages pour guider l'internaute.

A cette problématique « utilisateur grand public » de la BnF, nous avons ajouté nos propres contraintes dues aux « utilisateurs chercheurs » qui ne vont plus simplement butiner mais aussi jardiner, c'est-à-dire apporter leurs

propres analyses, leurs propres modifications, toujours au travers de l'interface sur la toile. L'interface utilisateur a été développée tout en essayant de répondre aux contraintes suivantes : Respect de la norme HTML pour rendre le site accessible à toutes versions et tous types de clients WWW. Utilisation de formats ouverts, éviter autant que possible les technologies « propriétaires ». Assurer un confort minimal aux clients WWW non graphiques, et plus particulièrement se conformer aux recommandations du WAIS pour rendre le site accessible aux non-voyants.

La page d'accueil permet l'accès à différents paratextes explicatifs (acteurs du projet, guide utilisateur, ...) et surtout à la « bibliothèque », qui présente la liste des ouvrages disponibles en ligne. Cette liste peut être triée par auteurs, par domaines ou par disciplines. Un accès multi-ouvrages est aussi proposé, celui-ci permet d'entrer dans la globalité du corpus en demandant l'étude d'une notion.

**Fonctionnalités de l'interface administrateur** Le premier mode d'emploi de la partie administrative consacrée à la modification de la page d'accueil est présentée annexe G. Il permet de se faire une idée des possibilités mises à la disposition de l'administrateur *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* pour modifier le contenu du site sans pour autant avoir à faire appel à des connaissances techniques. Un deuxième aspect intéressant de ce manuel est qu'il a été réalisé pour la première version du site *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, qui a ensuite évolué avec une nouvelle présentation graphique et aussi une nouvelle architecture hypertextuelle. Le manuel est donc en quelque sorte le témoin de l'évolution de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, évolution qui s'est faite en préservant l'existant de telle manière que le manuel peut toujours être utilisé, mais aussi évolution qui a ouvert de nouvelles perspectives de navigation en continuité avec les objectifs du projet.

Plutôt que d'être présenté tel quel, le manuel a été revu pour montrer ces évolutions, en mettant en avant les points de convergence comme ceux de rupture.

## 6.4 Notre architecture

Quand nous avons commencé à travailler sur *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, nous n'avons pas pu trouver de solution « toute faite » déjà prête ou abordable, adaptée à notre contexte de publication sur le Web d'un grand corpus de textes numérisés, présentés à la fois en plein texte et en images. En fait, la situation était même plus difficile encore, puisque nos textes étant numérisés sous format XML, nous avons besoin d'outils performants pour créer ces fichiers, les modifier, les accéder en sélectionnant certaines parties de leur contenu et finalement, pour tout simplement les stocker.

Ce dernier point nous a paru au premier abord critique, comme nous venons de l'aborder dans la section **6.3 - Le dispositif technique**, page 110, notre modèle de publication ne repose pas sur une Base de Données et les documents numérisés sont utilisés sous leur forme XML pratiquement sans préparation préalable à la mise en ligne. Ceci est le résultat à la fois du contexte technique dans lequel notre projet a démarré et d'une réflexion plus approfondie sur nos réels besoins. Nous aurions pu utiliser un *Système de Gestion de Base de Données XML*, le projet *Cocoon* géré par le groupe Apache existait déjà, mais il était paradoxalement à la fois complexe et pas assez mature. Sa complexité et la lourdeur qu'il risquait d'apporter à la fois au développement et ensuite au fonctionnement quotidien du site nous ont fait laisser cette solution de côté. L'autre solution aurait pu être de charger ces documents XML dans une Base de Données Relationnelle, mais un premier projet avait déjà permis de constater la lourdeur de cette solution. C'est finalement la solution la plus triviale et la plus criticable qui a été retenue, nous avons décidé de conserver les fichiers issus de la numérisation sous leur format XML et de les utiliser en faisant les transformations *à la volée*. Pour faire

court, nous avons commencé à « jouer » avec l'implémentation DOM de PHP et c'est toujours cette solution qui est mise en œuvre.

Pour la partie serveur, les logiciels que nous utilisons sont donc les traditionnels LAMP : Linux, Apache, MySQL et PHP. Le CMS (*Content Management System*) phpNuke a été repris et modifié pour accélérer le développement du site, il nous a permis d'avoir rapidement une infrastructure vide mais capable de gérer les appels à différents modules, ainsi que des utilisateurs avec des niveaux de privilèges différents. C'est ainsi par exemple que les outils d'administration restent invisibles aux lecteurs. En revanche, phpNuke ne participe pas réellement à la navigation hypertextuelle, il ne fait que l'héberger au même titre que les autres services qu'il nous apporte tels que le glossaire, les pages de description, etc.

Pour la préparation à la mise en ligne des documents qui ne sont pas issus du programme de numérisation de la bibliothèque André Deguine et mais disponibles en format numérique, OpenOffice est utilisé pour traduire des documents texte suivant notre DTD XML. Et finalement, pour l'édition des documents XML, quels que soit leur origine, plusieurs outils ont été utilisés successivement. Le logiciel XMeta<sup>116</sup> a d'abord été utilisé, outre l'aide qu'il a apporté dans la préparation de la DTD il offre en effet de très grandes possibilités pour l'édition XML. Son inconvénient est qu'il demande à ce que le document soit téléchargé localement sur la machine de la personne qui veut apporter une modification, et que le document doit être à nouveau transféré sur le site pour être mis en ligne après mise à jour. Cela demande que les documents puissent être « verrouillés » quand ils sont mis à jour, et un risque subsiste de voir petit à petit se multiplier un certain nombre de versions différentes d'un même document, sur les machines des correcteurs.

---

116. La société qui le commercialise est présente à l'adresse : <http://www.xmetal.com/>

Pour éviter cette situation, nous nous sommes attelés à la construction d'un outil rudimentaire mais efficace pour permettre de modifier les documents XML directement en ligne.

*« Nos techniques aujourd'hui, sont bien davantage que des outils. Ce sont des pensées, mais objectivées, mais instrumentalisées. Ce sont des sciences, mais appliquées. (Comte-Sponville, 2001, définition de la technique, page 576) »*

Notre modèle de publication est très spécifique sur deux aspects : d'une part il dépasse les problématiques des Bases de Données traditionnelles pour utiliser un modèle basé presque exclusivement sur l'utilisation de documents XML. Et d'autre part, les documents XML sont utilisés directement en ligne, sans passer par une phase de traduction, par exemple en HTML au moyen de XSLT. En effet, le contenu qui fait l'objet de la publication étant majoritairement stocké sous format XML, est directement affiché, modifié en ligne ce qui permet d'éviter de multiplier les interfaces et les traductions de formats qui sont souvent sources d'erreurs ou de perte d'information. Une Base de Données est tout de même utilisée pour la gestion administrative du site.

XML nous permet d'utiliser un format de stockage ouvert et donc d'être indépendant de la technologie de stockage. En particulier XSLT nous permettrait de passer facilement d'un format à un autre, XML ou quoi que ce soit.

## **A. Numérisation**

La mise en ligne d'un corpus qui n'existe que sous forme papier passera obligatoirement par une phase de numérisation, qui consiste à photographier les pages et ensuite à décoder ces images pour en extraire le texte qu'elles contiennent. La trivialité de cette description ne doit pas cacher la haute

technicité de cette opération. En effet, nous pouvons distinguer plusieurs phases dans la numérisation :

La première phase a lieu avant la numérisation proprement dite. Il s'agit de mettre au point le format de sortie de cette numérisation, c'est généralement une DTD (*Document Type Definition*) SGML qui sert à cette description.

La deuxième phase consiste en la manipulation des ouvrages pour les photographier, page par page, très délicate compte tenu de l'âge des ouvrages. Cette phase produit des fichiers images qui sont inexploitable au niveau textuel.

La troisième phase est l'analyse du contenu de ces photographies, dans le but de le transférer sous un format numérique au moyen de différents outils de reconnaissance visuelle de caractères (souvent identifiés sous l'acronyme anglais OCR *Optical Character Recognition*). Ceci demande de reconnaître non seulement les caractères constituant le texte, mais aussi la reconnaissance de la structure physique (mise en page et caractéristiques visuelles, segmentation du texte, identification de la typographie : italiques, gras, ... reconnaissance de la structure logique (DTD). Cette phase est aussi appelée *rétroconversion*.

La manipulation d'ouvrages anciens, donc fragiles, demande une expertise et des moyens importants. De plus, les outils d'OCR pour être fiables doivent être longuement mis au point, calibrés en fonction des ouvrages. Ils représentent un investissement difficile à amortir pour une « opération » ponctuelle telle que la mise en ligne d'un corpus d'ouvrages sur le Web. Pour toutes ces raisons, il est aisé de comprendre qu'il est préférable d'externaliser cette tâche.

Le produit final pourrait être un simple fichier numérique en mode texte, mais grâce à l'utilisation de XML dès cette étape, nous pouvons conserver

avec le texte des informations sur sa structure, sur sa provenance, ainsi que tous les autres attributs qui pourraient s'avérer utiles par la suite.

Le sujet dépasse largement le cadre de cette étude et a déjà été traité dans de nombreux ouvrages<sup>117</sup>

Comme pour la bibliothèque numérique de la BnF, nous avons choisi d'externaliser la numérisation (image + ROC) :

*« Dès le démarrage du projet, a été adopté le principe de l'externalisation de toutes les tâches assimilables à une production industrielle. Il fallait non seulement tenir compte des plannings d'exécution extrêmement courts au regard des volumes concernés, mais aussi éviter d'investir dans une logistique lourde et des techniques en évolution trop rapide pour ne pas être sujettes à obsolescence rapide. (Renoult, 1999, page 79) »*

## **B. Architecture serveur**

Comme tout projet mettant en jeu un système d'information, un projet de publication sur le Web doit faire face à des contraintes techniques et logicielles, assorties bien évidemment de contraintes budgétaires. Réaliser une bibliothèque en ligne telle que *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* nécessite non seulement de mettre à disposition des ouvrages mais aussi de mettre au point un *Scriptorium* moderne, en suivant la même démarche que la BNF et son PLAO (Poste de Lecture Assistée par Ordinateur). Notre contrainte par rapport à celui de la BNF, est que notre PLAO doit pouvoir traverser la toile pour aller « s'installer » sur n'importe quel poste client et s'afficher sur son écran avec un simple *browser web*. Parmi la palette de solutions à notre disposition, nous pouvons énumérer les choix habituels : mettre en œuvre un système « standard » du commerce, faire

---

117. Nous invitons le lecteur à consulter par exemple l'ouvrage de Alain Jacquesson et Alexis Rivier, *Bibliothèques et documents numériques : concepts, composantes, techniques et enjeux*, Paris, Editions du cercle de la librairie, 1999

appel à une société tierce pour personnaliser un système existant ou faire développer le notre, et enfin la solution que nous avons retenue qui est de développer nos propres outils et notre propre système. Ce dernier choix ne peut pas être fait sans une évaluation précises des ressources disponibles pour mener à bien le projet, les tâches sont nombreuses et complexes, il s'agit de ne pas gaspiller de l'énergie à résoudre des problèmes déjà résolus par d'autres. Parmi les possibilités offertes, la possibilité d'intégrer des outils existants dans notre solution paraît un moyen d'accélérer le développement, pour peu que des problèmes de droits ne viennent pas compliquer légalement une solution déjà complexe au niveau technique. Cette problématique de droits d'auteur vient rejoindre une autre réflexion, celle de la pérennité nous avait conduit vers l'utilisation d'outils « open source » et c'est à leur approche que nos problématiques ont trouvé leur solution. Le serveur du site *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* fait uniquement appel à des technologies « Open Source » et des formats ouverts et disponibles librement. En effet, dans notre contexte de pérennisation d'un patrimoine il nous est apparu indispensable de respecter ces contraintes pour que *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* soit toujours accessible dans le futur et qu'il puisse évoluer avec l'évolution des (toujours) Nouvelles Technologies de L'Information et de la Communication.

## **C. Le corpus disponible**

Le balisage doit permettre de rendre compte d'une part de la structure physique du texte et du découpage en parties, chapitre, sous-chapitre et paragraphes et d'autre part de la mise en valeur du texte : notes de bas de page, italiques et pagination de l'ouvrage original.

## D. Construction de la DTD

Nous avons essayé de tirer parti des enseignements en matière de balisage structurel produits par la *Text Encoding Initiative* (TEI), plus précisément de la TEI-Lite. Ceci s'est fait au travers d'une première expérience sur un prototype dédié à Claude Bernard, qui reprenait le site existant sans l'étendre. La représentation graphique de cette DTD « Claude Bernard » est visible sur la figure 6.12. Ce type de représentation est beaucoup plus explicite que le code, car il utilise la hiérarchie des balises pour les représenter suivant un arbre.

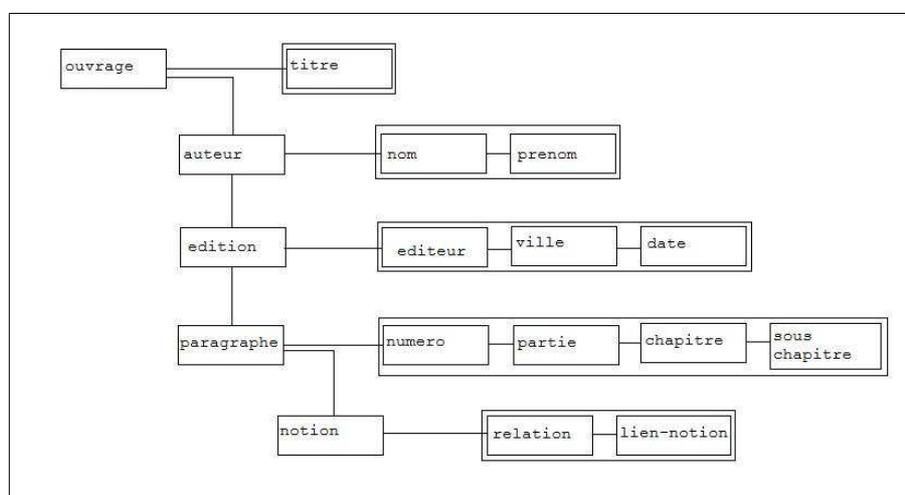


FIG. 6.12: DTD préparée pour le prototype Claude Bernard

Après avoir mené à bien un certain nombre d'expérimentations, nous avons convenu d'une DTD plus inspirée de la TEI et que nous baptiserons « DTD *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* ». C'est cette DTD qui a été communiquée à la société Jouve, responsable de la numérisation des ouvrages qui allaient constituer *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*. Elle est visible, représentée cette fois sous sa forme standard sur la figure 6.13. Les ouvrages papier ont donc été numérisés d'abord sous formes d'images (les fichiers jpeg que nous retrouverons ensuite sous la forme de facsimilé, après un mini traitement), ces images traitées par un OCR ont donné

lieu à la création d'un fichier contenant le texte, balisé en XML suivant notre « DTD *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* ».

---

```

<?xml version='1.0' encoding='UTF-8' ?>

<!ELEMENT ouvrage (titre , auteur+ , edition , tome+)>
<!ELEMENT titre (#PCDATA)>
<!ELEMENT auteur (nom , prenom)>
<!ELEMENT edition (editeur , ville , date)>
<!ELEMENT tome (chapitre* , biblio* , annexe* , toc*)>
<!ATTLIST tome numero CDATA #IMPLIED >
<!ELEMENT nom (#PCDATA)>
<!ATTLIST nom e-dtype NMTOKEN #FIXED 'string' >
<!ELEMENT prenom (#PCDATA)>
<!ATTLIST prenom e-dtype NMTOKEN #FIXED 'string' >
<!ELEMENT ville (#PCDATA)>
<!ATTLIST ville e-dtype NMTOKEN #FIXED 'string' >
<!ELEMENT date (#PCDATA)>
<!ATTLIST date e-dtype NMTOKEN #FIXED 'date' >
<!ELEMENT editeur (#PCDATA)>
<!ATTLIST editeur e-dtype NMTOKEN #FIXED 'string' >
<!ELEMENT chapitre (page+)>
<!ATTLIST chapitre titre CDATA #IMPLIED >
<!ELEMENT biblio (page+)>
<!ELEMENT annexe (page+)>
<!ELEMENT toc (page+)>
<!ELEMENT page (paragraphe* , folio* , titre.page? ,
  sous-titre.page? , note* , illustration* , numero.page)>
<!ELEMENT paragraphe (#PCDATA)>
<!ELEMENT folio (#PCDATA)>
<!ELEMENT titre.page (#PCDATA)>
<!ELEMENT sous-titre.page (#PCDATA)>
<!ELEMENT note (#PCDATA)>
<!ELEMENT illustration (#PCDATA)>
<!ATTLIST illustration e-dtype NMTOKEN #FIXED 'uri' >
<!ELEMENT numero.page (#PCDATA)>
<!ATTLIST numero.page e-dtype NMTOKEN #FIXED 'string' >

```

---

FIG. 6.13: DTD préparée pour la mise en ligne *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*

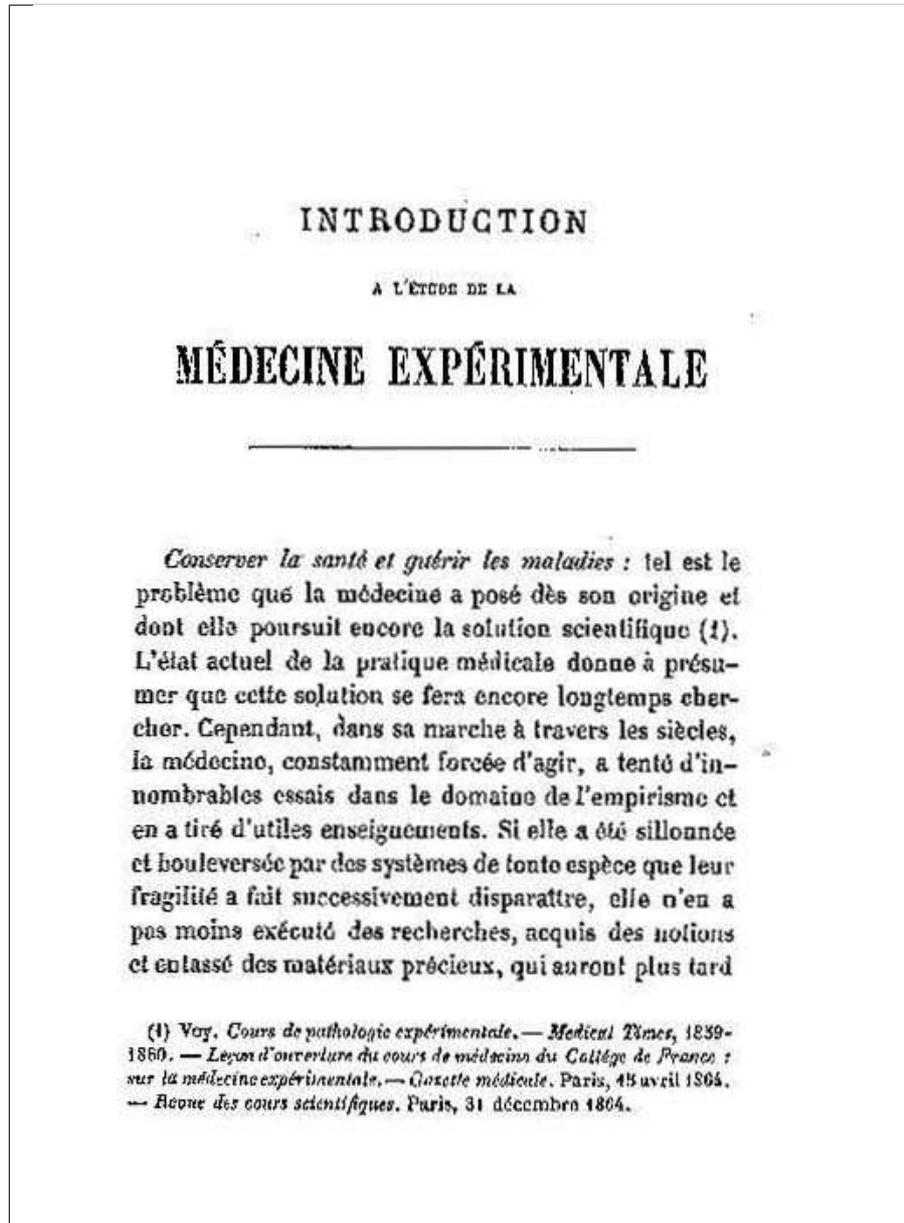
Sur la figure 6.14 est affiché le fac-similé de la page 1 de l'« *Introduction à l'étude de la médecine expérimentale* », cette page porte le titre de l'ouvrage,

un premier paragraphe ainsi qu'une note de bas de page. Nous allons devoir retrouver ces informations particulières au paragraphe en cours, assorties d'autres génériques à l'ouvrage, dans le fichier XML de description du premier paragraphe. Celui-ci est visible sur la figure 6.15, ce qui permet de voir comment le texte du paragraphe est enregistré, comment la note de bas de page est référencée et surtout, cela permet de voir les balises notionnelles qui ont été ajoutées après la numérisation.

Le texte ayant été lu et analysé, les notions ne peuvent être ajoutées que manuellement. L'édition d'un fichier XML de cette taille n'est pas particulièrement difficile, même avec un simple éditeur de texte, comme celui utilisé pour la figure 6.15, le fichier XML correspondant au premier paragraphe est ouvert dans SciTE<sup>118</sup> éditeur de texte, la seule fioriture qu'il autorise est de colorier les balises et leurs attributs pour les mettre en évidence. Toutefois la syntaxe très précise décrite par la DTD doit être scrupuleusement respectée sous peine d'erreurs de lecture lors de la mise en ligne. Il existe aussi des éditeurs spécifiques pour modifier les fichiers XML, qui prennent en charge la DTD d'une part et ne permettent que des modifications qui en tiennent compte, et bien entendu, la respectent. Dernière possibilité, nous avons développé pour un utilisateur enregistré avec des droits d'administrateur sur *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* une interface qui permet d'ajouter, d'enlever ou de modifier les notions. Bien entendu dans ce dernier cas, il n'est fait aucune allusion au stockage XML des informations, c'est uniquement les notions qui sont présentées.

---

118. Cet éditeur est distribué sous licence GNU, il peut être téléchargé depuis le site Scintilla <http://www.scintilla.org/>

FIG. 6.14: Le fac-similé de la page 1 de *IEME*

```

1 <?xml version="1.0" encoding="iso-8859-1" standalone="yes"?>
2 <?xml-stylesheet href="ouvrage.css" type="text/css" ?>
3 <!--generated by EB.-->
4 <ouvrage titre="Introduction à l'étude de la médecine expérimentale">
5 <!--(auteur+ , edition , tome+)-->
6 <auteur nom="Bernard" prenom="Claude"/>
7 <edition editeur="Librairie de l'académie impériale de médecine" ville="Paris" date="1865"/>
8
9 <paragraphe numero="1"
10 partie="Introduction" page="5" >
11 Conserver la santé et guérir les maladies : tel est le problème que la médecine a posé dès son origine
12 et dont elle poursuit encore la solution scientifique.
13 <note numappel="1">Voy. Cours de pathologie expérimentale. Médical Times,1859-1860. Leçon d'ouverture
14 du cours de médecine du Collège de France : sur la médecine expérimentale. Gazette médicale.Paris,15
15 avril 1864. Revue des cours scientifiques. Paris, 31 décembre 1864.</note>
16 L'état actuel de la pratique médicale donne à présumer que cette solution se fera encore longtemps
17 chercher. Cependant, dans sa marche à travers les siècles, la médecine, constamment forcée d'agir, a
18 tenté d'innombrables essais dans le domaine de l'empirisme et en a tiré d'utiles enseignements. Si elle
19 a été sillonnée et bouleversée par des systèmes de toute espèce que leur fragilité a fait successivement
20 disparaître, elle n'en a pas moins exécuté des recherches, acquis des notions et entassé des matériaux
21 précieux, qui auront plus tard leur place et leur signification dans la médecine scientifique. De notre
22 temps, grâce aux développements considérables et aux secours puissants des sciences physico-chimiques,
23 l'étude des phénomènes de la vie, soit à l'état normal, soit à l'état pathologique, a accompli des
24 progrès surprenants qui chaque jour se multiplient davantage.
25
26 <notion relation="opposition" lien-notion="médecine">empirisme</notion>
27 <notion relation="association" lien-notion="vie">physico-chimie</notion>
28 </paragraphe>
29 </ouvrage>

```

FIG. 6.15: Le paragraphe 1 de *IEME* sous forme de fichier XML

Fac-similé Texte Notions et Relations Lectochromie Signets Sommaire

Paragraphe: 1 Aller 636

<< < > >>

Introduction

Conserver la santé et guérir les maladies : tel est le problème que la médecine a posé dès son origine et dont elle poursuit encore la solution scientifique. <sup>1</sup> L'état actuel de la pratique médicale donne à présumer que cette solution se fera encore longtemps chercher. Cependant, dans sa marche à travers les siècles, la médecine, constamment forcée d'agir, a tenté d'innombrables essais dans le domaine de l'empirisme et en a tiré d'utiles enseignements. Si elle a été sillonnée et bouleversée par des systèmes de toute espèce que leur fragilité a fait successivement disparaître, elle n'en a pas moins exécuté des recherches, acquis des notions et entassé des matériaux précieux, qui auront plus tard leur place et leur signification dans la médecine scientifique. De notre temps, grâce aux développements considérables et aux secours puissants des sciences physico-chimiques, l'étude des phénomènes de la vie, soit à l'état normal, soit à l'état pathologique, a accompli des progrès surprenants qui chaque jour se multiplient davantage.

1 - Voy. Cours de pathologie expérimentale. Médical Times,1859-1860. Leçon d'ouverture du cours de médecine du Collège de France : sur la médecine expérimentale. Gazette médicale.Paris,15 avril 1864. Revue des cours scientifiques. Paris, 31 décembre 1864.

Notions et relations de ce paragraphe

empirisme opposition médecine  
physico-chimie association vie

FIG. 6.16: Le paragraphe 1 de *IEME* tel qu'il apparaît en ligne (parcours notionnel)

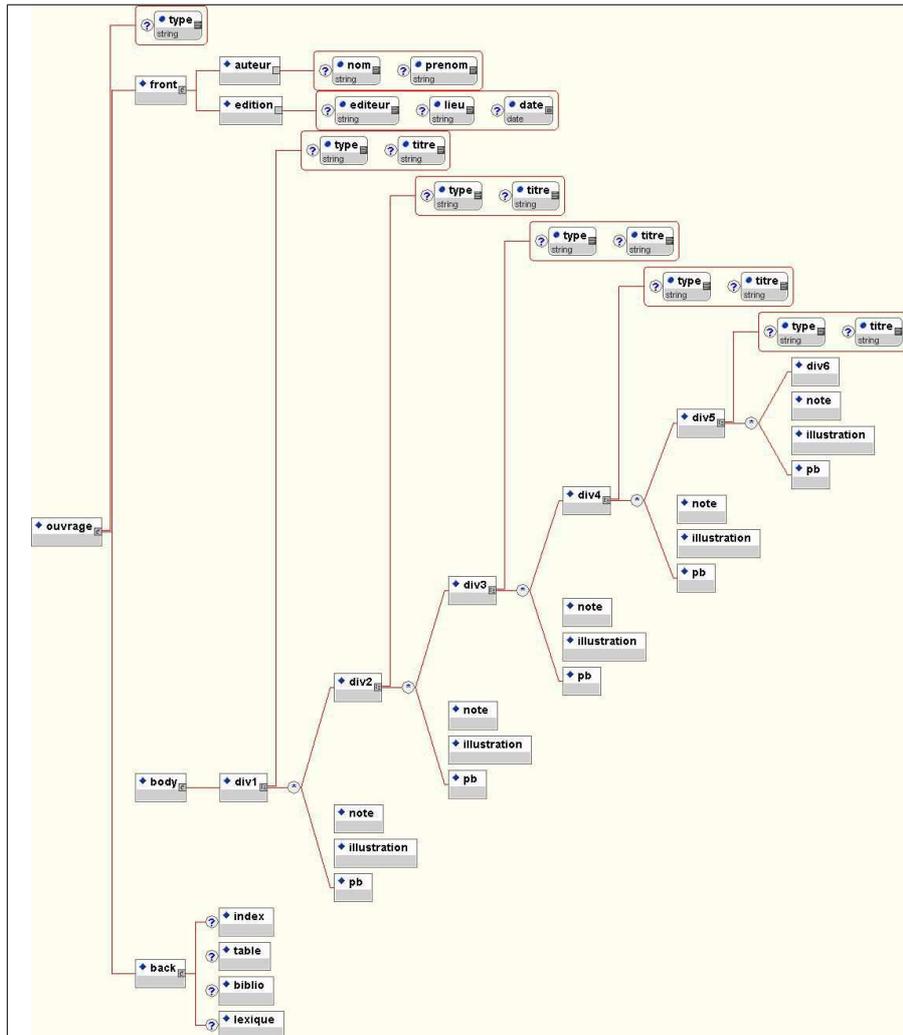


FIG. 6.17: DTD préparée pour la numérisation par Jouve



**REPRÉSENTATION DES PARCOURS DE LECTURE**

---

## Sommaire

---

<b>7.1</b>	<b>Navigation à travers <i>CoLiSciences</i></b>	<b>317</b>
<b>7.2</b>	<b>Parcours</b>	<b>326</b>
A.	Qu'est-ce qu'un parcours ?	326
B.	La co-construction des parcours de lecture	326
C.	Reconstructions du sens et parcours	327
D.	Lecture sur écran et parcours cognitifs	328
<b>7.3</b>	<b>Topologie</b>	<b>333</b>
A.	Hypertextes et repères pour la navigation	333
B.	Le cas de la bibliothèque numérique	335
C.	Représentation	337
<b>7.4</b>	<b>Lectochromie</b>	<b>342</b>
A.	Description	343
B.	Cartographie ou statistique ?	343
C.	Mise en œuvre	347
D.	Modélisation	360
E.	Analyse	362

---

## 7.1 Navigation à travers *CoLiSciences*

La première fois qu'un internaute se connecte sur *CoLiSciences* en entrant l'adresse `http://www.colisciences.net` dans le champ adresse de son navigateur, il voit s'afficher la page d'accueil de *CoLiSciences* telle qu'elle est représentée sur la figure 7.1. Cette page propose un court texte de présentation du projet *CoLiSciences*, et un certain nombre de liens répartis en deux catégories, sans que cela ne soit explicitement indiqué. Autour du ruban nous trouvons les cinq liens principaux, répartis comme les points d'une face de dé. Le point central permet d'accéder directement à la liste des ouvrages du corpus, tandis que les autres points donnent accès à diverses informations de présentation du site.

C'est toujours sur cette page que nous trouvons, à gauche de la barre bleue pâle supérieure, sous les logos des partenaires principaux du projet, un lien qui permet de s'identifier pour être reconnu et retrouver un certain nombre d'informations personnelles : signets, « étagères » d'ouvrages et enregistrement des parcours nominativement. L'objectif de proposer un grand nombre d'outils facilitant le travail collaboratif en ligne n'a pas encore été atteint et cette fonctionnalité n'est pas du tout exploitée comme elle aurait du l'être. Parmi les dispositifs que nous pouvons trouver traditionnellement sur les sites Web qui offrent ce type de service, nous pouvons citer : la lettre d'annonce, l'historique des lectures (encore que ceci existe d'une certaine manière dans la lectochromie qui sera vue page 151), la prise de notes, les commentaires.

La figure 7.2 montre cet écran de connexion qui permet dans un premier temps d'accéder à la page de modification d'un utilisateur enregistré que nous retrouvons sur la figure 7.3. Nous pouvons remarquer que le mot de passe

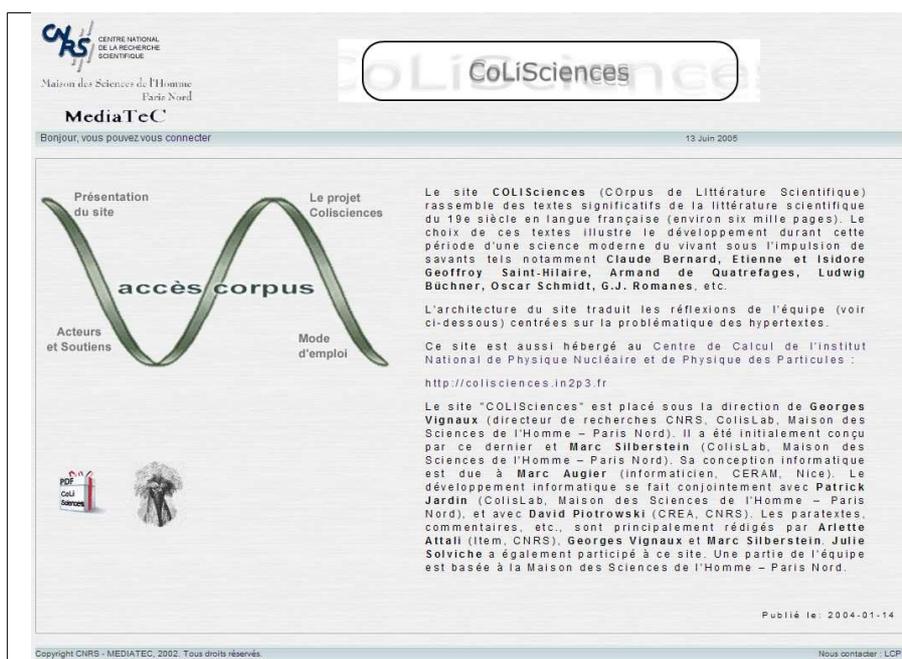


FIG. 7.1: Page d'accueil *COLiSciences*

apparaît masqué de manière à le protéger des regards indiscrets pendant sa frappe, en revanche il n'est pas fait appel à un protocole de communication sécurisé et crypté, ce mot de passe est donc transmis en clair au serveur qui héberge *COLiSciences*.



FIG. 7.2: S'identifier sur *COLiSciences*

Comme nous le disions, une fois connecté, le premier écran qui est affiché est celui qui permet à l'internaute de modifier ses informations personnelles : mot de passe, adresse de courriel, etc. Il est aussi possible d'effacer complètement son compte, pour ne plus apparaître parmi les utilisateurs

de *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*. La possibilité de « sélectionner un thème » permet de personnaliser l’affichage des pages de *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* parmi un certain nombre de propositions, cela permet d’adapter l’aspect visuel à ses goûts, mais aussi à des contraintes techniques ou personnelles : affichage plus contrasté, plus gros pour les mal voyants par exemple. Pour l’instant cette possibilité n’est pas complètement exploitée par *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* et les thèmes proposés restent très limités, seul le thème principal ayant fait l’objet d’un développement approfondi et réellement abouti.

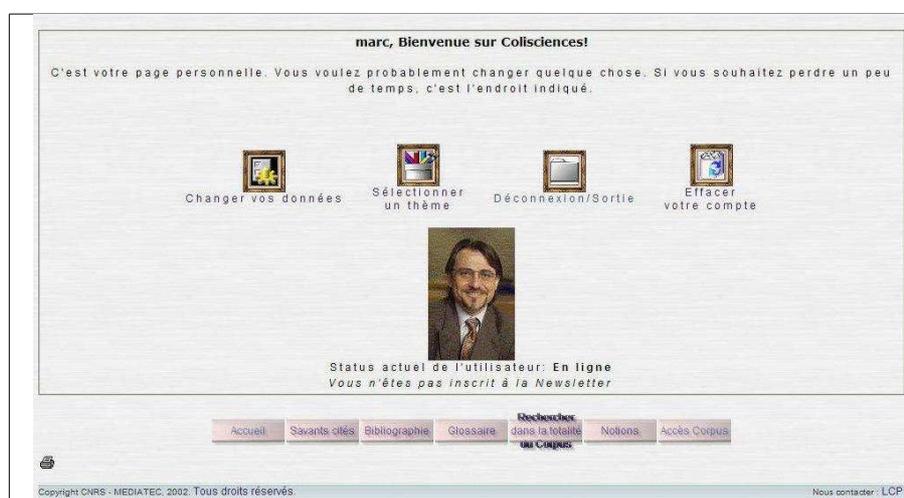


FIG. 7.3: Les informations personnelles

Si de là nous retournons à la page d’accueil, nous pourrions constater sur la figure 7.4 qu’elle est quasiment identique, au nom de la personne connectée près. Ce nom s’affiche à la place du lien qui suggérerait de se connecter, il permet de retourner sur la page de modification que nous avons vue figure 7.3. La fonction d’appropriation qui est sous-jacente dans la possibilité de s’identifier sur un site n’est pas menée jusqu’à son terme. Il faudrait sans doute que soit affichée sur cette page de garde personnalisée plus d’informations par rapport aux visites précédentes, proposer de reprendre une lecture en cours et toutes

autres informations qui permettraient à l'internaute-lecteur de se sentir chez lui.

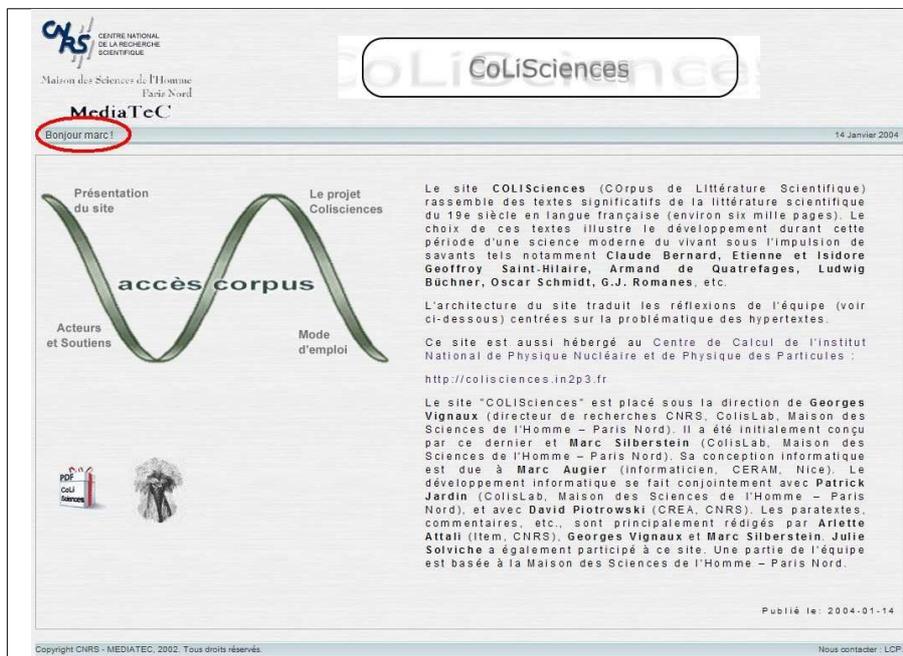


FIG. 7.4: Page d'accueil *COLiSciences* après identification

C'est donc à partir de cette page d'accueil qu'utilisateurs enregistrés et simples internautes anonymes vont pouvoir accéder au corpus. Si nous cliquons le lien central, nous obtenons la liste des ouvrages disponibles, qui se présente triée d'abord par auteur et ensuite par ordre chronologique pour chacun d'entre eux. Cette liste ne s'affiche pas dans un simple écran, nous l'avons représentée en entier sur la figure 7.5, mais un internaute ne pourra y accéder que de manière tronquée, en la faisant défiler sur son écran d'affichage, ce qui peut être perturbant pour appréhender le corpus dans sa totalité.

Pour essayer de faciliter cette navigation, il est possible de masquer une partie des ouvrages, auteur par auteur. Il suffit pour cela de cliquer le bouton (de couleur verte) situé à gauche du nom d'un auteur pour que la liste de ses ouvrages se trouve masquée; seul le nom de l'auteur est alors apparent

Fac similé, Textes, Notions & Relations  
Accès au corpus hypertextuel

Les Auteurs	Les Disciplines	Les Domaines	Les Notions
-------------	-----------------	--------------	-------------

**Les ouvrages par auteur**

- **Claude Bernard**
  - "De l'origine du sucre dans l'économie animale" (1848, article)
  - "De la présence du sucre dans le foie" (1848, article)
  - "Influence du système nerveux sur la production du sucre dans l'économie animale" (1849, article)
  - "Présence du sucre dans les matières vomies par un diabétique" (1849, article)
  - "Autopsie d'un diabétique" (1850, article)
  - "Chiens rendus diabétiques" (1850, article)
  - "De l'assimilation du sucre de canne" (1850, article)
  - "Du sucre dans l'oeuf" (1850, article)
  - "Influence de la section des pédoncules cérébelleux moyens sur la composition de l'urine" (1850, article)
  - "Note sur la présence du sucre dans l'urine du fœtus et dans les liquides amniotique et allantoïdien" (1851, article)
  - "Sur les causes de l'apparition du sucre dans l'urine" (1851, article)
  - *Recherches sur une nouvelle fonction du foie considéré comme organe producteur de matière sucrée chez l'homme et les animaux* (1853, livre)
  - "Expériences instituées pour déterminer dans quelles conditions certaines substances qui sont habituellement gardées par le sang passent dans les urines" (1854, article)
  - "Note sur la présence du sucre dans le sang de la veine porte et dans le sang des veines hépatiques" (1855, article)
  - "Remarques à propos de la communication de M. Lehmann" (1855, article)
  - "Remarques sur la sécrétion du sucre dans le foie faites à l'occasion de la communication de M. Lehmann" (1855, article)
  - "Sur le mécanisme de la formation du sucre dans le foie" (1855, article)
  - "Sur les phénomènes glyco-géniques du foie" (1856, article)
  - "Remarques sur la formation de la matière glyco-génique du foie" (1857, article)
  - "Sur le mécanisme physiologique de la formation du sucre dans le foie (suite)" (1857, article)
  - "Nouvelles recherches expérimentales sur les phénomènes glyco-géniques du foie" (1858, article)
  - "De la matière glyco-gène considérée comme condition de développement de certains tissus, chez le fœtus, avant l'apparition de la fonction glyco-génique du foie" (1859, article)
  - "Sur une nouvelle fonction du placenta" (1859, article)
  - "De la matière glyco-gène chez les animaux dépourvus de foie" (1860, article)
  - "Recherches sur l'origine de la glyco-génie dans la vie embryonnaire. Nouvelles fonctions du placenta" (1860, article)
  - *Introduction à l'étude de la médecine expérimentale* (1865, livre)
  - "Critiques expérimentales sur la glycémie des conditions physico-chimiques et physiologiques à observer pour la recherche du sucre dans le sang" (1876, article)
  - "Critique expérimentale sur le mécanisme de la formation du sucre dans le foie" (1877, article)
  - *La Science expérimentale* (1878, livre)
- **Armand De Quatrefages**
  - *L'espèce humaine* (1877, livre)
  - *Introduction à l'étude des races humaines* (1887, livre)
  - *Les émules de Darwin, tome 1* (1892, livre)
  - *Les émules de Darwin, tome 2* (1892, livre)
  - *Préface : Les émules de Darwin* (1892, livre)
- **Isidore Geoffroy Saint-Hilaire**
  - *Traité de tératologie, tome 1* (1832, livre)
  - *Préface : Traité de tératologie* (1836, livre)
  - *Traité de tératologie, tome 2* (1836, livre)
  - *Traité de tératologie, tome 3* (1836, livre)
- **George John Romanes**
  - *L'intelligence des animaux. Tome premier - Les animaux inférieurs* (1887, livre)
  - *L'intelligence des animaux. Tome second - Les vertébrés* (1887, livre)
  - *Préface : l'intelligence des animaux* (1887, livre)
- **Louis Büchner**
  - *Conférences sur la théorie darwinienne* (1869, livre)
- **Oskar Schmidt**
  - *Descendance et darwinisme* (1873, livre)
- **Rodolphe Radau**
  - "L'origine de l'homme d'après Darwin" (1874, article)

[Accueil](#)
[Savants cités](#)
[Bibliographie](#)
[Glossaire](#)
[Recherches dans la totalité du corpus](#)
[Notions](#)
[Accès Corpus](#)

Copyright CNRS - MEDIATEC. 2002. Tous droits réservés. Nous contacter : LCP

FIG. 7.5: L'accès au corpus par la liste des auteurs

et le bouton devient de couleur rouge. Les noms des auteurs sont des liens qui permettent d'accéder à une page contenant une biographie, une bibliographie et d'autres informations le concernant. Nous pouvons aussi remarquer, toujours sur la figure 7.5, les quatre onglets qui permettent de modifier la manière d'appréhender le corpus : par auteur, par domaine ou par discipline, et enfin ce qui constitue la vraie innovation de *C<sup>o</sup>Li<sup>s</sup>ciences*, par les notions qui sont traitées à l'intérieur du corpus. En revanche, les possibilités de personnalisation de l'affichage de chacune de ces listes sont assez pauvres. En particulier, il n'est pas possible de changer l'ordre d'apparition des ouvrages, par exemple les trier par titre plutôt que par date dans les listes. Sur les écran de listes par discipline (visible figure 7.6) ou par domaine (visible figure 7.7) il n'est pas possible non plus de trier d'abord par date ou titre, c'est obligatoirement l'auteur qui est utilisé comme deuxième critère.



FIG. 7.6: L'accès au corpus par disciplines

Ces détails montrent qu'au delà de l'aspect visuel, les contraintes techniques pour générer toutes sortes de listes, afficher les résultats avec de



FIG. 7.7: L'accès au corpus par domaines

nombreux critères de tris et de sélection, ne sont pas simples à résoudre et demandent du temps et des ressources. L'accès au corpus est une sorte de goulot d'étranglement à travers lequel tous les lecteurs doivent obligatoirement passer et il est indispensable d'en faciliter la fluidité du transfert vers l'intérieur, c'est-à-dire vers un ouvrage.

Il reste enfin la possibilité d'entrer dans le corpus en suivant l'étude d'une notion et le parcours érudit qui aurait été préparé autour d'elle. C'est là une véritable nouveauté qui ne met plus ni l'auteur, ni l'ouvrage en avant, mais bien des idées et leur construction à travers différentes œuvres. La figure 7.8 montre l'écran d'accueil de ce dispositif.

Les possibilités de lancer un parcours notionnel sont les mêmes que celles que nous retrouverons ensuite à l'intérieur d'un ouvrage, mais ici elles s'appliquent à la totalité du corpus. Il est donc possible de démarrer un parcours sur une notion qui apparaît soit comme principale, soit comme secondaire, soit indifféremment principale ou secondaire. Le résultat de cette

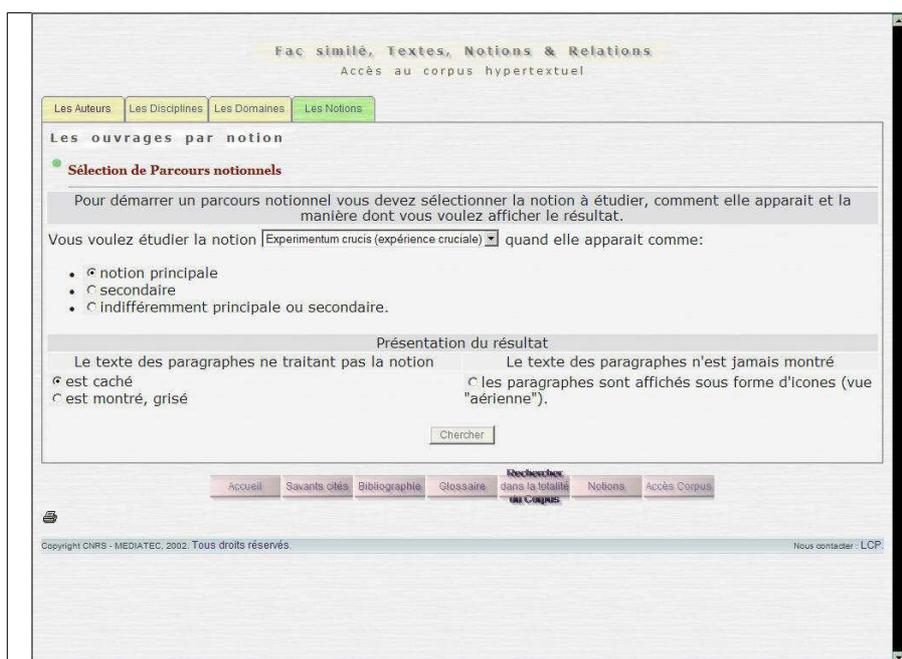


FIG. 7.8: L'accès au corpus par parcours notionnel

requête peut être affiché de différentes manières : soit seulement le texte traitant de la notion est affiché, soit tout le texte est affiché mais le texte qui ne traite pas de la notion est grisé, soit une vue « aérienne » est proposée, c'est-à-dire que les ouvrages sont affichés symbolisés avec un carré de couleur par paragraphe. Si le paragraphe ne traite pas de la notion étudiée il est affiché en noir, si la notion est principale le carré est bleu et si la notion est secondaire le carré est blanc. Ce dernier dispositif permet donc d'avoir une vision d'ensemble du corpus et de chaque ouvrage dans le corpus, dans la pratique le nombre d'ouvrages en ligne a finalement rendu cette option inutilisable sur l'ensemble du corpus, mais elle peut être utilisée sur une « étagère », c'est-à-dire un nombre d'ouvrages plus réduits, sélectionnés par un utilisateur identifié et enregistrés au sein d'une étagère virtuelle.

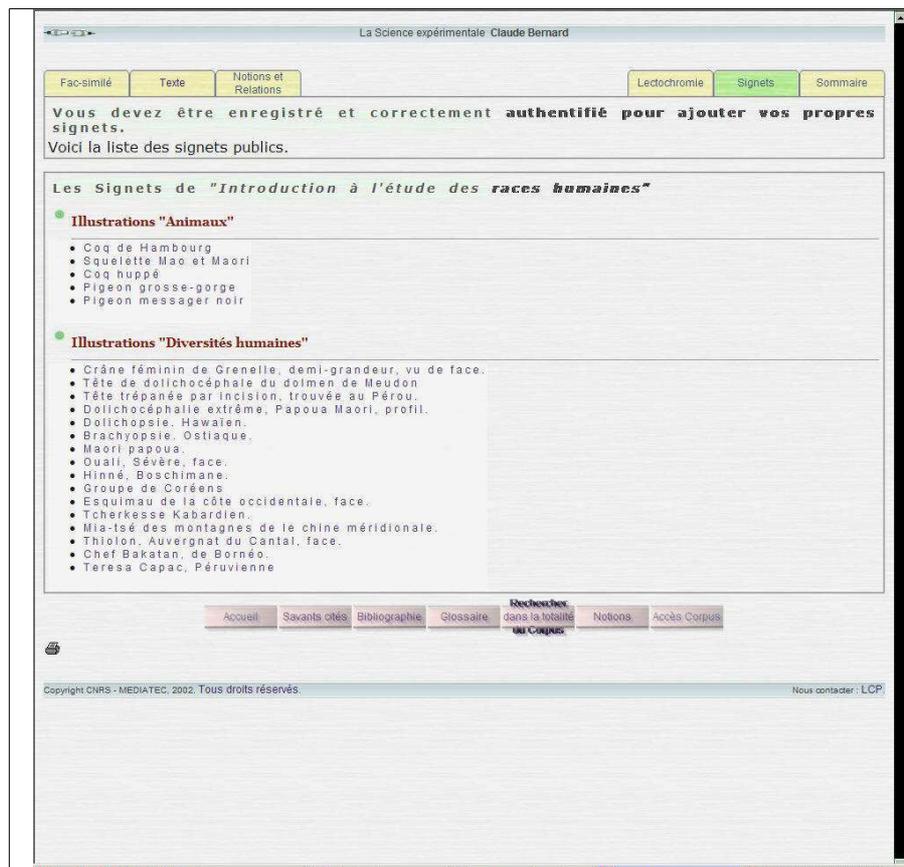


FIG. 7.9: Le liste des signets disponibles sur le corpus

## 7.2 Parcours

### A. Qu'est-ce qu'un parcours ?

Le parcours de lecture semble être une notion qui va de soit avec un système hypertextuel, qu'en est-il réellement ?

Nous devons nous poser plusieurs questions : quand est-ce que le parcours de lecture commence ? A partir du moment où le lecteur arrive sur le site ou après qu'il ait choisi un ouvrage à lire ? Quand le lecteur navigue d'un ouvrage à l'autre, s'agit-il du même parcours, ou devons nous considérer que plusieurs parcours s'enchaînent ? Enfin, de nombreux sites présentent leur contenu sous la forme d'un « plan de site », c'est-à-dire qu'ils utilisent une typologie cartographique pour aider les internautes à retrouver leur chemin parmi le contenu. Au-delà des termes : cartographie, plan, parcours, en quoi la métaphore du déplacement physique peut-elle aider à la lecture ?

### B. La co-construction des parcours de lecture

Nous pensons comme Francisco Varela que

*« L'idée fondamentale est donc que les facultés cognitives sont inextricablement liées à l'historique de ce qui est vécu, de la même manière qu'un sentier au préalable inexistant apparaît en marchant. (Varela, 1996, page 111) »*

et par voie de conséquence, le parcours à travers un corpus hypertextuel est un témoin de l'historique des interactions du lecteur à travers l'ouvrage, en quelque sorte le témoin du sens construit par le lecteur. Les recherches menées sur la compréhension de textes montrent que toute lecture est une individuation sémantique, c'est-à-dire une réécriture interne du texte lu. Lire

un texte signifierait donc le réécrire intérieurement par une série d'allers-retours entre l'anticipation de ce qui va suivre et l'intégration de ce qui précède. L'enregistrement des parcours de différents lecteurs doit mettre en évidence, doit faire émerger ce sens qu'ils ont construit par leur lecture. Ce sont donc les parcours, ceux établis *a priori* par l'auteur et ceux créés à postériori par les lecteurs qui vont témoigner du réseau cognitif mis en place par l'hypertexte. Une remarque s'impose immédiatement : les parcours des lecteurs étant constamment en évolution, l'hypertexte vu sous ce jour n'est plus un objet technique figé mais une entité cognitive toujours remise en jeu.

### **C. Reconstructions du sens et parcours**

Le lien hypertexte permet au lecteur de découvrir quelle réalité se dissimule derrière un titre ou un document et de créer en conséquence, parfois immédiatement, une connexion vers un nouveau sous-ensemble cognitif. En conséquence, le concepteur d'hypertexte doit savoir contrôler le contenu de chaque document hyperlié en le rapportant non pas à une progression linéaire mais à un parcours qui peut être aléatoire, fragmenté, décontextualisé. Il est donc nécessaire de penser par avance ces différents scénarios de lecture et stratégies d'exploration. Enfin, un problème structurel relatif aux hyperliens est celui de la surcharge cognitive : il faut éviter au lecteur-utilisateur de se perdre dans l'arborescence du site. Il s'agit d'une préoccupation centrale dans nos expérimentations à venir : quelles stratégies d'économie cognitive peut-on observer selon les lecteurs et leur finalités ? A quel moment se perd-on dans un hypertexte ? Comment s'y retrouver ?

La psychosociologie des interactions montre que les situations de communication comme l'entretien ou l'audition d'une émission impliquent chez le récepteur des opérations mentales telles que le filtrage des informations en

fonction de ce qu'il juge pertinent, le remodelage de ces informations selon ses propres présupposés et éventuellement l'ajout d'unités de sens destinées à « compléter les blancs ». Les études comportementales<sup>119</sup> ont montré qu'un très grand nombre de lecteurs sélectionnaient les pages à lire et à écarter en fonction des toutes premières lignes. C'est pourquoi, dans la mise en page hypermédia, la présence de liens répartis de façon équilibrée dans la page est un facteur puissant de dynamisation de la lecture. Le lien hypertexte permet au lecteur de découvrir quelle réalité se dissimule derrière un titre ou un document et de créer une connexion vers un nouveau sous-ensemble cognitif. En conséquence, le concepteur d'hypertexte doit savoir contrôler le contenu de chaque document hyperlié en le rapportant non pas à une progression linéaire mais à un parcours qui peut être aléatoire, fragmenté, décontextualisé. Il est nécessaire de penser par avance ces différents scénarios de lecture et stratégies d'exploration. Enfin, un problème structurel relatif aux hyperliens est celui de la surcharge cognitive : il faut éviter au lecteur-utilisateur de se perdre dans l'arborescence du site.

#### **D. De la lecture sur écran aux parcours cognitifs**

Avec la lecture sur écran, ce sont régulièrement les problèmes d'ergonomie et de surcharge cognitive qui sont évoqués comme freins à l'adoption de ce nouveau média. Bien entendu, nous retrouvons là encore le clivage habituel entre technophiles et technophobes, qui a accompagné la grande majorité des

---

119. Il est possible de consulter sur le Web les conclusions d'une étude de Jakob Nielsen, éditée par Sun Microsystems : *Designing Web Usability* à l'adresse <http://www.sun.com/980713/webwriting/wftw9.html>.

L'étude de l'Université de Stanford, précitée, est complètement accessible ici : <http://www.poynter.org/eyetrack2000/index.htm>.

Dans un article bien documenté de l'Ergonome, « Quelques recommandations pour la rédaction de contenus Web », disponible à l'adresse [http://www.lergonome.org/pages/detail\\_articles.php?indice=14](http://www.lergonome.org/pages/detail_articles.php?indice=14), il est question de lisibilité des textes à l'écran

innovations. Les premiers affirmant que cette évolution ouvre de nouvelles perspectives à l'écrit et les seconds redoutant qu'une *fracture numérique* vienne s'installer entre ceux qui sont capables d'assimiler le nouveau média et les autres. Dans ce contexte, les anglo-saxons ont forgé un nouveau terme, difficile à traduire en français, il s'agit de *computer literacy*, c'est-à-dire la capacité qu'aurait une personne à savoir lire et écrire avec un ordinateur. Une sorte d'*alphabétisation informatique*, qui serait l'étape suivant l'alphabétisation telle qu'elle nous a préoccupé jusqu'à aujourd'hui.

Les transformations précédentes de l'écrit (passage du volumen au codex, du manuscrit à l'imprimé, invention de la typographie et de convention d'impression) ont conduit à des transformations dans la manière de lire. L'exemple le plus probant étant le passage de l'oralisation à la lecture silencieuse. Ce qui est la norme aujourd'hui ne l'était pas au IV<sup>e</sup> siècle de notre ère, comme le rapporte Augustin dans ses Confessions (VI,3), après une visite à l'évêque de Milan Ambroise il fait part de son étonnement devant un fait pour lui extraordinaire :

« *Quand il lisait, ses yeux couraient les pages dont son esprit perceait le sens ; sa voix et sa langue se reposaient. Souvent en franchissant le seuil de sa porte, dont l'accès n'était jamais défendu, où l'on entrait sans être annoncé, je le trouvais lisant tout bas et jamais autrement.* (Augustin, 1864, Livre VI, Chapitre 3) »

Mais avec le recul, il nous semble aujourd'hui que la transition se soit effectuée sans rupture et qu'elle était positive, puisqu'elle a conduit au modèle du livre actuel, que nous acceptons donc tous aujourd'hui comme étant le meilleur puisque nous n'en connaissons pas d'autre. Le livre électronique, pour convaincre, doit franchir plusieurs obstacles :

- Pour l’instant il est cher et repose sur des technologies qui se démodent très rapidement ;
- il s’utilise différemment d’un livre imprimé, ce qui fait dire qu’il est peu pratique ;
- il n’a pas une réelle objectivation parce qu’il reste un produit commercial diffusé par des entreprises en concurrence qui n’ont pas encore bien cerné la cible, ni décidé d’une plate-forme commune ;
- et pour terminer avec l’obstacle principal, la lecture sur écran est encore un terme qui fait peur.

La citation suivante permet de mettre en lumière que même les auteurs qui cherchent à ne pas entourer la lecture sur écran d’une aura négative restent très circonspects sur la manière dont l’évolution doit se faire ou se fera.

*« La numérisation des textes est trop souvent perçue comme une évolution naturelle et anodine du mode d’écriture. Le texte n’est plus sur une page mais sur un écran, et il convient de faire défiler le curseur comme on fait défiler les feuillets.*

*Un tel bouleversement des mentalités de lecture s’est déjà opéré lors du passage du volumen au codex.[...] Cependant les enjeux sont autres. On n’a pas seulement changé de support mais aussi changé de regard. En effet, on doit pouvoir attendre d’un texte numérisé la même organisation que celle d’un texte « papier ». Libéré de son support d’in quarto ou d’in octavo, le texte numérisé ne doit pas pour autant perdre sa hiérarchie, sa logique. (Hudrisier et Hudrisier, 2000, page 45) »*

Le modèle de texte numérique présenté ici est « monopage », ce qui est bien réducteur. Comme nous l’avons déjà discuté<sup>120</sup>, un ouvrage numérique peut être lu à travers divers dispositifs, certains présenteront le texte sur une

---

120. Dans la partie I - **Du livre au Web**, qui commence page 53.

seule page qu'il faudra faire défiler, d'autres proposeront une interface qui permettra une lecture page par page, ou plutôt écran par écran, le but de ce deuxième type de dispositif étant justement d'avoir à éviter de faire défiler le texte sur l'écran pour ne pas perturber la lecture. Le but premier du dispositif de lecture, qu'il soit livre ou qu'il soit ordinateur, est de s'effacer derrière le contenu, d'être au service des idées qu'il contient. La structuration qui a été mise en place sur les livres n'est pas intrinsèquement liée au livre. Elle est liée à notre manière de concevoir la lecture et destinée à la faciliter. Nous pensons que si la hiérarchie et la logique des ouvrages doivent évoluer ce ne sera pas *à cause* mais **grâce** au numérique, car cette évolution se fera pour tirer parti des nouvelles possibilités du médium, sous la volonté des éditeurs et des lecteurs, comme cela s'est déjà fait lors du passage du volumen au codex.

### 1. Typologie des parcours et Navigation

L'hypertexte permet un processus d'exploration et éventuellement d'appropriation des relations spatiales et cognitives entre différentes informations : c'est un véritable mode d'exploration. Prenons l'exemple d'une découverte urbaine : on peut visiter une ville à partir d'un plan avec un programme de lieux où se rendre. Mais en présence d'une intersection ou d'un carrefour inattendus, certains visiteurs se laisseront guider par le contexte et changeront de direction : c'est la déambulation qui crée alors une nouvelle topographie de l'exploration. De même, lors d'une requête dans un hypertexte, tout élément de réponse induit une forme supérieure d'élaboration du questionnement et le lecteur collabore avec le programme pour générer un processus évolutif. L'interface assume ainsi un rôle de miroir en obligeant le chercheur à reformuler sa problématique et à en éliminer les aspects non pertinents, ce qui

permet de poser la question des interactions entre types de requêtes et types de représentations construites sur l'hypertexte.

## **2. La compréhension de textes**

Les travaux consacrés à l'étude de la compréhension de textes au cours de la lecture ont connu trois périodes historiques. Dans les années 70, l'accent portait sur les produits de cette compréhension notamment sous la forme de résumés demandés aux enfants ou aux adultes et supposés exprimer ce qu'ils avaient compris ou ce dont ils se rappelaient. Vers le milieu des années 1980, l'intérêt s'est porté vers le traitement des mots, des phrases, des inférences que les lecteurs produisent et aussi vers la capacité attentionnelle ou mémorielle requise. Nombre de travaux vont faire alors référence à une entité particulière, la mémoire de travail, censée maintenir actives les informations et les traiter (Baddeley, 1986). Ces perspectives de recherche ont établi la nécessité d'explorer les interactions entre les caractéristiques « fines » des textes, celles des lecteurs définies en termes de connaissances préalables et l'élaboration progressive de la représentation mentale d'ensemble. Il apparaît ainsi que les organisations conventionnelles de textes ne facilitent la compréhension que dans la mesure où les individus ont eu fréquemment l'occasion de telles structures au cours d'activités antérieures et significatives. Ces organisations textuelles sont souvent associées à des configurations de marques linguistiques qui en permettent le repérage, lequel repérage de marques déclenche la mise en œuvre de modalités spécifiques de traitement. C'est ce rapport entre systèmes de marques et représentations cognitives qui en résultent ou qu'elles traduisent que nous entendons explorer et modéliser. Méthodologiquement en particulier, nous entendons nous livrer à des confrontations expérimentales de l'expertise humaine avec des moteurs sémantiques, susceptibles de procé-

der à des découpages sémantiquement pertinents des textes, afin de tendre vers une semi-automatisation de la procédure d'extraction des termes ou des thèmes en vue de la réalisation d'un hypertexte pleinement développé.

## 7.3 Topologie

Si nous parlons de parcours, nous parlons aussi de voyage à travers un territoire. Avant d'aborder la cartographie au travers de la lectochromie que nous décrivons dans la section 7.4, nous devons nous attarder sur les différentes topologies de sites Web, comment ces topologies peuvent s'adapter à une bibliothèque numérique en ligne et enfin comment sont-elles représentées.

### A. Hypertextes et repères pour la navigation

Pour éviter que l'utilisateur ne se perde dans un site, il importe d'élaborer une structure de visite cohérente et soigneusement hiérarchisée. Un document papier permet à tout moment de revenir à une vue d'ensemble de l'ouvrage. Cela est impossible avec l'écriture électronique qui se présente généralement par pages distinctes et induit parfois le sentiment à tel ou tel point de lecture, de ne plus pouvoir se situer par rapport à une problématique d'ensemble. A l'instar du plan d'une ville qui devrait comporter une indication permettant au lecteur de se localiser par rapport à un espace plus vaste (« vous vous trouvez ici »), chaque page d'un site devrait permettre à l'internaute non seulement d'identifier les liens qu'il a déjà ouverts, mais de se situer par rapport à l'arborescence. Notamment, de pouvoir à tout moment revenir au menu général ou au menu précédent afin de réorganiser son circuit précédent. Une confusion courante consiste à renvoyer en guise de plan du site vers la home page qui contient le menu général. Or, celle-ci ne comporte en principe

que le sommaire des grands chapitres et non pas l'arborescence en profondeur. Un plan de site devrait en vérité s'apparenter à la table des matières analytique d'un ouvrage plutôt qu'au sommaire. Or de nombreux sites se contentent de se présenter sous la forme présentée à la figure 7.10.

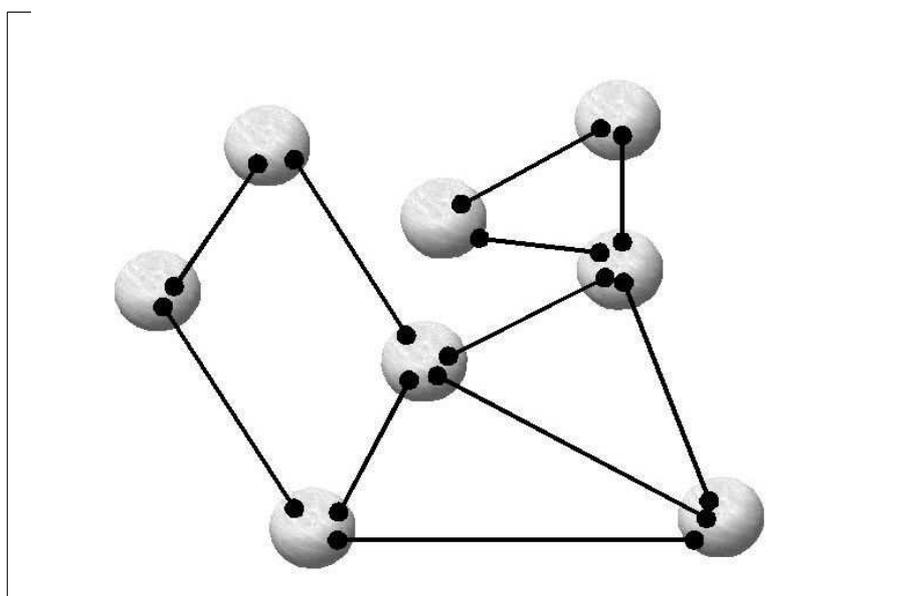


FIG. 7.10: Arborescence sans organisation hiérarchique

Dans un réseau tel que ci-dessus, les pages sont liées de manière factuelle sans aucune possibilité d'accéder à des niveaux inférieurs ou supérieurs comme il en va entre classes thématiques subordonnées les unes par rapport aux autres. Rien n'y ressort des pratiques et méthodes de structuration de la pensée. En vérité, s'agissant de constituer un site dont le corpus est déjà existant (documentation, dossier, corpus d'auteur), la première étape consiste à définir des zones de frontière thématiques même flexibles. L'exemple de la figure 7.11 montre comment schématiquement pourraient s'articuler différentes sections, chapitres et pages.

Observons qu'un organigramme d'hyperliens peut développer des lignes de communication verticales ou transversales. D'où les questions :

1. Peut-on se déplacer à loisir dans cette structure ?

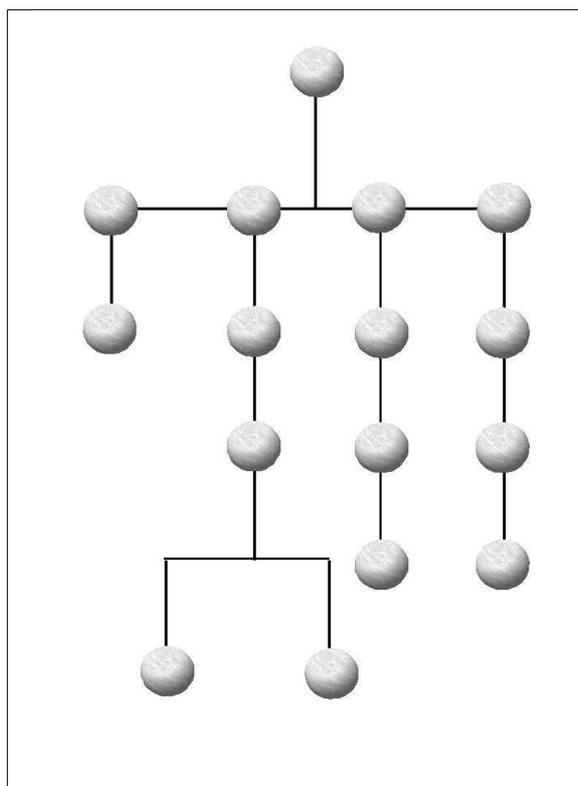


FIG. 7.11: Arborescence en organigramme

2. Est-ce que telle page de niveau secondaire peut renvoyer à une rubrique primaire qui ne lui est pas directement supra-ordonnée ? Etc.
3. Un émiettement excessif des rubriques du sommaire ; cette structure de choix qui se veut panoramique n'a-t-elle pas pour inconvénient une trop grande non-directivité.

## B. Le cas de la bibliothèque numérique

La bibliothèque numérique est en prise directe avec la problématique de la rupture entre ouvrage papier et ouvrage numérique. Parce qu'elle représente la porte de passage vers la nouvelle médiatisation des ouvrages, c'est à la bibliothèque numérique de résoudre la question de savoir comment présenter un ouvrage, une fois qu'il a été numérisé. Il est clair que le format d'un écran

n'est pas celui d'une page de livre, livre qui présente d'ailleurs deux pages à la fois. Il est moins clair qu'il n'est pas possible de contrôler complètement comment les pages sont affichées dans un navigateur Web, à moins d'imposer de lourdes contraintes aux internautes mais de plus en plus de sites commencent à prendre aussi cette dimension en compte. La bibliothèque numérique va ainsi proposer à la lecture des documents numériques sous différents formats, suivant des choix éditoriaux d'un nouveau type, introduits par l'utilisation du Web comme support d'édition. Nous pouvons diviser ces deux formats en deux grandes catégories : les formats « natifs » et les formats « exogènes ».

- Les formats natifs sont ceux que le navigateur Web est capable de supporter sans besoin d'installer de programmes de visualisation complémentaires, appelés *plug-ins*. Par exemple les images, quand elles sont d'un type supporté par le navigateur, les plus courants étant JPEG, GIF et PNG. Différents formats textes sont aussi supportés nativement, le plus naturel sur le Web étant HTML. La contrepartie de la facilité d'implémentation est qu'il est pratiquement impossible, ou tout du moins techniquement très complexe, d'allier la présentation originale de l'ouvrage aux possibilités plein texte que le numérique invite à utiliser. Si ce sont des images qui sont affichées, nous avons une parfaite reproduction de l'original, mais il est impossible de copier des bouts de textes et les pages ne seront pas interprétées comme des pages de texte par les moteurs de recherche. Si nous utilisons des pages HTML, le document sera la plupart du temps très différent du document original et la pagination ne pourra être respectée que difficilement. De même que les placements d'images. Il faut noter que si la pagination originale n'est pas respectée, il devient très difficile d'utiliser l'ouvrage numérique pour des citations dans un travail académique. Pour certains ouvrages cela ne

- pourrait pas avoir une importance majeure, mais il est tout de même difficile au Webmaster qui construit le site d'avoir la maîtrise de tous les champs de recherche pour identifier ceux qui sont susceptibles ou pas d'utiliser les ouvrages dans leurs travaux.
- Les formats exogènes sont ceux qui demandent l'installation d'un complément logiciel, ce qui permettra de visualiser un document PDF directement dans le navigateur, ou un document à base de Macromedia ou de Java, ou autres. Il est difficile d'établir une liste exhaustive des formats possibles, ce qui donne une idée de la complexité qui peut être imposée à un internaute qui devrait modifier son système à chaque découverte d'un nouveau site si les Webmasters ne s'imposent pas un minimum de rigueur dans leur choix. Même en utilisant un format très répandu comme PDF, il reste un certain nombre de problèmes à résoudre : transférer tout l'ouvrage en une seule fois, ce qui peut entraîner un délai de téléchargement assez long et interdire le « feuilletage », c'est-à-dire prendre un ouvrage sur une étagère électronique, lire la quatrième de couverture et le reposer. L'autre solution serait de transférer les données à transférer lors de chaque changement de page, cela entraîne une complexité technique sur la « fabrication » de l'ouvrage numérique et surtout, cela rend la lecture beaucoup moins fluide.

## **C. Représentation**

Quand la topologie est définie, il reste encore le problème de la représenter à l'écran. Les stratégies de représentations peuvent se décliner selon trois grand schémas :

1. Simplicité et abstraction
2. Information pléthorique

### 3. Fac-similé

Ces schémas peuvent être déclinés suivant des attributs techniques qui vont soit les conforter soit les tempérer :

- Conformité aux standards
- Hi-tech

## 1. Les Schémas de représentation

**Simplicité** La simplicité de la présentation est une tendance qui est née avec Internet du temps où la bande passante disponible était réellement réduite. Même si elle a été considérablement augmentée ces dernières années, le trafic ayant augmenté dans de grandes proportions lui aussi, la bande passante reste une « denrée » rare du cyberspace. La même « philosophie » de construction existe donc toujours, qui prône la simplicité pour économiser les ressources électroniques, pour faciliter l'accès aux documents, du fait que le fond est mis en valeur au lieu d'être masqué par une forme omniprésente et enfin pour permettre l'accès au plus grand nombre, ce type de présentation étant plus facilement interprétable par un plus grand nombre de navigateurs : graphiques ou textuels, voire même audio, utilisés par les non-voyants. Un exemple est proposé sur la figure 7.12.

**Encyclopédisme** Les possibilités de stockage, de liens croisés entre plusieurs sites, de commentaires dynamiques de la part des lecteurs font qu'une façon de penser la bibliothèque numérique se situe dans la construction d'un espace en ligne regroupant le maximum d'information qu'il soit possible de compiler. C'est le rêve de bibliothèque universelle de Ted Nelson qui a donné naissance au projet Xanadu, le même que celui qu'avait fait Paul Otlet quelques décen-



FIG. 7.12: Public@tions Électroniques de Philosophi@ Scient@e

nies auparavant avec le Mundaneum<sup>121</sup>. L'idée est de reprendre la typologie développée par les grand portails informationnels comme l'incontournable Yahoo!<sup>122</sup>. La figure 7.13 montre la page d'accueil du site Revues.org<sup>123</sup> qui est un exemple de portail de bibliothèques numériques assez représentatif de cette catégorie en cherchant à fédérer des revues scientifiques (35 à ce jour).

**Mimétisme du réel** Une façon de mettre à l'aise l'internaute et de l'aider à se repérer est de lui représenter dans la fenêtre du navigateur une vue plus ou moins réaliste de la réalité équivalente. Pour une bibliothèque virtuelle, il peut être tentant de vouloir représenter une bibliothèque réelle : les salles vont servir à identifier les différentes sections ainsi que les domaines, des rayonnages vont présenter les ouvrages. Le nec plus ultra de ce type de représentation est de la rendre configurable de manière à ce que les rayonnages présentent les volumes arrangés suivant les requêtes de l'internaute. Nous

121. Nous avons évoqué ces projets précédemment dans le cadre du chapitre traitant de l'hypertexte, page 67

122. Faut-il préciser qu'il se trouve à l'adresse <http://www.yahoo.com> pour le portail international, une déclinaison localisée existant pour la plupart des pays.

123. <http://www.revues.org>



FIG. 7.13: Bibliothèque et Portail

avons ainsi la fusion des avantages de la représentation traditionnelle avec une configuration dynamique. L'exemple le plus abouti de représentation virtuelle est visible sur le site de ActiveWorlds<sup>124</sup>, ce dispositif n'est pas spécialisé dans la représentation de bibliothèques uniquement mais dans la construction de « réalités virtuelles ». Pour revenir aux bibliothèques, il faut souligner que ce type de représentation est finalement assez peu présent sur le Web du fait de sa lourdeur de mise en œuvre, il reste plutôt utilisé dans des applications autonomes, du type de celle proposé par KeeBook<sup>125</sup> dont on peut voir un exemple sur la figure 7.14.

Pour terminer sur ce mode de représentation, il faut aussi noter que pour l'utilisateur l'interrogation à distance revient à se promener dans les rayons d'une bibliothèque, à saisir et à parcourir les livres qui l'attirent. La

124. De nombreux mondes virtuels d'exemple, permettant de se familiariser avec cet outil, sont accessibles à l'adresse <http://www.activeworlds.com/>

125. Une application est disponible en version d'essai sur le site de l'éditeur à l'adresse <http://www.keebook.com/>

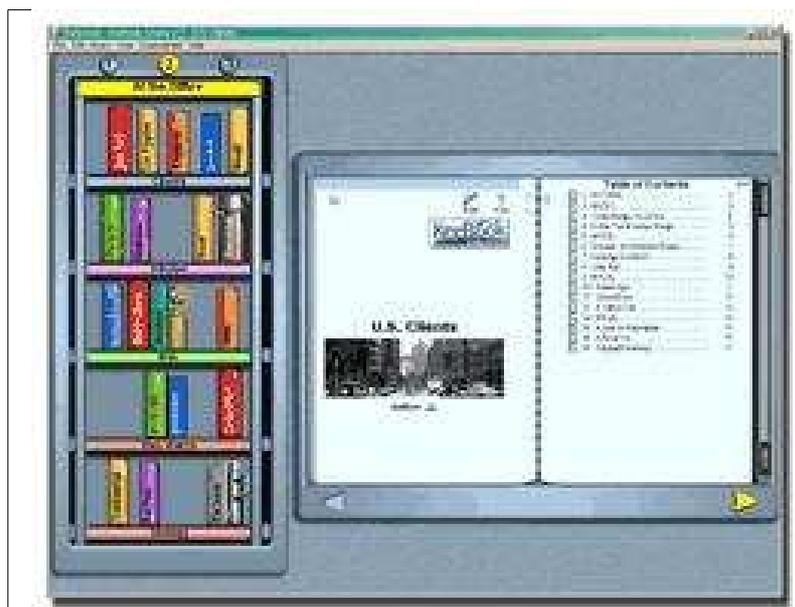


FIG. 7.14: Bibliothèque virtuelle KeeBook

représentation spatiale d'une bibliothèque réelle n'apporte rien de plus, au contraire, que ce dont un catalogue électronique bien construit est capable.

## 2. Les attributs techniques

Les attributs techniques influent très souvent de manière directe sur la représentation, de plus ils sont chargés d'un message à l'attention de l'internaute. Le message est à la fois de l'ordre du subliminal et de l'autre du message secret à l'adresse des adeptes de la même « secte » technique.

Pendant la fin du XX<sup>e</sup> siècle et le tout début des années 2000, nous avons assisté à une multiplication de sites Web : du particulier à la multinationale tout le monde se devait d'avoir une présence virtuelle sur la Toile. Dans le même temps la bataille technologique autour des standards du Web faisait rage : Netscape et Microsoft, le W3C, Macromédia et d'autres encore essayent d'imposer leurs technologies, ouvertes contre propriétaires, propriétaires entre elles. C'est à ce moment que nous avons vu souvent fleurir sur la page

d'accueil de nombreux sites des messages d'information, qui ressemblaient en fait plus à des messages de mise en garde :

« Les pages de site doivent être visualisées avec un écran de 800x600  
avec le plug-in Flash Macromedia version ... »

Ce qui semblait dire d'une part : soit vous avez le matériel et les logiciels du club et vous pouvez entrer, soit vous passez votre chemin. D'autre part, implicitement ce message était une sorte de condescendance à l'adresse de ceux qui n'étaient pas capables de se doter des outils adéquats (voire, encore plus humiliant, de les maîtriser).

## 7.4 Lectochromie

La lectochromie est un procédé de cartographie mis en place spécifiquement pour *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* pour permettre de synthétiser sur un écran les nombreuses informations de parcours qui ont été collectées au cours du temps, c'est-à-dire pour rendre visibles sur un seul document des parcours construits au cours du temps.

La lectochromie a été bâtie par combinaison de deux idées : la première est l'utilisation d'un procédé cartographique pour aider un lecteur à se situer dans ses lectures et sa navigation à travers le site *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* ; la seconde est d'utiliser un spectre dynamique de couleurs, qui traduise par un affichage chromatique l'usage, pour ne pas dire l'*usure*, qui est fait des textes numériques. Par ailleurs, il est aussi proposé un rendu des parcours de lecture dans les paragraphes du corpus qui met en scène non seulement l'usage, mais aussi la chronologie de cet usage.

## A. Description

### 1. Un système commun de référence

La lectochromie propose un système commun de références dans lequel chacun peut retrouver à la fois ses propres parcours, à travers ses différentes visites, mais aussi les parcours des autres lecteurs de *C<sub>o</sub>Li<sup>s</sup>ciences*. Chaque acteur peut ainsi partager ses expériences de lecture, sans pour autant que ce soit une démarche active comme avec l'actualisation et la mise en commun de signets par exemple.

### 2. Co-élaboration d'une matrice commune d'identification des parcours

Ce dispositif produit des lectochromes qui sont des représentations particulières de l'état d'utilisation du corpus à un instant donné. Les particularités d'un lectochrome sont qu'il est non permanent, imparfait et incomplet.

- Non permanent parce que chaque nouveau visiteur, dès sa première lecture va contribuer à bouleverser l'ordre préalablement établi.
- Incomplet parce que toujours en attente d'une nouvelle lecture qui va mettre en place de nouvelles connexions et créer un nouvel équilibre éphémère.
- Imparfait par sa construction même et parce que toute chose est imparfaite.

## B. Cartographie ou statistique ?

### 1. Statistiques de lecture

Au départ, le problème est d'enregistrer des statistiques de lecture, puis le problème devient d'afficher ces statistiques.

**Statistique ou usage?** Qu'entendons-nous par « statistique de lecture » ?

Il s'agit en fait de données accumulées au cours de leur navigation à travers le corpus sur le ou les usages que les internautes font de *CoLiSciences*. Lecture séquentielle ou butinage, voire lecture savante en faisant référence aux encyclopédies qui accompagnent le corpus textuel. Un des objectifs de la conservation de ces données est de pouvoir reconstruire un parcours spécifique, pour ensuite le proposer à d'autres lecteurs.

## 2. La cartographie pour rendre visible

Le dispositif de cartographie dynamique permet de rendre visible l'utilisation qui est faite du corpus. En effet, un des problèmes que nous essayons ici est lié à la dématérialisation des ouvrages. Le lecteur ayant perdu le contact physique avec l'ouvrage qu'il lit, perd aussi le contact avec les lecteurs précédents et ceux qui lui succéderont. La cartographie permet de redonner au lecteur la perception des ouvrages les plus ouverts, et pour chaque ouvrage quels sont les paragraphes les plus lus. Plusieurs cartes sont proposées, que nous appellerons « lectochromes ».

- Une carte des visiteurs, représentés par leur nom de connexion s'ils ont pris la peine de s'identifier, ou par leur adresse Internet s'il ne l'ont pas fait.<sup>126</sup>
- Une carte des ouvrages.

---

126. Cette gestion des noms d'utilisateurs et de leurs adresses nécessite une petite clarification technique. En fait, par adresse Internet nous entendons l'adresse IP, c'est un identifiant numérique de la forme 192.168.10.5 qui représente de manière unique une machine connectée au réseau. La lecture directe de ces adresses est presque dénuée de tout intérêt pour un humain parce qu'il est difficile de les retenir et d'identifier à travers elle une personne ou une institution. Sur la carte des visiteurs dans *CoLiSciences*, nous affichons à la place le nom de la machine, tel qu'il est connu par les serveurs DNS (*Domain Name Server*, ce sont ces serveurs qui nous permettent d'utiliser des noms de domaines tels que *www.colisciences.net* en traduisant automatiquement et de façon transparente pour les internautes les noms comme *www.colisciences.net* en l'adresse IP correspondante 213.245.116.125). Ce nom est assez générique et peut prêter à confusion. Par exemple, tous les abonnés d'un même fournisseur d'accès Internet apparaîtront avec le même nom qui est celui de ce fournisseur.

- Une carte des paragraphes d'un ouvrage.

Il est aussi possible d'approfondir une recherche, de « zoomer » en quelque sorte, depuis un visiteur, vers les ouvrages qu'il a lu, puis vers les paragraphes qu'il a lu dans un ouvrage. Il est aussi possible de visualiser le parcours d'un visiteur, c'est-à-dire la succession des étapes qu'il a suivies durant sa lecture sur *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* : lecture d'un fac-similé, puis des notions d'un paragraphe, puis saut vers un autre paragraphe en suivant une notion, puis saut au paragraphe suivant, etc ...

Les lectochromes peuvent être construits en prenant comme unité de base le « clic », mais la durée de la lecture est aussi une information pertinente, pour autant que nous soyons capables de la détecter puis de l'enregistrer.

### 3. La couleur pour donner du sens

Plusieurs cartes géographiques représentant la même région ou pays peuvent avoir des interprétations très différentes : routières, administratives, économiques, reliefs, ... Outre quelques différences graphiques, la différence la plus notable est l'utilisation des couleurs, pour représenter sur un même territoire des informations différentes. La carte plane n'ayant que deux dimensions, des couleurs vont restituer la troisième sur une carte en relief. Par extension, nous pouvons dire que les couleurs ajoutent une dimension administrative, ou économique, ou autre à la carte plane du territoire.

Si nous transposons l'utilisation de la couleur en cartographie géographique à notre problème de cartographie des parcours, cela semble vouloir dire que nous pouvons ajouter une dimension temporelle grâce à la couleur. Pour permettre une lecture plus efficace, en quelque sorte intuitive, la graduation colorée utilisée est proche de celle de l'arc-en-ciel, dans un modèle qui est censé aller empiriquement du froid vers le chaud. La finalité de

ce dispositif est de pouvoir comparer les parcours en valeur relative sans avoir à faire une analyse approfondie. Quand le lecteur désire comparer deux parcours, que ce soit des paragraphes, des ouvrages ou quoi que nous décidions d'implémenter par la suite, il devrait pouvoir intuitivement trouver lequel est le plus « chaud », donc le plus lu.

L'échelle est donnée sur le côté parce qu'elle est recalculée pour chaque lectochrome, de manière à ce que l'élément le moins utilisé soit le plus froid et l'élément le plus utilisé soit le plus chaud. Ceci permet de donner chaque fois une vision adaptée aux valeurs de « température » en présence. Mais comme ces valeurs apparaissent aussi en valeur absolue, il est toujours possible de les comparer ou de les réutiliser en dehors de l'écran où elles apparaissent.

Pour récapituler, sur un lectochrome, le lecteur dispose d'une échelle colorée graduée qui lui permet d'attribuer visuellement une « chaleur » à chaque lecture, c'est-à-dire une échelle qui lui permet de faire correspondre pour chaque couleur un nombre de lectures. Un exemple de représentation de cette échelle est proposé figure 7.15. Cette échelle affiche des valeurs relatives, c'est-à-dire qu'à la même valeur peut être associée une couleur différente suivant les valeurs qui doivent être représentées. Ceci permet de comparer sur un même lectochrome les valeurs entre elles, mais pas d'un lectochrome à l'autre.

Cette façon de représenter une valeur et son importance par une couleur peut-être facilement critiquée, nous n'avons effectivement pas tous la même représentation et cet « échauffement » peut ne pas être du tout visible pour certains lecteurs. Comme le montrent (Maturana et Varela, 1994) l'interprétation des couleurs est finalement assez aléatoire, même en dehors de pathologies qui affectent la vision. Souvent, nous ne voyons pas que nous ne voyons pas certaines choses et s'il a pu être démontré une corrélation de noms

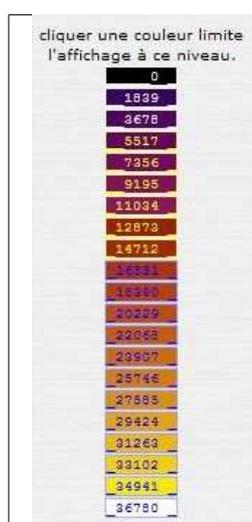


FIG. 7.15: Exemple d'échelle de couleur pour la lectochromie

de couleurs avec états neuronaux, en revanche il semble ne pas y en avoir avec les longueurs d'ondes correspondantes. Nous ne voyons pas les couleurs du monde, nous construisons notre propre espace chromatique.

### C. Mise en œuvre

L'appel de l'outil « lectochromie » de *CoLiSciences* se fait de manière implicite ainsi que de manière explicite lors d'une consultation. En effet, toutes les lectures engendrent des parcours qui sont enregistrés pour pouvoir être analysés et affichés sous forme de lectochromes. Le simple fait d'entrer dans le corpus active donc la lectochromie, d'ailleurs nous pouvons voir sur la figure 7.16 qui présente l'affichage d'un paragraphe en mode texte que la lectochromie est présente de ces deux manières. Dans la bulle marquée « A » se trouve un carré de couleur qui indique le nombre de fois que ce paragraphe a été affiché, ceci en donnant la valeur et en présentant le carré suivant l'échelle colorimétrique de l'ouvrage. Il nous semble difficile d'afficher cette échelle sur toutes les pages affichant un paragraphe sans alourdir la présentation mais en contrepartie, il est assez difficile d'interpréter la signification de ce carré tant

que l'internaute n'a pas activé le lien qui le conduira vers la lectochromie, lui donnant les explications nécessaires pour l'interprétation de ces échelles colorées. Nous voyons d'ailleurs ce lien toujours sur la figure 7.16, il est représenté sous la forme d'un onglet, qui apparaît dans la bulle marquée « B » et qui est visible sur n'importe quelle page affichant un paragraphe du corpus, ceci en mode de navigation fac-similé, texte ou notion.

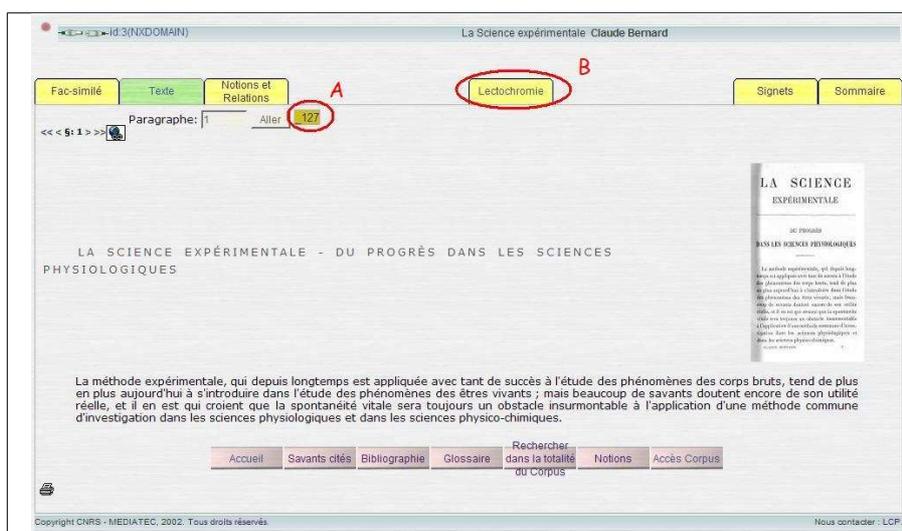


FIG. 7.16: Un paragraphe du corpus *C<sup>o</sup>L<sup>i</sup>s<sup>c</sup>iences*

Nous allons maintenant entrer dans le dispositif lectochromie pour voir plus en détail ce qu'il permet d'afficher et comment il peut aider l'internaute à appréhender le corpus. L'activation du lien lectochromie de l'onglet que nous venons de voir transporte l'internaute sur le menu principal de la lectochromie tel qu'il est visible sur la figure 7.17. Les choix disponibles vont par paires, parce qu'en fait pour une même sélection à l'intérieur des parcours enregistrés sur le corpus il est possible d'afficher soit les fréquences d'utilisation, c'est-à-dire le nombre de fois que l'entité a été affichée, soit les durées d'utilisation, c'est-à-dire le temps pendant lequel l'entité est restée affichée. Bien entendu, il n'existe pas de choix parfait pour représenter les parcours. D'une part avec les fréquences d'utilisation nous donnons moins

d'importance à un long paragraphe qui est lu longuement, mais qui est à la fin ou sur le chemin d'une suite cohérente de paragraphes à lire pour l'étude d'une notion, par rapport à un autre très court paragraphe qui serait au début d'un chemin de lecture et de ce fait affiché systématiquement, sans doute plus souvent que le premier. Mais ce court paragraphe, s'il existe, a certainement un rôle à jouer. Peut-être articule-t-il une idée déjà développée avec une nouvelle qui va être introduite ? Même s'il est court et facile à lire, il apporte alors sans doute un sens à la lecture et sa fréquence d'utilisation reflétera la réalité de son importance, ce qui n'apparaîtrait pas de manière aussi claire avec la durée de lecture.

D'autre part avec les durées de lecture, nous avons une difficulté incontournable qui consiste à savoir ce que fait l'internaute derrière son écran. Il n'est pas possible de faire un système tel celui qui s'assure que les conducteurs de locomotive sont toujours éveillés : « cliquez ce bouton toutes les 10 secondes pendant que vous lisez un paragraphe ». Ce qui est enregistré est donc tout simplement la durée pendant laquelle un paragraphe reste affiché à l'écran, mais cela ne signifie pas qu'il soit lu. Pour éviter d'enregistrer des valeurs aberrantes, le dispositif n'enregistre pas les durées trop longues. Que signifie « trop » dans ce contexte ? C'est un réglage, une supposition, basée empiriquement sur une valeur moyenne du temps qu'il faut pour lire un paragraphe, majorée de manière à obtenir un temps « maximum » durant lequel un internaute est susceptible de lire un paragraphe. Au delà de ce délai, le temps d'affichage n'est plus enregistré.

La figure 7.17 nous permet de voir ce qui est affiché sur le menu principal du dispositif de lectochromie. Il comporte cinq entrées principales, permettant à l'internaute d'afficher les parcours de lecture pour :

- lui-même dans l'ouvrage en cours.

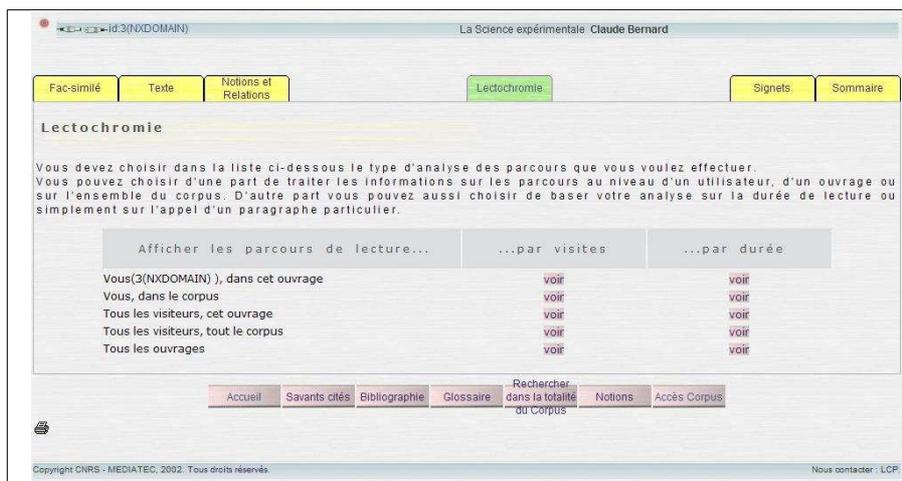


FIG. 7.17: Le menu d'accueil de la lectochromie

- lui-même dans tout le corpus.
- tous les utilisateurs de  $C_o^L i^S c i e n c e s$  dans l'ouvrage en cours.
- tous les utilisateurs dans le corpus.
- tous les ouvrages de  $C_o^L i^S c i e n c e s$ .

Ces parcours pouvant être affichés en prenant pour base de l'analyse des interactions entre l'internaute et  $C_o^L i^S c i e n c e s$ , soit l'affichage d'un paragraphe (première colonne), soit la durée de lecture (deuxième colonne).

Il est ainsi possible pour chaque internaute d'analyser visuellement et rapidement l'utilisation qu'il fait d'un ouvrage grâce à un écran du type de celui qui est affiché sur la figure 7.18 qui lui donne son utilisation personnelle. Sur cette figure il apparaît très facilement que ce sont les deux premiers paragraphes qui sont les plus lus, nous pouvons ensuite noter que le dernier et le paragraphe 127 ont été relativement plus accédés que les autres. La dernière remarque qui peut être faite est que tous les autres paragraphes lus l'ont été de manière isolée. Nous sommes là complètement dans une lecture hypertextuelle et non séquentielle, mis à part deux exceptions : les paragraphes 76 et 77 d'une part et les paragraphes 126, 127 et 128 d'autre part.



FIG. 7.18: Les parcours de l'internaute dans l'ouvrage en cours

Le lecteur pourra ensuite utiliser cette vue pour la comparer avec celle qui présente les lectures faites par l'ensemble des personnes qui viennent consulter les ouvrages de *C<sup>o</sup>LiSciences* et que nous pouvons trouver sur la figure 7.19.

Sans aller très loin dans l'analyse, quelques informations peuvent être facilement extraites de la comparaison des parcours d'un lecteur en particulier par rapport à la moyenne de l'ensemble des lecteurs de cet ouvrage. Dans les deux cas les premiers paragraphes sont visiblement très souvent activés, plus souvent que n'importe quel autre paragraphe, ce qui voudrait dire que les lecteurs commencent leur cheminement conventionnellement à partir du début de l'ouvrage plutôt qu'en utilisant les signets qui les feraient arriver directement sur une figure caractéristique par exemple. Cette fois nous voyons qu'en moyenne les lecteurs poursuivent une lecture séquentielle jusqu'au paragraphe 6, voire 7 pour un nombre plus restreint, alors que notre lecteur

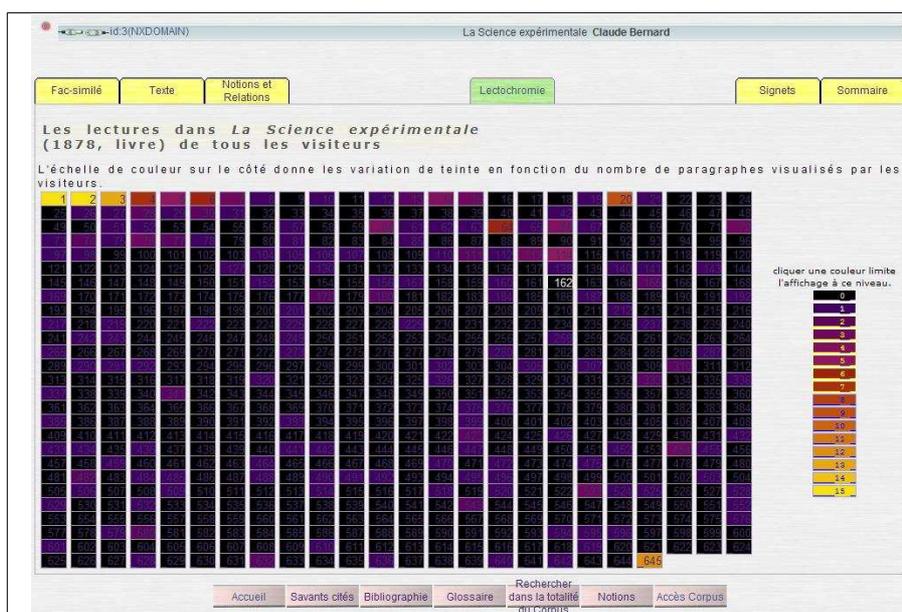


FIG. 7.19: les parcours de tous dans l’ouvrage en cours

ne dépassait pas le 2. Plus précisément, quand nous prenons les parcours de tous les visiteurs, ils vont tous globalement jusqu’au troisième et un grand nombre « pousse » même les explorations jusqu’au sixième, un petit nombre va jusqu’au septième.

Le paragraphe 127 n’apparaît pas comme quelque chose d’important aux yeux des lecteurs et c’est une différence intéressante facilement identifiable au niveau de ce paragraphe, qui a été très souvent affiché par l’internaute, alors que cela n’apparaît pas sur la figure reprenant les parcours de tous. Comme les parcours de cet internaute sont inclus dans les parcours de tous, cela confirme d’autant plus que ce paragraphe 127 n’a pas eu globalement d’attrait pour l’ensemble des visiteurs, même pas pour un petit groupe d’entre eux.

Le même type de vue qui permet à la fois d’analyser une situation et de permettre une comparaison pourra être fait sur l’ensemble du corpus, ce qui permet de donner un niveau d’étude plus élevé que celui d’un simple ouvrage et permet d’évaluer l’usage qui est fait du corpus dans sa globalité. Il sera ainsi possible de voir si tous les ouvrages sont utilisés de façon égale ou si certains

sont plus souvent lus. Il est aussi possible de voir si un ou plusieurs ouvrages sont complètement délaissés. Ce ne sont donc plus véritablement des parcours mais des lectures qui sont affichées, en fait si un ouvrage montre être souvent lu, cela voudra dire qu'il est possible de retrouver un plus grand nombre de parcours à l'intérieur de celui-ci.

Sur les écrans que nous allons voir (figure 7.20 et figure 7.21), ce ne sont plus des numéros de paragraphes qui sont affichés, mais les titres des ouvrages. En revanche c'est bien toujours pour représenter le nombre de paragraphes (ou la durée de lecture de ceux-ci) que le nom de l'ouvrage est utilisé, c'est donc en quelque sorte la même entité qui est représentée, mais avec un niveau d'échelle différent. Cette notion de mise à l'échelle et de « zoom » sur les parcours est mal mise en valeur et n'apparaît explicitement, le dispositif n'utilise pas encore les signes adéquats pour mettre le lecteur en position d'avoir cette interprétation, mais c'est une piste qu'il faudra développer. Pour l'instant cette possibilité est mise en pratique uniquement à travers les liens hypertextuels. Les noms des ouvrages sont actifs, c'est-à-dire qu'un clic sur un de ces noms permet de revenir à des affichages équivalents à ceux présentés figures 7.18 et 7.19, mais pour l'ouvrage dont le nom aurait été sélectionné plutôt que pour l'ouvrage en cours.

La figure 7.20 représente un exemple de lectochrome qu'il est possible d'obtenir pour illustrer quels sont les ouvrages les plus lus par une personne. La liste comporte tous les ouvrages du corpus et pas uniquement ceux qui ont été ouverts au moins une fois par cet internaute, à l'avenir il pourra être nécessaire d'ajouter des filtres permettant de ne voir que les ouvrages ouverts plus d'un certain nombre de fois.

De la même façon que précédemment, il est aussi possible de voir sur un autre lectochrome comme celui présenté sur la figure 7.21 les lectures de tous

les internautes à travers tout le corpus, ce qui permet d'identifier les ouvrages les plus ouverts et les plus parcourus.

La même type de comparaison que celle que nous avons conduit pour un ouvrage peut être effectuée pour l'affichage des lectures ou parcours du corpus dans sa globalité, d'un internaute par rapport à l'ensemble des internautes. Nous pouvons voir sur la figure 7.20 que l'ouvrage le plus souvent ouvert par cet internaute est « *Introduction à l'étude des races humaines* », c'est même le seul qui ait été réellement utilisé, d'après la couleur affectée aux autres ouvrages ils ont été ouverts tout juste une fois, quand ils l'ont été. En comparaison, sur la figure 7.21 nous pouvons constater que cette fois plusieurs ouvrages sont utilisés, deux beaucoup plus que les autres, un certain nombre ayant aussi été ouverts sans pour autant qu'ils semblent avoir réellement été étudiés. C'est cette fois « *Introduction à l'étude de la médecine expérimentale* » qui est le plus ouvert, « *Introduction à l'étude des races humaines* » arrivant en deuxième position.

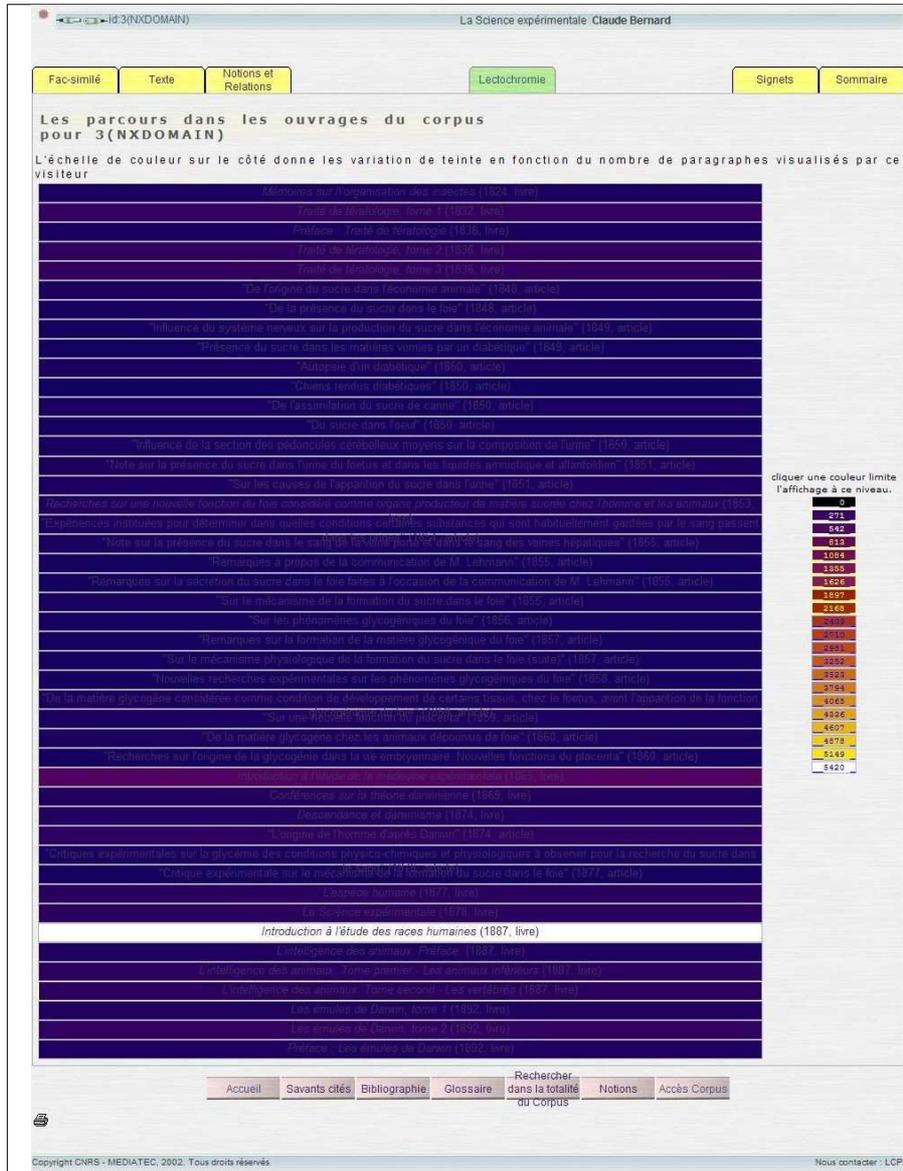


FIG. 7.20: Les parcours de l'internaute dans le corpus

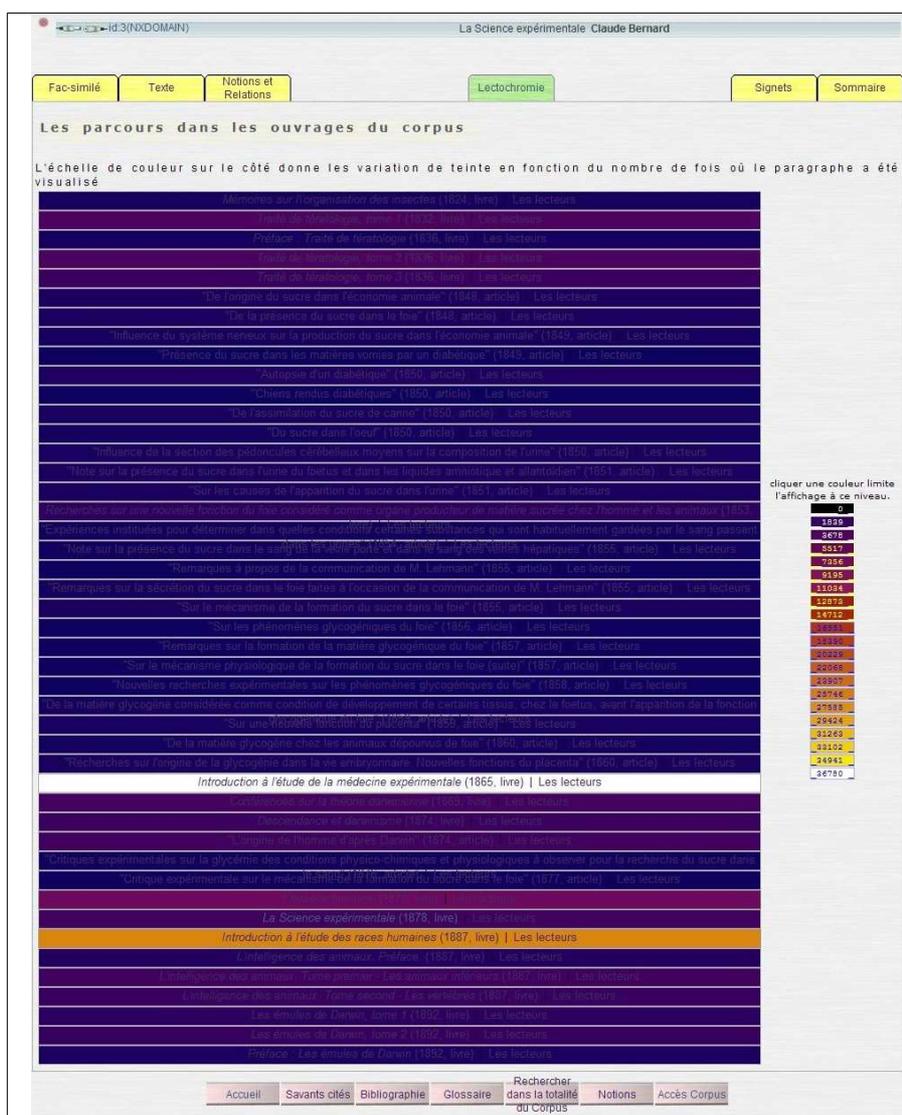


FIG. 7.21: Les parcours de tous dans le corpus





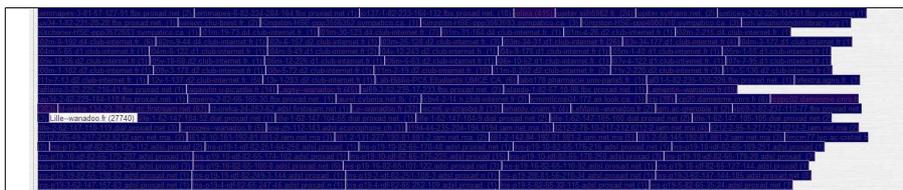


FIG. 7.25: les parcours de tous dans le corpus (4/5)

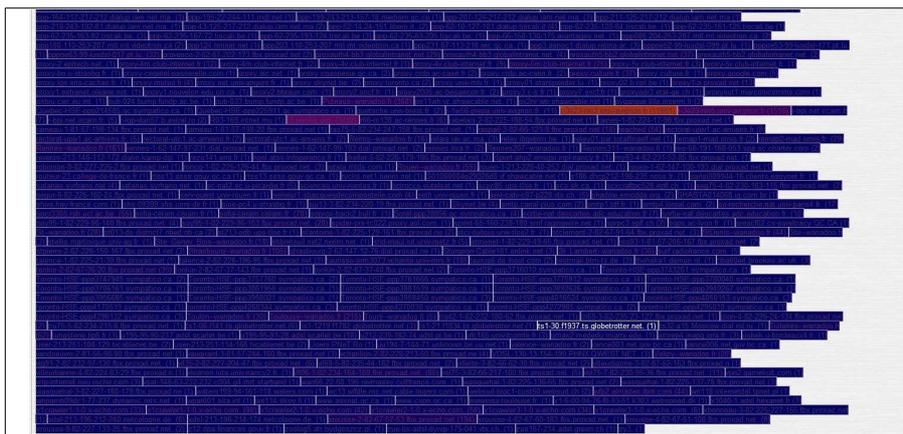


FIG. 7.26: les parcours de tous dans le corpus (5/5)

## D. Modélisation

Un essai de représentation systémique de la lectochromie demande de prendre en considération comme unité élémentaire une opération de lecture sur un paragraphe du corpus. La combinaison de ces unités élémentaires dans les parcours à travers le corpus donnera lieu à l'élaboration d'une carte conceptuelle qui regroupe par inférences ascendantes les représentations élémentaires que sont les opérations de lecture sur un paragraphe. Ainsi, les parcours à travers le corpus sont définis par la conjonction des sous-systèmes de représentation, ce qui est en rupture avec la logique analytique d'Aristote qui repose sur trois axiomes :

1. L'axiome d'identité : A est A.
2. L'axiome de non-contradiction : A n'est pas non-A.
3. L'axiome du tiers exclu : il n'existe pas un troisième terme T qui est à la fois A et non-A.

Ce type de raisonnement supposerait qu'à un signe, à un mot, ne puisse être attribué qu'une seule réalité, et que cette représentation est vraie ou fausse en toutes circonstances. Mais dans la lectochromie, un paragraphe peut prendre des interprétations différentes :

- Il peut intervenir dans plusieurs parcours différents.
- Il peut représenter plusieurs notions, liées ou pas,
- Il existe sous trois instanciations différentes : plein texte, texte balisé notionnellement et fac-similé.

Ce qui va le faire apparaître comme un nœud de passage important du réseau hypertextuel reposera sur l'abstraction qui sera faite des motivations qui étaient à l'origine de sa lecture.

La logique conjonctive permet de rendre compte des phénomènes que nous percevons dans et par leurs conjonctions complexes. Elle introduit un caractère relatif et contingent des représentations : nous ne cherchons pas une vérité absolue pré-existante, nous construisons ensemble notre représentation de notre expérience des phénomènes au travers d'une systémographie (l'isomorphe) qui garantit la fiabilité de cette représentation.

*« Le système est récursif, s'établissant dans l'interaction entre le système modélisé (le phénomène perçu complexe) et le système modélisant... La modélisation systémique postule que l'action de modéliser n'est pas neutre et que la représentation du phénomène n'est pas disjoignable de l'action du modélisateur. L'idéal de la modélisation ne sera plus dès lors l'objectivité du modèle, comme en modélisation analytique, mais la projectivité du système de modélisation : on caractérise la projectivité par la capacité du modélisateur à expliciter ses « projets de modélisation », c'est-à-dire les finalités qu'il propose au modèle(...) le système de modélisation (...) se comprend comme auto-finalisant : il élabore ses projets, il est projectif. (Le Moigne, 1990, page 65) »*

L'étude du comportement des lecteurs, qui sont d'une part les participants du système modélisant doit ainsi tenir compte du fait qu'il sont aussi d'autre part des acteurs directs du processus de conception de ce système. La construction des lectochromes devra tenir compte du comportement des lecteurs, l'intégrer et évoluer avec les analyses qu'elle permet de faire. Nous verrons plus loin dans la section concernant l'analyse que l'arrivée d'un lecteur inattendu nous conduira à modifier le comportement du dispositif de production des lectochromes, devenu inutilisable depuis son arrivée. Les lecteurs sont des acteurs directs dans le processus continu de création des lectochromes. Le dispositif ne permet pas de découvrir une réalité cachée, qui

préexisterait au lectochrome construit par les lecteurs, il permet seulement de mettre en commun leurs choix de mots, de parcours et d'agencement des idées. À un instant donné, le lectochrome est un résumé visuel de ces choix, ce n'est pas le corpus, mais simplement la carte des traces des voyages à travers celui-ci.

## **E. Analyse**

### **1. Construction de parcours**

L'analyse des diagrammes fournis par la lectochromie doit nous permettre de visualiser globalement l'usage qui est fait du corpus, et quand nous utilisons les outils d'analyse plus fins, de mettre en évidence quels chemins sont suivis par les lecteurs pour arriver sur un paragraphe en particulier, ou encore quels chemins sont utilisés par les lecteurs intéressés par une notion précise.

Ce qui est intéressant à essayer de faire émerger, c'est ce qui donne du corps au réseau hypertextuel. Les paragraphes sont liés entre eux, de manière plus ou moins forte et nous avons vu que ce lien pouvait être renforcé par la lecture qui en est faite. Il serait de ce fait important de pouvoir comprendre comment d'un paragraphe le lecteur peut-il choisir de continuer vers un autre et lequel, cela donnerait une information nouvelle et très pertinente pour alimenter le moteur hypertextuel.

Ces remarques expriment en fait le besoin d'une nouvelle mise en situation du texte par rapport à ses lecteurs au travers de l'hypertexte en général et de la lectochromie en particulier. L'analyse des lectures que permet la lectochromie va nous permettre de faire un premier pas dans cette direction. Nous sommes en fait engagé dans la finalisation d'un dispositif qui s'auto-alimente, comme

cela se fait dans la mise au point d'un prototype entre réflexion théorique et essais pratiques, pour nous les deux parties du dispositif qui sont à l'origine de ce mouvement de balancier entre deux situations instables sont les réflexions sur les résultats de la lectochromie d'une part et la finalisation de l'agent hypertextuel, la stabilisation serait obtenue par l'aboutissement à un agent hypertextuel qui serait capable de guider un lecteur à travers n'importe quel texte.

## 2. Les lecteurs

**Qu'est-ce qu'un lecteur ?** Pour définir ce qu'est un lecteur, nous n'entrerons pas dans les détails de la théorie de la communication qui nous amènerait à présenter les études sur la réception et la définition d'un lecteur-type. Nous nous contenterons d'une définition plus triviale, qui consiste à présenter le lecteur d'un ouvrage papier comme un être humain qui prend connaissance du contenu d'un ouvrage. En effet, le fait que ce lecteur soit humain est sous-entendu et nous a amené à faire des hypothèses erronées, qui nous ont conduit à modifier notre dispositif pour pouvoir l'utiliser.

### Qui sont les lecteurs ?

**Les lecteurs attendus** Bien évidemment, chacun peut retrouver ses propres parcours. Nous pouvons ainsi retrouver tous les parcours des partenaires significatifs du projet.

**Le lecteur inattendu** Dans l'élaboration d'un système tel que *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, la question de l'IHM, *Interface Homme Machine* revient de manière assez récurrente. Pour chaque fonction ou outil que nous avons ajouté, nous avons toujours essayé de faire des choix ergonomiques simples, de manière à faciliter

le butinage, à maintenir une cohérence dans les parcours ou les lectures que les utilisateurs de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* seraient amenés à faire. Comme nous l'avons dit, la lectochromie fait partie de ces outils qui aident le lecteur à se réorienter.

La surprise a donc été grande, et le paradoxe intéressant, de constater dans les statistiques d'utilisation et la création de parcours que le lecteur le plus assidu de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* n'était autre qu'une machine. C'est le robot d'indexage du moteur de recherche Google qui arrive effectivement en tête des lecteurs. Infatigable, inexorable, il a patiemment suivi tous les liens, il est revenu régulièrement, refaire les mêmes parcours et en démarrer d'autres.

### 3. Les possibilités d'évolution du système

L'évolution de ce dispositif pourrait se faire suivant deux directions :

- Rendre l'interface plus lisible, de manière à ce que le lecteur retrouve cette impression de survol du corpus, puis de zoom qui le rapproche d'abord des ouvrages, ensuite du texte et de ses notions, pour enfin le mener pas à pas à travers un parcours déjà effectué.
- Pour ce qui est de ces parcours, la finalisation technique de cette partie du dispositif doit permettre de les modifier, de les nommer et de pouvoir les regrouper par thématique.
- Faciliter les navigations « transcorpus », c'est-à-dire d'un ouvrage à l'autre, et rendre encore une fois lisibles les parcours de ce type.
- Proposer une carte de France ou du Monde pour situer les lecteurs et contribuer à la création d'une communauté de pratique autour de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*.

## Troisième partie

# Genèse du dispositif *CoLiSciences*

Retour d'expérience et mise en œuvre



**LES LOGICIELS LIBRES**

---

*« Il y a assez de lumière pour ceux qui ne désirent que de voir, et assez d'obscurité  
pour ceux qui ont une disposition contraire »*

Pascal / *Pensées* (430) .

*« Le désir de puissance consacre la machine comme moyen de suprématie, et fait  
d'elle le philtre moderne. L'homme qui veut dominer ses semblables suscite la  
machine androïde. Il abdique alors devant elle et lui délègue son humanité. »*

Gilbert Simondon / *Du mode d'existence des objets techniques* (page 10) .

## Sommaire

---

<b>8.1</b>	<b>Introduction</b>	<b>369</b>
A.	Définitions	376
B.	Les licences libres	386
C.	Les textes fondateurs	392
D.	La rencontre	396
<b>8.2</b>	<b>Nouveaux cadres pratiques, nouveaux modèles théoriques</b>	<b>399</b>
A.	Quel statut?	399
B.	Programmeur de Logiciel Libre?	401
C.	Logiciel libre et innovation	405
<b>8.3</b>	<b>Logiciels Libres et recherche scientifique</b>	<b>409</b>
A.	Introduction	409
B.	Une impasse scientifique	410
C.	Le domaine public en danger	411
<b>8.4</b>	<b>Pourquoi des Logiciels Libres dans un projet de numérisation</b>	<b>419</b>
A.	Nos objectifs et contraintes	420
B.	Composants logiciels	426
C.	Méthode de développement	430

---

## 8.1 Introduction

Avant d'aborder un chapitre qui semble entièrement dédié à la technologie, il nous paraît nécessaire de faire un certain nombre de mises au point, en particulier sur la place de la technologie dans un projet de recherche. Tout d'abord il est important de noter que le principal enjeu des Logiciels Libres n'est pas technique mais philosophique, nous y reviendrons bien entendu plus largement au cours du chapitre. Il nous faudrait aussi sans doute expliquer la pertinence d'une intrusion technologique dans notre travail de recherche. Pour cela il est intéressant de noter comme l'a fait Pierre Lévy :

*« Les dirigeants des multinationales, les gestionnaires avertis et les ingénieurs créatifs savent parfaitement que les stratégies victorieuses passent par les moindres détails « techniques », dont aucun n'est méprisable, et qui sont tous indissolublement politiques et culturels en même temps que techniques (Lévy, 1990, page 10) »*

Pour bien préciser la teneur de ce lien, nous allons tout d'abord commencer par situer le contexte de notre recherche. La technologie, le numérique dans le cas présent, n'est bien entendu pas une fin en soi. Mais comme c'est le cas pour toutes les technologies, le numérique est perçu comme une boîte noire, un outil que l'on utilise ou pas et de ce fait souvent relégué au plan de simple accessoire. Pourtant, quand Gutenberg a inventé l'imprimerie, il n'a pas simplement inventé un outil technique, une machine mécanique apte à reproduire des feuillets écrits avec plus de facilité. Il a en fait révolutionné la diffusion du savoir et des idées.

Pour illustrer cela, nous pouvons prendre pour exemple deux tentatives de faire émerger une nouvelle interprétation des textes sacrés, une avant et

une après Gutenberg<sup>127</sup>. Wycliffe croyait que chaque individu avait le droit d'avoir une expérience avec Dieu à travers les témoignages bibliques, en devenant personnellement familier avec les écritures. Le moyen d'acquérir cette connaissance résidait dans la possibilité d'avoir la Parole de Dieu dans une langue que le peuple pouvait réellement comprendre et c'est pour cette raison qu'il entrepris de traduire la Bible en anglais, vers 1380-1381. Rendre la Bible disponible à la masse, dans un anglais courant, était considéré comme une hérésie. Cela menaçait grandement l'autorité de l'église et elle considérait donc Wycliffe comme dangereux. Le mouvement Lollard qu'il avait fondé fut facilement démantelé et il fallu plus d'un siècle avant que Martin Luther n'entreprenne la même démarche.

Lorsque Tyndale traduisit le nouveau testament en 1524-1525, grâce à l'imprimerie il put imprimer plus de 50000 copies avant d'être rattrapé et brûlé vif. Le siècle des Lumières et par là l'effondrement de l'ancien régime en Europe doivent sans doute beaucoup plus qu'il ne semble à première vue aux mécaniciens qui ont construit des machines à imprimer.

Aujourd'hui les enjeux sont les mêmes, même si les acteurs sont bien entendu différents et ce n'est plus la circulation des idées religieuses qui est généralement en cause. La lecture, les images, musiques et vidéos, médiatisées numériquement par l'ordinateur et Internet circulent plus facilement que sous leurs formats précédents. Surtout, ils échappent pour l'instant à tout contrôle, ce qui remet en question une économie qui s'est construite sur le contrôle de la diffusion et la rétribution des auteurs proportionnellement à l'utilisation qui est faite de leurs œuvres. Plutôt que de remettre en cause ce modèle économique et de chercher à en penser un nouveau qui s'appuierait sur les nouveaux outils, pour l'instant la réponse a été principalement judiciaire, ou

---

127. La Bible de Gutenberg a été numérisée et mise en ligne par la British Library. Elle est visible à l'adresse : <http://www.bl.uk/treasures/gutenberg/homepage.html>

purement technique. Réponse judiciaire d'une part avec des actions en justice à l'encontre de quelques personnes qui échangent illégalement de trop nombreux fichiers soumis à copyright. Ces procès, par leur médiatisation, envoient un message au grand public pour leur rappeler que ce qu'ils font est illégal et que nul n'est censé ignorer la loi. Même lorsqu'il est si facile de l'enfreindre, elle doit être scrupuleusement respectée. Réponse technique d'autre part, avec la mise en place d'un certain nombre de verrous technologiques<sup>128</sup> qui sont régulièrement venus renforcer ou plutôt rétablir le contrôle qui ne peut plus être fait sur la duplication des œuvres. Mais ces verrous sont souvent trop facilement contournés par les personnes qui sont décidées à enfreindre la loi. Si un contre-verrou ou « passe partout » électronique est créé, encore une fois Internet et la facilité de duplication permettent de le diffuser à un grand nombre de personnes et de rendre le verrou obsolète. Ainsi, les individus qui ont fait le choix de l'illégalité ont toujours accès à ces contenus protégés, alors que dans le même temps, les utilisateurs normaux se trouvent face à de nouveaux problèmes de restitution de l'œuvre dont ils ont acquis légalement les droits de diffusion mais pour laquelle ils utilisent un matériel trop ancien ou qui n'est pas exactement conforme (l'exemple le plus typique étant le CD de musique qui refuse d'être lu par un auto-radio); de plus ce type de verrou interdit aux usagers de faire état de leur droit à la copie privée.

Il reste une troisième voie, qui mettrait enfin tout le monde d'accord : artistes, diffuseurs et public, c'est celle qui consisterait à trouver un nouveau modèle économique qui prenne en compte la nouvelle réalité et les nouvelles possibilités de diffusion. Si les intermédiaires de diffusion actuels de contenus

---

128. Il s'agit par exemple de l'actuelle gestion des droits assurée par la technologie DRM\*, l'annexe K.1 présente la complexité face à laquelle se trouve un lecteur potentiel du fait de l'utilisation de ce type de verrou. Une nouvelle technologie du même type, mais encore plus verrouillée et opaque pour les utilisateurs est représentée par le processeur Palladium et TCPA\*. Le plus grave est que ce type de blocage peut très bien se produire avec des ouvrages du domaine public. C'est d'ailleurs déjà le cas comme cela est montré dans l'annexe L.1.

culturels ne sont plus adaptés, ils devraient de toute manière finir par disparaître, comme cela a déjà été le cas lors de précédentes mutations. Parce que ce n'est bien entendu pas la première, en particulier dans le domaine musical.

- Quand sont apparus les tous premiers CD (compact disques), ce nouveau produit a tout de suite fasciné le grand public. Le CD apportait une qualité de restitution inconnue jusqu'alors, accompagnée d'une robustesse et d'une compacité qui semblaient résoudre les problèmes majeurs des disques vinyles. Il était devenu possible d'écouter ses morceaux favoris des centaines de fois sans que ce support ne se détériore, comme lors de la simple écoute d'un fragile disque vinyle et il était devenu possible de les transporter sans (trop de) risques. Des baladeurs et des auto-radios allaient rapidement être équipés pour pouvoir écouter plus d'une heure de musique de qualité tout en voyageant, rendant ainsi au passage les cassettes audio obsolètes elles-aussi, puisque c'était le format adopté pour la musique « mobile » du fait de sa robustesse. Le CD représentait donc un vrai bon en avant pour le consommateur final et les galettes vinyles ont été mises au rebut. Le basculement a été très rapide, en moins de dix ans tout au plus les fabricants de disques ont renouvelé tous leurs outils de presse. Toutefois, ni les maisons de disques, ni les distributeurs (qui ont simplement dû faire quelques ajustements physiques dans leurs magasins) n'ont réellement eu à souffrir de cette rupture technologique puisque parallèlement le public a renouvelé et racheté tous ses titres préférés en CD. D'ailleurs, la dépense ne s'est pas arrêtée là puisqu'il a fallu aussi investir dans de nouveaux lecteurs. Les fabricants d'équipement ont vu parallèlement s'ouvrir de nouveaux marchés (notamment avec les lecteurs de voiture et les baladeurs).

- Mais ce n’était pas la première fois qu’une telle rupture se produisait. A peine plus d’un siècle auparavant, nous trouvons un exemple analogue avec l’industrie de la musique mécanique, i.e. les limonaires et autres orgues de barbarie. C’est une industrie qui s’est développée et a connu une assez grande diffusion sur une période d’environ deux siècles, que nous avons maintenant oublié si ce n’est le sympathique joueur d’orgue de barbarie avec son singe sur l’épaule. A côté de cela il y avait des machines dont les fonctions sont très proches de nos lecteurs de CD et autres juke-boxes. Parce qu’effectivement, des juke-boxes mécaniques avec monayeurs existaient déjà, parce qu’il existait des machines à même de remplacer un orchestre pour animer un bar, de remplacer un pianiste pour sonoriser un film. Tout cela a disparu avec l’avènement du phonographe<sup>129</sup>, remplacé par l’électrophone et les vinyles, détrônés par le CD qui est à son tour largement menacé aujourd’hui par le téléchargement.

Les ateliers qui préparaient la musique sur feuilles de carton, ou sur disques métalliques ont soit disparus, soit se sont reconvertis dans la fabrication des disques de cire d’abord, puis de vinyle. Ils n’avaient pas le choix : s’adapter ou disparaître, comme cela est normalement le cas avec l’arrivée d’une nouvelle technologie qui rend la précédente obsolète. N’ayant pas peur des mots, la musique électronique rend obsolète le CD, il suffit de comparer les possibilités d’un baladeur CD « traditionnel » à celui d’un iPod<sup>130</sup> pour se rendre à l’évidence que n’importe quel consommateur, s’il en a les moyens, donnera la préférence à l’iPod. Maintenant que le dispositif

---

129. L’inventeur américain Thomas Edison a déposé le brevet du premier phonographe le 19 décembre 1877.

130. Le iPod est un baladeur fabriqué par la société Apple. Suivant le modèle il permet de stocker de plusieurs centaines à plusieurs milliers d’heures de musique, qui peuvent être téléchargées depuis le site associé iTunes <http://www.apple.com/fr/itunes/>.

émergent commence à être relativement bien identifié, il faudrait que nous sachions comment vont devoir évoluer les maisons de disques, par qui risquent-elles d'être remplacées et qui aura la charge de rétribuer les auteurs dans cette nouvelle organisation. Comme nous venons de le mentionner, une illustration de ce type de réponse est le lecteur de musique iPod, associé à un site Web marchand où il est possible d'acheter en la téléchargeant de la musique à un tarif sans commune mesure avec celui pratiqué par la vente de CD. Ce nouveau phénomène est complexe, difficile à traiter si les droits de toutes les parties en présence sont pris en compte et nous nous trouvons finalement dans une situation de relative attente, sans savoir si la réponse sera judiciaire, technique ou commerciale. Pour cette raison, ce phénomène inquiète les auteurs, mais avant eux il inquiète tous les intermédiaires dont l'économie est fondée sur la diffusion de ces contenus. Ils n'ont pas encore trouvé le modèle qui leur permette de continuer à travailler dans de bonnes conditions. Ils subissent de plein fouet un phénomène de « désintermédiation », les nouveaux outils de communication ayant supprimé leur étape dans la chaîne de distribution. Ils ne savent pas pour l'instant développer un nouveau modèle, créer une nouvelle économie qui leur permette de s'inscrire à nouveau dans cette chaîne. Plutôt que de s'adapter aux nouveaux médias, ces grandes sociétés, généralement multinationales, essaient de verrouiller les nouveaux médias pour pouvoir continuer à contrôler les flux d'information et la diffusion de contenus culturels.

Notre recherche est en relation directe avec un objet technique. Pour qu'elle soit complète, il nous paraît indispensable de commencer par présenter et de faire un point sur les technologies ouvertes, en particulier sur la relation étroite qu'elles devraient entretenir avec tout travail de recherche, ceci de façon encore plus nette quand il s'agit d'un travail de recherche action tel

que le nôtre. Mais l'enjeu est aussi politique. Á la croisée de tous ces aspects, la frontière est difficile à ne pas franchir. Nous nous efforcerons de rester sur une analyse objective et scientifique de la relation entre la technologie, les Logiciels Libres et le travail du chercheur. Il nous semblait important de préciser comment recherche et techniques sont liées, pour pouvoir rendre ce lien le moins contraignant possible. A trop vouloir s'affranchir du discours technique, certains finissent par s'aliéner à une technologie, ceci à l'inverse de leur aspirations premières. Les Logiciels Libres nous ont permis de regarder la technique véritablement comme un outil, c'est-à-dire à notre disposition, sans que cela ne remette en question nos choix de recherche.

Pour donner la teneur d'ensemble et le contexte de ce chapitre, il faut commencer par préciser que le projet *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* s'appuie largement sur des logiciels issus de la communauté « Open Source », c'est-à-dire des logiciels dont le code source est disponible et modifiable, appelés aussi Logiciels Libres, ceci accompagné d'un certain nombre de critères que nous allons voir. Ces termes « Open source », « Logiciel Libre », sont de plus en plus utilisés dans toute la presse qui touche à l'informatique et même souvent au-delà, sans pour autant être définis correctement. Il sont souvent interprétés comme synonymes même si leur sens est très proche et font partie d'un même courant de pensée dans le milieu de la création numérique, ceci pris au sens large puisque maintenant ce mode de diffusion est repris pour d'autres types de création comme la musique, la littérature ou la photographie pour lesquels on parlera de « licences de contenus libres ». Toutefois, leur sens est défini de manière très précise à l'intérieur d'un cadre légal, il nous faut donc commencer par préciser d'abord ce que ces définitions désignent avant de commenter les relations entre la communauté des Logiciels Libres et notre projet. Nous verrons que ce qui pourrait sembler n'être qu'un environnement

technologique est en fait un choix guidé d'abord par l'idée que nous nous faisons des logiciels, mais qui répond aussi à des soucis budgétaires et permet d'accélérer le processus de développement. Les premières licences de Logiciels Libres sont apparues à la fin des années quatre-vingt dans le milieu de la recherche. Ce qui n'a rien d'étonnant dans la mesure où ces licences ont pour objectif de formaliser une pratique répandue entre laboratoires : l'échange des données en toute transparence. Grâce aux Logiciels Libres plusieurs de nos problématiques ont trouvé leur solution. Ils nous ont aussi permis d'assouplir les contraintes habituelles de ce type de projet, ainsi que celles dues à l'utilisation d'outils technologiques quelquefois trop spécifiques.

## A. Définitions

### 1. Logiciel

Nous devons passer par une courte série de définitions pour être certain d'avoir bien clarifié le contexte technique des Logiciels Libres.

**Logiciel :** Un logiciel est un ensemble d'ordres écrits dans le langage de l'ordinateur sur lequel il doit s'exécuter pour traiter des données. Le logiciel, ou programme, est aussi quelquefois appelé exécutable : *qui peut être exécuté* par opposition au code source, que nous détaillons plus loin et qui lui ne le peut pas.

**Langage de programmation :** Un ordinateur est une machine électronique munie d'un processeur qui, pour simplifier, ne sait faire la différence qu'entre le fait que *le courant passe* et le fait que *le courant ne passe pas*. Le raccourci est un peu simpliste et nous nous en excusons auprès des spécialistes, mais il nous suffit pour expliquer que les ordinateurs ne savent seulement qu'interpréter un langage binaire, c'est-à-dire constitué de 0

et de 1, courant ou pas courant. Aux premiers temps des ordinateurs, les programmeurs devaient coder les programmes directement sous la forme de séquences d'ordres dans le langage de la machine (appelé aussi assembleur, code assembleur, langage d'assemblage sans doute à la suite d'une série d'anglicismes). Chaque processeur, ou famille de processeur, possède son propre langage, de plus ces ordres se situant au niveau du processeur sont très simples, voire minimalistes : déplacer une valeur d'une case mémoire à une autre, additionner deux cases mémoire, ... L'assembleur étant à la fois très complexe à manipuler et très dépendant du matériel pour lequel il est prévu, les ingénieurs ont développé des langages dits évolués, plus proches du langage humain, pour coder les logiciels.

**Code source :** Pour résumer, un logiciel se présente sous deux formes : le « binaire » ou « exécutable » représente le logiciel tel que l'ordinateur pourra l'interpréter ; le « code source » représente le logiciel tel qu'un être humain pourra l'interpréter (pourvu qu'il connaisse le langage de programmation dans lequel le logiciel a été écrit).

## 2. Logiciel Libre

L'expression *logiciel libre* correspond au terme anglais *free software*. Du fait de l'ambiguïté du terme *free*, qui en anglais peut signifier à la fois libre ou gratuit, il a existé, et il existe sans doute toujours, une certaine confusion dans la place que ce type de logiciel pouvait jouer dans notre société, en particulier du point de vue économique. De plus, à une époque encore récente, une méthode de distribution très populaire parmi certains programmeurs était le *shareware*. Cela consistait à distribuer librement le code exécutable d'un programme, en demandant aux utilisateurs de bien vouloir régler une somme,

souvent modeste, auprès du programmeur s'ils jugeaient le programme utile. A cela est venu s'ajouter le *freeware*, qui est donc un modèle où le code exécutable est distribué gratuitement. Avant d'aller plus loin commençons donc par noter que les Logiciels Libres n'ont absolument rien à voir avec les *freewares* ni avec les *sharewares*, et qu'en aucun cas un logiciel libre n'est un logiciel gratuit.

Pour présenter les Logiciels Libres le plus simple est de faire référence au site de la FSF, *Free Software Foundation* <sup>131</sup> où nous pouvons lire la définition qu'en fait son fondateur Richard Stallman :

*« L'expression « Logiciel libre » fait référence à la liberté et non pas au prix. Pour comprendre le concept, vous devez penser à la « liberté d'expression », pas à « l'entrée libre ».*

*L'expression « Logiciel libre » fait référence à la liberté pour les utilisateurs d'exécuter, de copier, de distribuer, d'étudier, de modifier et d'améliorer le logiciel. Plus précisément, elle fait référence à quatre types de liberté pour l'utilisateur du logiciel : Un programme est un logiciel libre si les utilisateurs ont toutes ces libertés.*

- La liberté d'exécuter le programme, pour tous les usages (liberté 0).*
- La liberté d'étudier le fonctionnement du programme, et de l'adapter à vos besoins (liberté 1). Pour cela, l'accès au code source est une condition requise.*
- La liberté de redistribuer des copies, donc d'aider votre voisin, (liberté 2).*

---

131. Le site de la FSF se trouve à l'adresse : <http://www.fsf.org> et la description en français des Logiciels Libres peut être consultée à : <http://www.fsf.org/philosophy/free-sw.fr.html>

- 
- *La liberté d'améliorer le programme et de publier vos améliorations, pour en faire profiter toute la communauté (liberté 3). Pour cela, l'accès au code source est une condition requise.*

*Un programme est un logiciel libre si les utilisateurs ont toutes ces libertés. Ainsi, vous êtes libre de redistribuer des copies, avec ou sans modification, gratuitement ou non, à tout le monde, partout. Être libre de faire ceci signifie (entre autre) que vous n'avez pas à demander ou à payer pour en avoir la permission. »*

Par extension, on parle de contenu libre pour des documents dont le contenu est librement consultable et copiable, voire parfois modifiable par tout le monde, notamment les documents régis par la GFDL, comme l'encyclopédie libre, ouverte et en ligne Wikipédia<sup>132</sup>.

Richard Stallman explique comment il est venu au « Libre » en introduction à la plupart de ses conférences, histoire que nous retrouvons dans sa biographie (Williams, 2002) « Free as in Freedom ». C'est à cause d'une colère qui date des années 70, alors qu'il travaillait au laboratoire d'intelligence artificielle du prestigieux Massachusetts Institute of Technology. Xerox fournissait son laboratoire en imprimantes, au départ ils avaient reçu un modèle et le logiciel qui permettait de la gérer, code source compris. A l'utilisation, ils s'étaient aperçus que ce logiciel réclamait plusieurs modifications, pour gérer les impressions par plusieurs utilisateurs, éviter les bourrages papiers à cause de travaux qui s'enchaînaient trop rapidement, etc. Disposant du code source, ils avaient adapté le pilote de l'imprimante et même si elle était assez lente, elle répondait parfaitement à leur besoins : quand une impression était demandée, il suffisait d'attendre suffisamment longtemps avant de passer la récupérer dans le bac de sortie de l'imprimante et elle s'y trouvait à coup sur. Puis Xerox

---

132. <http://fr.wikipedia.org/wiki/>

a remplacé leur vieille imprimante par un tout nouveau modèle, beaucoup plus rapide, mais cette fois le logiciel de contrôle de l'imprimante était fourni sans son code source. L'imprimante était très rapide, mais mal gérée elle se bloquait souvent et il fallait à chaque fois se déplacer pour la débloquer. Au final, il était beaucoup plus long d'imprimer avec cette imprimante rapide qu'avec l'ancienne. L'imprimante ne lui obéissait pas, et il ne pouvait accéder au logiciel qui la contrôlait pour le modifier lui-même, selon ses désirs et ses besoins. « C'est secret, protégé par des droits et des brevets », lui rétorquait le constructeur. Sans accès au code source, Richard Stallman ne pouvait adapter lui-même le logiciel rétif. Et aurait-il pu y accéder qu'il n'aurait pas eu le droit d'intervenir sur leur contenu, ni de diffuser ses perfectionnements à d'autres utilisateurs de la même imprimante. Quelques années plus tard, en 1985, Stallman fondait la Free Software Foundation, avec comme objectif de créer des logiciels diffusés avec leur code source. Des logiciels que tout un chacun pouvait copier à l'infini, modifier au gré de ses besoins, et redistribuer à sa convenance. Des logiciels que personne ne pouvait s'approprier. Des logiciels libres développés dans le cadre du projet GNU<sup>133</sup>. Nous verrons plus loin dans la section B., page 188, traitant des composants logiciels utilisés dans *CoLiSciences* quelques exemples d'outils qui sont les fruits de ce projet.

---

133. Il existe plusieurs plaisanteries autour de ce nom. Tout d'abord il se prononce comme l'animal, le « Gnou », qui est par ailleurs l'emblème du projet. Mais c'est aussi un acronyme « récursif », qui signifie *GNU Not Unix*, c'est-à-dire « GNU n'est pas Unix », ce qui est un double jeu de mots. D'une part les acronymes récursifs sont un des jeux de mots favoris des *hackers*, parce qu'ils s'apparentent à une technique de programmation difficile à maîtriser par les néophytes : un programme s'appelle lui-même, comme dans la parabole de l'oeuf et de la poule. Bien entendu dans le cas de la programmation récursive il faut qu'une condition à un moment donné décide de l'oeuf à partir duquel la lignée de poules est née, c'est-à-dire une condition de sortie qui fait sortir le programme de cette boucle *a priori* infinie d'auto-appels. Pour revenir aux acronymes, il sont construits tels que la première lettre est une répétition de l'acronyme lui-même, GNU est GNU Not Unix, PHP est PHP Hypertext Processor. D'autre part, GNU se voulant un système « tel que » Unix mais sans en reprendre les contraintes de propriété intellectuelle et commerciales, GNU n'est pas Unix est un nom effectivement plein d'ambiguïté.

Il est très intéressant de mettre en parallèle cette recherche de liberté, qui a amené Richard Stallman à créer la *Free Software Foundation* et l'analyse que Gilbert Simondon nous donne concernant le danger d'aliénation à un objet technique, ceci dès l'introduction « *Du mode d'existence des objets techniques* » :

« *La plus forte cause d'aliénation dans le monde contemporain réside dans cette méconnaissance de la machine, qui n'est pas une aliénation causée par la machine, mais par la non-connaissance de sa nature et de son essence, par son absence du monde des significations, et par son omission dans la table des valeurs et des concepts faisant partie de la culture. (Simondon, 1969, pages 9-10)* »

La plupart des utilisateurs d'un ordinateur expriment ce sentiment d'aliénation, ils butent sur une opération qui ne veut pas aller à son terme, ils ne trouvent pas le chemin ou les opérations à mettre en œuvre pour réaliser un travail particulier. Ce sentiment d'aliénation et d'incompréhension conduit le plus souvent à la technophobie, un rejet complet de la technologie. Pour ramener à eux les sceptiques et les déçus, la réponse des marchands de technologie est de la rendre plus « transparente », au sens d'invisible, ils ajoutent encore plus de technologie, pour renforcer les processus d'automatisation et rendre les machines plus simples d'emploi, car plus complètement automatisées. C'est ainsi que nous voyons fleurir les dispositifs et technologies qui devraient nous libérer de la technique : le « *plug and play* » qui permet d'ajouter et de connecter n'importe quel périphérique à un ordinateur sans devoir passer par de plus ou moins longues étapes de configuration, plus trivialement le « *wysiwyg* » qui permet de taper son texte sans se préoccuper de ce qui va se passer à l'impression, puisque « ce que l'on voit, c'est ce qui va être

imprimé<sup>134</sup> ». Mais en fait, ils ne font que renforcer la frustration devant une machine qui échappe de plus en plus à son utilisateur.

*« Les idolâtres de la machine présentent en général le degré de perfection d'une machine comme proportionnel au degré d'automatisme. (Simondon, 1969, page 11) »*

Une autre façon d'éviter la technophobie peut conduire aussi au besoin d'en savoir plus. Cette fois c'est du côté des utilisateurs que nous trouvons la réponse. Certains vont se donner les moyens de maîtriser leur outil en passant du temps à se former à son maniement. Le problème de « se former à l'utilisation d'un ordinateur » est que cela est à la fois très vague et très complexe. Très vague parce que nous n'avons pas dit s'il s'agissait de se former à son utilisation, comme utilisateur d'applications déjà prêtes ou à la programmation, pour être capable de développer de telles applications. Dans le premier cas, il va falloir passer par le choix des applications sur lesquelles il est nécessaire de se former, suivant que ces applications sont génériques (traitement de texte, tableur, ...) ou spécifiques (gestion de stock, paye, ...) le parcours sera différent, d'autre part il est aussi nécessaire de se former à la méta-application que constitue le système d'exploitation. Dans le deuxième cas, le problème est encore plus complexe puisque cette fois nous pouvons aller jusqu'à la construction de la machine comme dispositif matériel, puis pourquoi pas, imaginer de développer un système d'exploitation ou enfin une application qui va s'exécuter dans l'environnement d'un système précis. Le développement d'un applicatif peut en fait se faire dans des conditions assez différentes suivant qu'il sera mono ou multi utilisateur, suivant qu'il nécessitera de se connecter à une autre machine ou non, suivant qu'il utilisera des bases de données ou non, etc. Mais, dans tous ces cas, il faut pouvoir

---

134. traduction littérale de « *wysiwyg* ».

« soulever le capot », accéder au moteur pour en analyser les organes, comme cela est dit à nouveau par Gilbert Simondon :

*« Les objets techniques qui produisent le plus d'aliénation sont ceux qui sont destinés à des utilisateurs ignorants. De tels objets se dégradent progressivement : neufs pendant peu de temps, ils se dévaluent en perdant ce caractère, parce qu'ils ne peuvent que s'éloigner de leurs conditions de perfection initiale. Le plombage des organes délicats indique cette coupure entre le constructeur qui s'identifie à l'inventeur et l'utilisateur qui acquiert l'usage de l'objet technique uniquement par un procédé économique. (Simondon, 1969, page 250–251) »*

### 3. Open Source

Bien que l'expression Open Source soit une marque appartenant à l'Open Source Initiative, elle est employée à tort et à travers pour désigner tout ce qui se rapproche peu ou prou du libre. Ceci a contribué à en faire une notion marketing aux contours flous et mous. En toute rigueur, le terme Open Source ne devrait être employé que pour les licences qui répondent aux critères de l'OSD et qui ont fait l'objet d'une certification mise en place par l'OSI. La traduction française du terme *Open Source* n'est pas très explicite : Source Ouvert littéralement, le sens est plus proche de « Source Disponible » ou « Source Lisible ». Ce terme caractérise les logiciels informatiques dont les codes sources ne sont pas secrets et provient de l'Open Source Initiative (OSI)<sup>135</sup>, association fondée par Eric S. Raymond qui a déposé la marque Open Source (TM) et qui a défini les critères nécessaires afin de pouvoir l'utiliser. Comme pour les Logiciels Libres, il existe une définition précise de l'Open Source, dont voici les critères essentiels :

---

135. Le site de l'Open source Initiative se trouve à l'adresse : <http://www.opensource.org/>

**Libre redistribution :** la licence ne doit pas prévoir de redevance en échange de l'accès au logiciel, mais ne doit pas non plus interdire qu'une redevance puisse-être exigée ;

**Code source disponible :** la licence doit prévoir la possibilité d'accéder au code source. Le programme doit inclure le code source, et la distribution sous forme de code source comme sous forme compilée doit être autorisée. Si un mode de distribution n'inclut pas le code source, il doit exister un moyen clairement indiqué de se procurer ce dernier pour un coût raisonnable destiné à compenser les frais de reproduction, ou, de préférence, de le télécharger sans frais supplémentaire depuis l'Internet. Par «code source» nous désignons ici la forme la plus adéquate pour un programmeur qui souhaiterait modifier le programme. Il n'est pas autorisé à proposer un code source rendu volontairement difficile à comprendre. Il n'est pas autorisé à proposer des formes intermédiaires, comme ce qu'engendre un préprocesseur ou un convertisseur ;

**Travaux dérivés :** la licence doit offrir les libertés de modifier et de distribuer les travaux dérivés en vertu des mêmes conditions que le permis du logiciel initial ;

**Intégrité du code source de l'auteur original :** la licence, tout en encourageant les modifications et donc l'évolution des programmes, doit garantir la paternité de l'auteur, pour cela, elle peut prévoir que les modifications du logiciel ne pourront être distribuées que sous la forme de fichiers correctifs (patch files). Par ailleurs, la licence doit expressément permettre la distribution du logiciel combiné provenant du code source modifié, enfin, la licence peut requérir que les travaux dérivés portent des noms ou des numéros de versions distincts du logiciel original ;

**Pas de discrimination :** la licence ne doit pas permettre de discrimination contre des personnes, des groupes ou contre certains secteurs d'activités (notamment commerciaux). La licence ne doit pas limiter le champ d'application du programme. Par exemple, elle ne doit pas interdire l'utilisation du programme à des fins commerciales ou dans le cadre de la recherche génétique ;

**Distribution de la licence :** la distribution du logiciel ne doit pas être soumise à des exigences supplémentaires tel qu'un accord de non-révélation ;

**La licence ne doit pas être spécifique à un produit :** Les droits accordés à l'utilisateur du programme ne doivent pas dépendre du fait que le programme appartient à un ensemble donné. Si le programme est extrait d'une distribution et utilisé ou distribué selon les conditions de sa licence, toutes les parties auxquelles il est ainsi redistribué doivent bénéficier des droits accordés lorsque que le programme se trouve au sein d'un ensemble ;

**La licence ne doit pas contaminer d'autres logiciels :** La licence ne doit pas imposer de restrictions à d'autres logiciels distribués avec le programme. Il n'est pas nécessaire que les logiciels qui accompagnent celui soumis à la licence soient également open source ;

Dès lors, comme pour les logiciels libres, le fait de disposer des sources d'un logiciel ne suffit pas à dire qu'il est Open Source. Dans tous les cas, on se référera à la licence d'utilisation du logiciel pour vérifier les droits de redistribution et de travaux dérivés qui sont les compléments indispensables de la disponibilité du source.

## B. Les licences libres

Pour le non spécialiste il est difficile au premier abord de ne pas faire l'amalgame entre Logiciels Libres et Open Source, leur point commun le plus apparent étant la mise à libre disposition du source des logiciels, il faut aller dans les détails des deux licences pour comprendre quelles sont les différences. Certaines licences « Logiciels Libres » sont souvent reprises comme étant des licences « Open Source » qui de nom déposé est devenu un nom générique comme le « walkman » de la société Sony. Dans les deux cas, le fait que le nom soit d'une consonance anglaise a sans doute facilité son introduction et son adoption pour l'identification d'une nouveauté. Ce sont les deux « courants » majeurs de cette nouvelle manière de penser la création et la diffusion à l'ère du numérique, plusieurs sociétés ou organismes les ont reprises à leur compte telles quelles ou avec des adaptations, ce qui a amené à la création de nombreuses licences dérivées. A y regarder de plus près, l'esprit qui les anime toutes est le même et ce sont les règles de mise en application qui les composent qui diffèrent : la diffusion ou la transmission peut se faire dans un cadre plus ou moins restrictif, le cadre d'application est adapté à la législation d'un pays, à un objet technique particulier. Au-delà de la différence sémantique des deux termes Open Source et Logiciel Libre, qui engendre quelques débats, il faut savoir que ces deux mouvements sont en accord sur la plupart des recommandations pratiques et qu'ils travaillent ensemble sur de nombreux projets spécifiques. Pour simplifier, la volonté de l'Open Source Initiative est de fournir un programme de marketing pour le logiciel libre, notamment en établissant un label de certification reconnu et estimé. La position de la Free Software Foundation est plus « idéologique » en mettant l'accent sur la signification et la précision du mot « libre ». En principe tous les Logiciels Libres sont Open Source, mais la réciproque n'est

pas toujours vraie. La différence provient du fait que les libertés inhérentes sont plus accentuées dans les licences « Logiciel Libre ».

## 1. Licences pour les logiciels

La liste suivante est celle des licences disposant du label de certification OSI qui assure le respect de l'Open Source Definition<sup>136</sup>. Cette liste n'est là que pour montrer d'une part la complexité engendrée par le nombre de licences libres existantes mais aussi pour mettre en évidence que l'ouverture permet l'adaptabilité : tous les grands projets libres ont souvent développé une licence spécifique et adaptée à leur problématique de développement et de diffusion. Nous n'avons pas systématiquement relevé l'adresse du site qui donne les détails de chacune de ces licences, mais une requête dans un moteur de recherche comportant le nom de la licence donnera généralement cette information en premier dans sa liste de réponse.

- GNU General Public License<sup>137</sup>.
- GNU Lesser General Public License<sup>138</sup>.
- The BSD licence<sup>139</sup>.
- CeCILL : première licence française de logiciel libre.
- MIT license
- Artistic license
- Mozilla Public License v. 1.0 (MPL)

---

136. Cette liste est en constante évolution, la faire apparaître ici permettra de constater son large spectre de couverture et la complexité qui est en fait derrière. Une liste mise à jour régulièrement est disponible à l'adresse <http://www.opensource.org/licenses/index.html>. Un autre document sur le sujet est très intéressant à prendre en compte, il s'agit du *Guide de choix et d'usage des licences de logiciels libres pour les administrations* publié par l'ADAE à l'adresse [http://www.adae.gouv.fr/article.php3?id\\_article=172](http://www.adae.gouv.fr/article.php3?id_article=172).

137. <http://www.fsf.org/copyleft/gpl.html> ou en français <http://www.linux-france.org/article/these/gpl.html>

138. <http://www.fsf.org/copyleft/lesser.html> ou en français <http://www.linux-france.org/article/these/licence/lgpl/>

139. <http://www.debian.org/misc/bsd.license>

- Qt Public License (QPL)
- IBM Public License
- MITRE Collaborative Virtual Workspace License (CVW License)
- Ricoh Source Code Public License
- Python license
- Python Software Foundation License
- zlib/libpng license
- Apache Software License
- Vovida Software License v. 1.0
- Sun Industry Standards Source License (SISSL)
- Intel Open Source License
- Mozilla Public License 1.1 (MPL 1.1)
- Jabber Open Source License
- Nokia Open Source License
- Sleepycat License
- Nethack General Public License
- Common Public License
- Apple Public Source License
- X.Net License
- Sun Public License
- Eiffel Forum License
- Motosoto License
- Open Group Test Suite License
- W3C software notice and license<sup>140</sup>.

---

140. <http://www.w3.org/Consortium/Legal/copyright-software-19980720>  
et <http://www.w3.org/Consortium/Legal/ipr-notice-20000612>

- La licence « artistique »<sup>141</sup>, publiée sur le site du langage Perl. Il existe plusieurs versions de la licence artistique, incluant la Clarified Artistic License.

Les quatre premières licences sont actuellement les plus répandues. La licence Artistic qui, entre autre, concerne le langage PERL et ses modules, ainsi que la licence Mozilla sont également jugées comme importantes.

## 2. Licences pour les documents et créations artistiques

Parce que les œuvres numériques peuvent se copier, se modifier et évoluer à la manière des logiciels, des auteurs et des artistes se sont inspirés des licences logiciels pour protéger leur droit d’auteur tout en facilitant la diffusion de leur créations à travers le Web. Ils ont rédigé des licences plus précises pour leur domaine et adaptées à d’autres types de réalisations, cette fois du domaine artistique et culturel. Il faut noter que c’est un juste retour des choses, puisqu’à l’origine les programmeurs désirent protéger leurs créations par un simple droit d’auteur, faisant valoir qu’un programme n’est rien d’autre qu’un travail intellectuel. Le même foisonnement existe là encore, pour preuve suit une liste non exhaustive des licences principales. Encore une fois, du fait de leur reconnaissance sur le Web, une simple requête dans un moteur de recherche permet d’identifier la page sur laquelle trouver plus de détail sur chacune d’elles.

- Academic Free License
- Licence Art Libre<sup>142</sup>
- Licence de Libre Diffusion des Documents – LLDD version 1 et sa version anglaise : Free Document Dissemination Licence – FDDL version 1.

---

141. <http://www.perl.com/pub/a/language/misc/Artistic.html> ou en français <http://www.linux-france.org/article/these/licence/artistic/fr-artistic.html>

142. <http://artlibre.org/licence/lal.html>

- GNU Free Documentation License Version 1.1 (mars 2000) par la Free Software Foundation.
- Les licences du groupe Opencontent :
  - OpenContent License (OPL)<sup>143</sup>, pour du contenu ouvert.
  - Open Publication License.
- W3C document notice and license.
- EFF Open Audio License (Version 1.0) : Licence Libre Audio par l'Electronic Frontier Foundation.
- Le contrat social de l'Open Directory Project (ODP), index libre du Web.
- proposition de GNUArt pour appliquer la licence GNU GPL à une œuvre d'art.
- Design Scientific License par Michael Stutz <sup>144</sup>, pour la science et la création artistique.
- Creative Commons<sup>145</sup> : Il s'agit là sans doute du portail d'information sur les licences libres le plus connu. en plus d'information générales sur pourquoi et comment se protéger, surtout dans quel but. Un artiste peu connu n'aura pas la même attitude qu'un autre qui déjà reconnu, un particulier qui désire partager des photos et obtenir des critiques et des conseils n'aura pas les mêmes restrictions qu'un photographe professionnel dont c'est le gagne pain. Tous devraient trouver sur ce self-service la licence qui leur convient.
- The Free Music License par Ethymonics (Tony Hardie-Bick) : <sup>146</sup> Pour la musique.

---

143. <http://opencontent.org/opl.shtml>

144. <http://dsl.org/>

145. <http://www.creativecommons.org/>

146. <http://www.ethymonics.co.uk>

- The Free Music Public License par vnatrc-bortch.org<sup>147</sup> Pour la musique (en français).
- The Open Audio Licence par l'Electric Frontiere Foundation :<sup>148</sup> Pour la musique.
- La Licence Publique Multimédia par Vidéon :<sup>149</sup> Pour la vidéo.
- La Licence Publique de Traduction (LPT-FR) par I.R.Maturana<sup>150</sup> Pour toutes les traductions de toutes les langues.
- La Licence HyperNietzsche par Paolo D'Iorio :<sup>151</sup> Elle est destinée à la recherche intellectuelle et permet aux chercheurs travaillant sur Nietzsche de publier leurs travaux en ligne sur le site HyperNietzsche<sup>152</sup>, mais elle pourrait être réutilisée dans n'importe quel autre contexte de publication scientifique en ligne.
- La Licence Ludique Générale par la société des jeux libres<sup>153</sup> Pour les jeux de rôles.
- la GFDL (GNU Free Documentation License)<sup>154</sup> Pour la diffusion de documents, sans droit de modifier.
- La Licence de Libre Diffusion de Documents de Bernard Lang<sup>155</sup> Pour les documents écrits (avec un souci particulier pour les bases de données).
- La Licence de libre diffusion Mnémosyne Projet Mnémosyne contre la haine<sup>156</sup>

---

147. <http://www.musique-libre.com>

148. [http://www.eff.org/IP/Open\\_licenses/20010421\\_eff\\_oal\\_1.0.html](http://www.eff.org/IP/Open_licenses/20010421_eff_oal_1.0.html)

149. <http://www.videontv.org/licence/>

150. <http://www.in3activa.org/doc/fr/LPT-FR.html>

151. <http://www.hypernietzsche.org/doc/puf/chevet2/>

152. <http://www.hypernietzsche.org/>

153. <http://jeuxlibres.free.fr/llg.htm>

154. <http://www.gnu.org/copyleft/fdl.html>

155. <http://pauillac.inria.fr/~lang/licence/lldd.html>

156. [http://www.contenuouvert.org/article.php3?id\\_article=11](http://www.contenuouvert.org/article.php3?id_article=11)

- La Licence du GESI (organe qui relie tous les départements GEII des IUT) <sup>157</sup> Destinée à la publication de documents pédagogiques et de travaux de recherche.

## C. Les textes fondateurs

Plus intéressant que ces longues listes de textes juridiques, nous allons maintenant tenter de nous approcher de cet esprit commun à tous les projets de type « libre ». Pour cela nous allons commencer par revenir sur quelques textes fondateurs de ce mouvement, qui à son origine ont permis à des individus isolés de se reconnaître dans cette démarche et qui ainsi en constituent le ciment.

### 1. La cathédrale et le bazar

Cette expression est le titre d'un texte de Eric S. Raymond dans lequel la communauté des hackers s'est immédiatement reconnue. Il met en opposition de façon imagée deux modèles de développement du logiciel<sup>158</sup>.

D'un côté nous trouvons le modèle traditionnel de la « cathédrale » : l'écriture du logiciel y est une procédure fortement structurée avec des étapes bien séparées et bien hiérarchisées. De la détermination des besoins du produit (ou DSL : dossier de spécification du logiciel) aux essais in situ, le logiciel se construit d'une manière globalement linéaire. De l'autre côté, le modèle de type « bazar » : même si un groupe possède la direction d'un projet, il est à l'écoute des propositions surgissant à l'horizontale (c'est-à-dire de l'ensemble potentiel des utilisateurs), et se sert même des utilisateurs

---

157. <http://www.gesi.asso.fr/documentlibre/CharteGe.pdf>

158. Une version française de ce texte (traduit par Sébastien Blondeel) est disponible à l'adresse <http://www.linux-france.org/article/these/cathedrale-bazar/cathedrale-bazar.html>.

volontaires en tant que co-développeurs. Ce mode de développement n'est plus pyramidal et linéaire, le logiciel ne se développe pas par des mécanismes disposés en série, mais par des voies travaillant en parallèle. Cette disposition en parallèle des efforts des co-développeurs permet d'être plus efficace dans l'implémentation des diverses fonctionnalités du logiciel, d'accélérer son débogage (phase qui est d'ordinaire grande consommatrice de temps), et de laisser libre l'expression de nouveaux besoins, la suggestion d'idées nouvelles... La condition de la réussite est liée à la capacité du noyau dur du projet de savoir distinguer les contributions utiles des agitations vaines.

*« Le monde Linux sous de nombreux aspects, se comporte comme un marché libre ou un écosystème, un ensemble d'agents égoïstes qui tentent de maximiser une utilité, ce qui au passage produit un ordre spontané, auto-correcteur, plus élaboré et plus efficace que toute planification centralisée n'aurait pu l'être. (Raymond, 1998) »*

## 2. Le chaudron magique

Dans le chaudron magique(Raymond, juin 1999, traduit en juillet), Eric Raymond fait référence au mythe gallois de la déesse Ceridwen pour lancer une analyse du fonctionnement du substrat économique du phénomène « Open Source ».

*« Dans le mythe gallois, la déesse Ceridwen possédait un grand chaudron qui produisait magiquement une riche nourriture – quand elle le lui ordonnait par un sort connu par elle seule. Dans les sciences modernes, Buckminster Fuller nous a donné le concept de « l'éphéméralisation », une technique qui devient d'autant plus efficace et moins chère que les ressources physiques investies dans les premiers efforts de développement sont remplacées par de plus en plus de contenu informatif. Arthur C.*

*Clarke a relié ces deux phénomènes en observant que « toute technique suffisamment avancée est indiscernable de la magie<sup>159</sup> ».*

*Pour beaucoup de personnes, les succès de la communauté du source ouvert ont l'apparence d'une improbable magie. Des logiciels de haute qualité se matérialisant à partir de « rien », c'est bien tant que ça dure, mais cela semble peut praticable dans le monde de la concurrence, où on ne dispose que de ressources limitées. Quel est le piège ?*

*Le chaudron de Ceridwen n'est-il qu'un tour de passe-passe ?*

*Dans la négative, comment « l'éphéméralisation » fonctionne-t-elle dans ce contexte - quelle est la formule magique de la déesse ? »*

Pour clarifier cette citation tirée de l'introduction du chaudron magique, il faut rappeler le contexte faisant référence à Fuller. Toutes les choses finissent par disparaître, et tout particulièrement celles produites par la main de l'homme. L'architecte Buckminster Fuller nous a donné le concept de « l'éphéméralisation » pour décrire la disparition d'une technologie en tant que telle dans la société, nous aurions aussi pu parler d'assimilation. C'est le processus qui petit à petit transforme une technologie que nous devons apprendre (voiture, téléphone, ordinateur, ...) en quelque chose qui « va de soi » et qui fait partie des acquis de la société. Ces technologies disparaissent en tant que technologie parce que nous ne les percevons plus comme quelque chose que nous devons apprendre, maîtriser, étudier. L'interface en devient si intuitive, si prégnante dans la société qu'elles semblent naturelles. Qui oserait expliquer dans quel sens le volant d'une voiture doit-il être tourné pour la faire changer de direction comme désiré ? Le concepteur doit prendre en compte l'environnement, les besoins humains, les ressources énergétiques mondiales et les capacités de l'industrie avant d'entreprendre tout projet.

---

159. La citation de Arthur C. Clarke provient de (Clarke, 1964), où elle apparaît comme la « troisième loi de Clarke ».

Le monde contemporain est un univers en perpétuel changement. Pour Fuller la technologie doit amenuiser sa propre présence et s'alléger jusqu'à la disparition.

Raymond nous rappelle ensuite qu'il existe de grandes différences entre les biens manufacturés traditionnels et les logiciels. Nous pouvons citer par exemple le coût de duplication comme élément différenciateur, il en existe de nombreux autres. Un autre aspect très spécifique aux logiciels est que nous sommes finalement restés dans un état largement artisanal, non pas que tout soit uniquement réalisé manuellement, mais la plupart sont réalisés « sur mesure » pour résoudre un problème particulier.

*« Le code écrit pour la vente n'est que la partie émergée de l'iceberg de la programmation. Dans l'ère qui a précédé les micro-ordinateurs, tout le monde savait que 90 % du code était écrit en interne par des banques et des compagnies d'assurances. Ce n'est probablement plus le cas – d'autres industries sont devenues bien plus dépendantes du logiciel depuis, et la part de l'industrie de la finance a sans doute baissé – les faits empiriques montrent que 95% du code est encore produit en interne par ceux qui en ont besoin. »*

L'idée de Raymond est de montrer que de développer ces logiciels « internes » sous la forme de projets à code ouvert est le moyen le plus fiable et le plus efficace, sans que cela n'engendre de risques supérieurs à ceux qui existent de toutes façons aussi avec l'utilisation de logiciels à sources fermés. Il va ainsi jusqu'à conclure sur la prédiction d'un futur où la majorité des logiciels seraient à code ouvert. Les effets de l'éphémérialisation tendent à faire intégrer petit à petit dans un corpus global de logiciels tous les logiciels ouverts déjà largement utilisés et acceptés, ils se trouvent finalement à faire partie d'une infrastructure informatique globale par intégration progressive.

En revanche, les monopoles des logiciels à code fermé disposeront d'une espérance de vie limitée par le fait que les catégories de produits qu'ils embrassent seront progressivement qualifiées d'infrastructure.

## D. La rencontre

Pour conclure cette introduction et revenir à notre projet *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, la rencontre avec les Logiciels Libres s'est faite après une première étape de définition de nos besoins quand nous avons sélectionné les technologies qui seraient mises en œuvre pour réaliser notre projet. C'est par les deux premiers services, Serveur Web et Gestionnaire de Base de Données, fondations de tout le reste de l'édifice logiciel que nous voulions bâtir, que nous avons lié nos besoins et notre budget à « l'éthique logicielle » telle qu'elle est décrite par la FSF, et surtout que nous avons pris conscience des enjeux sous-jacents dans la sélection d'une technologie, enjeux qui dépassent largement le technologique puisqu'ils nous amènent à nous questionner sur des notions de domaine public, de propriété intellectuelle et de pérennisation de patrimoine.

Nous avons déjà abordé<sup>160</sup> la question de la pérennisation dans le contexte de la numérisation d'un corpus, nous ne reviendrons donc pas en détail sur notre problématique de pérennisation. La chose importante à relever pour faire la transition avec l'objet de ce chapitre, est que cet objectif de pérennisation nous poussait à utiliser un format de stockage qui soit le plus standard, ou le plus utilisé, ou tout simplement, qui soit ouvert et dont nous puissions maîtriser tous les rouages internes si « techniques » soient-ils. Il existe sans doute de nombreux exemples où l'utilisation de logiciels « propriétaires », et surtout de formats de documents « propriétaires » a conduit à une perte de temps ou de ressources, voire à la perte des données elles-

---

160. section 1. *Pérennisation et patrimoine*, page 145

mêmes. Malheureusement, cette information n'est pas facilement accessible car elle met en évidence un dérapage, une erreur d'évaluation, sur laquelle les entreprises n'ont aucun intérêt à communiquer.

*« [...], le département de la Défense américain a dû mettre en place une coûteuse opération de « migration » en vue de restaurer les fichiers relatifs aux coordonnées géographiques des bombes lâchées durant la guerre du Vietnam. Ces données, stockées à l'époque dans une base de données « propriétaire », étaient devenues illisibles : plus aucun logiciel du marché ne permettait d'y accéder. (Boydens, 2004, page 14) »*

Dans cette citation, Boydens fait référence à (Ruggiero, 2003) qui avait présenté lors de la conférence « The Future of Digital Memory and Cultural Heritage<sup>161</sup> » six études de cas qui montraient les difficultés de la préservation ou de la sauvegarde de données numériques. Souvent c'est la disponibilité d'un matériel spécifique qui pose problème et dans la plupart des cas c'est un lecteur de bande d'un format ancien et particulier qui risque d'être la cause de la disparition d'un patrimoine.

Pour répondre à nos besoins de services Web coté serveur comme base de données, les logiciels Apache et MySQL étant les deux standards de fait sur la toile, comme nous le verrons page 187 dans la section 3. se sont imposés d'eux-mêmes. Loin des batailles commerciales, sans matraquage publicitaire mais avec derrière eux des centaines de développeurs prêts à s'investir personnellement, sur leur nom, pour assurer leur bon fonctionnement.

Partant de là, GNU/Linux s'imposait tout naturellement comme système d'exploitation et il ne nous restait plus qu'à définir la nature des services Web

---

161. *The Future of Digital Memory and Cultural Heritage* : conférence qui s'est tenue à Florence, en Italie les 16 et 17 Octobre 2003. Les présentations sont disponibles en ligne sur le site de la conférence à l'adresse [http://galileo.imss.firenze.it/memorie\\_digitali/](http://galileo.imss.firenze.it/memorie_digitali/)

qui seraient développés. En effet, même dans le monde du libre, plusieurs solutions s'offraient à nous : CGI, Perl, PHP, Python, ...

La trivialité de la réponse pourrait décevoir, mais notre choix s'est tout simplement porté sur le langage *à priori* le plus simple à mettre en œuvre pour quelqu'un qui n'avait jamais touché à aucun d'eux avant le début du projet, qui soit largement utilisé et pour lequel la communauté Internet, par son dynamisme, puisse nous fournir en applications. C'est donc le langage PHP que nous avons retenu pour animer les scripts du serveur et le gestionnaire de contenu PhpNuke pour servir de coquille vide.

L'utilisation de différents outils accessibles à travers Internet (messagerie, outils de collaboration pour le suivi des modifications, CVS pour le suivi des sources, le site lui-même qui grâce à ses outils d'administration en ligne peut être géré à plusieurs et à distance) nous a permis d'avancer rapidement sans être sur le même lieu géographique, dans le même esprit que celui qui anime généralement les projets « Open Source ».

Dans un futur proche et pour assurer une véritable pérennité au site, il serait important de stabiliser le code source de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* pour pouvoir le partager et trouver une nouvelle dynamique avec de nouveaux partenaires, non institutionnels mais qui partageraient nos mêmes préoccupations sur des corpus différents. Cela pourrait être interprété comme une manière de rendre à la communauté l'aide qu'elle nous a fourni, mais ces ont aussi des objectifs plus égoïstes qui pourraient nous animer. En effet, c'est aussi un moyen pour continuer les développements et les améliorations en créant une synergie de développeurs autour de la problématique des bibliothèques en ligne.

De nombreux sites proposent des ouvrages en ligne, chacun utilisant ses propres outils ou des assemblages différents d'outils disponibles sur le marché (commerciaux ou « open source »). En distribuant le moteur *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*

accompagné d'une licence de Logiciel Libre, nous pourrions espérer le voir traduit en différentes langues, le voir amélioré au niveau des fonctionnalités et des performances.

La modularité de phpNuke nous a permis d'ajouter nos propres outils à la coquille existante, cette modularité mieux maîtrisée permettrait de rendre le moteur *Co<sup>L</sup>i<sup>S</sup>ciences* adaptable à de nombreux problèmes de publication en ligne.

## 8.2 Nouveaux cadres pratiques, nouveaux modèles théoriques

### A. Quel statut ?

Les logiciels libres font souvent questionnement par rapport à leur statut économique et politique. Les idées qui sous-tendent ce nouveau modèle de développement sont fondées sur le travail collaboratif, la libre circulation des idées et le partage. Nous restons là dans des domaines qui restent assez éloignés de l'économie, pourtant comment ne pas faire le parallèle entre logiciels libres et le marché ? Dans les deux cas, le moteur qui anime et donne de la cohérence, c'est...

*« La rencontre de l'offre et de la demande, autrement dit la libre convergence – par la médiation de l'échange et sous réserve de la concurrence – des égoïsmes. Chacun est utile à l'autre, sans qu'on ait besoin de le forcer. Tous ne cherchent que leur propre intérêt, mais ne peuvent le trouver qu'ensembles. (Comte-Sponville, 2001, définition du marché, page 356) »*

Tout en restant dans le même registre, nous retrouvons ce parallèle formulé toutefois différemment dans un article en ligne qui présente le travail de « *celui par qui le code est parlé* », le programmeur de logiciels libres. Cette fois nous voyons une référence à la division du travail sur laquelle nous allons revenir.

*« Dès lors que nous entrons dans un domaine de production où l'objet n'a plus de forme matérielle, nous pouvons parfaitement envisager que cet objet soit, dans son essence même un objet instable, en perpétuelle évolution. Cette instabilité aurait alors pour conséquence de remettre en cause son statut de marchandise. On ne peut en effet vendre, comme marchandise, qu'un objet fini, qu'un produit ayant une existence propre, distincte de celui qui l'a produit. C'est loin d'être le cas d'un produit immatériel. Instable par essence, à moins qu'il ne soit protégé par un dispositif technique ou par une licence d'utilisation, cet objet peut donc être, à tout moment, modifié par quiconque a l'envie et la capacité technique de le faire. Dans ces conditions, les frontières inhérentes à la division du travail sont évidemment brouillées, mais c'est aussi la frontière entre producteur et consommateur qui est elle-même aussi remise en cause de manière fondamentale. Ayant le droit et la possibilité technique d'enrichir un savoir qui lui pré-existe, le consommateur, l'utilisateur devient lui-même acteur, producteur de savoir. (Blondeau, 2003) »*

Nous pouvons continuer cette description en la complétant par une présentation équivalente des logiciels propriétaires. Comme cela vient tout juste d'être évoqué par Olivier Blondeau, ils sont ce que Marx appelle la division du travail : une entreprise se spécialise dans la fabrication de logiciels et en décharge complètement ses clients, qui dans un même temps se déchargent eux-mêmes de toute participation à leur élaboration.

## B. Programmeur de Logiciel Libre ?

Pour certains, le programmeur de logiciel libre est un individu étrange : ses motivations restent floues, son rôle dans la société mal compris et de ce fait il est difficile de se le représenter autrement qu'à travers les images d'Épinal du *hacker* – bitouilleur<sup>162</sup>.

- Est-ce un jeune étudiant qui passe tout son temps libre avec son ordinateur parce qu'il ne supporte pas de se mêler aux autres ?
- Est-ce un idéaliste ou pire un anarchiste qui cherche par ce biais à déstabiliser les entreprises représentant pour lui un « satan » capitaliste ?

### 1. Motivation

Il est vrai que quelqu'un qui participe à un projet de développement de Logiciel Libre est quelqu'un qui investit beaucoup de son temps sans forcément avoir en face de chaque heure travaillée une heure payée sur un bulletin de salaire. Se pose donc le problème de sa motivation. Pourquoi donnent-ils gratuitement ce qu'ils pourraient facturer des centaines d'Euros de l'heure ? Tout cela procède de la quête d'une réputation, c'est un étonnant mélange d'altruisme et d'individualisme.

**Individualisme** Souvent, un programmeur-bitouilleur écrit des programmes parce que c'est ce qu'il aime faire, comme un pianiste fait ses gammes et finit par trouver une jouissance dans l'aisance que cela lui procure. Un bitouilleur est donc quelqu'un qui a programmé pendant des années, et pour qui la programmation n'est pas une chose ennuyeuse ni même un travail, mais une

---

162. Hacker : « bitouilleur » bien que le grand public fait généralement l'amalgame entre hackers et pirates. En fait ce qualificatif représente plutôt des « bidouilleurs de bits », voire des « touilleurs de bits », des « bitouilleurs », qui sont plongés dans le code des programmes pour essayer de les optimiser et pousser leur performance. Le terme bitouilleur est construit en référence à bidouilleur et comme un touilleur de bits, c'est-à-dire quelqu'un qui sait mélanger les bits de manière à construire des bidouilles

activité intellectuelle qui lui apporte une satisfaction suprême. La plupart des projets Open Source prirent leur essor grâce à un sentiment de frustration, que Éric Raymond qualifie de « gratter une frustration » et qui aboutit à une prise de pouvoir. Généralement, après avoir recherché vainement un outil capable d'accomplir une tâche donnée, ou en avoir découvert un qui n'était pas au point, abandonné, ou mal maintenu, le bitouilleur se met au travail pour développer son propre outil. Aux débuts de l'informatique, du fait que les solutions commerciales étaient très rares ou hors de prix, voire les deux, c'était en fait la seule solution. Ce phénomène s'est produit avec les premiers systèmes professionnels et s'est répété ensuite à nouveau avec les systèmes personnels. C'est de cette façon qu'Eric Raymond commença à développer Fetchmail, que Larry Wall amorça le projet Perl, que Linux Torvalds créa Linux. La prise de pouvoir, pour plusieurs raisons, était le concept le plus important sous-jacent à la motivation de Stallman pour démarrer le projet GNU.

**Altruisme** Cette prise de pouvoir est aussi un moyen de se faire connaître, et surtout reconnaître des autres. La valeur du travail d'un programmeur provient du fait qu'il le partage avec d'autres, parce que c'est l'intérêt que les pairs pourraient porter à ce travail qui lui donne de la valeur. Il s'agit là d'une forme de reconnaissance par les pairs, un compagnonage moderne. Là où un horloger ne s'arrêterait pas au cadran de la pendule, mais en ouvre le coffre pour étudier les rouages, le programmeur a besoin du code source pour examiner ces mêmes rouages qui animent le logiciel. En France aussi, dans les années 80, sans Internet et pour de simples micro-ordinateurs personnels, de nombreuses revues proposaient les sources des programmes envoyés par d'autres lecteurs, dans certaines revues le programme du mois

était listé avec une courte présentation de « l'auteur » et sa photographie. Le bitouillage devenait de cette façon une sorte d'ascenseur social, le programme retenu faisant la gloire de son auteur. Chaque programmeur d'un logiciel libre est connu et reconnu pour son travail en tant qu'individu ; dans une grande compagnie éditrice de logiciels, d'après son contrat de travail tout ce qu'il produit serait la propriété de sa compagnie et il ne serait en fait qu'un anonyme au sein d'une armée de fourmis laborieuses. Ces contributeurs n'ont sans doute pas les poches pleines de stocks options ou d'actions, mais chacun bénéficie d'une réputation qui devrait lui assurer des occasions lui permettant de payer le loyer et de nourrir ses enfants.

## **2. Éthique du programmeur**

Pour le programmeur éthique, la programmation est une culture du don : si vu de l'extérieur, le fonctionnement du modèle « Logiciel Libre » peut paraître paradoxal, la réponse se trouve, en partie, au-delà de la notion de rétribution du travail. La volonté de participer dans un projet de Logiciel Libre s'assimile à la volonté de participer à un projet humanitaire, l'enjeu est encore une fois individuel et altruiste. Tout en respectant l'individu « programmeur », le projet le fera s'engager dans une démarche qui va bien au delà du simple fait d'assurer des revenus à son entreprise. Plutôt qu'une fourmi laborieuse, le programmeur devient un des maillons d'un grand projet mondial.

## **3. Plus simplement**

La dernière explication est tout à fait triviale, le développeur d'un logiciel libre peut aussi être tout simplement le salarié d'une entreprise qui lui demande de développer ce logiciel pour répondre à un besoin interne. Comme

cela est dit dans (Raymond, juin 1999, traduit en juillet), une société qui souhaite développer un nouveau système de comptabilité spécialisé et qui emploie un développeur pour cela peut trouver intéressant de faire appel à un développement partagé suivant un modèle Logiciel Libre.

Quel avantage tirerait-elle de fermer le code ?

- Ce n'est pas de pouvoir revendre le logiciel, il n'est pas développé dans ce but et la société n'a pas la structure pour se transformer en « usine » logicielle.
- Fermer le code et ne pas partager le logiciel va permettre d'en empêcher l'utilisation par un concurrent. Le but est d'éviter qu'une autre société concurrente puisse bénéficier des services du logiciel et en l'utilisant puisse améliorer son chiffre d'affaires aux dépens de la première. Mais en fait, la question n'est pas juste de compter ce qui est perdu. Il faut comparer ce que la société qui est à l'origine du logiciel peut gagner d'un côté et ce qu'elle risque de perdre de l'autre. En ouvrant le code de son logiciel, elle a bénéficié d'améliorations qu'elle n'aurait sans doute jamais été en mesure d'apporter seule, elle a donc bénéficié d'un gain de productivité qu'elle n'aurait pas pu obtenir sans l'aide de la communauté qui a participé à l'écriture du logiciel. En contre-partie, le gain de productivité du passager clandestin est sans doute très inférieur à la somme de ce qu'elle a économisé en développement et ce qu'elle a gagné avec utilisation du logiciel optimisé.

Un certain nombre d'entreprises sont capables de tenir ce raisonnement, et financent ainsi chacune une petite partie du développement des outils nécessaires à leur fonctionnement, alimentant au passage la communauté des logiciels libres.

## C. Logiciel libre et innovation

Quinze ans après la création de la FSF par Richard Stallman, Linux ou plutôt GNU/Linux pour faire référence à sa double origine, système d'exploitation libre, fruit du travail de Stallman, de Linus Torvalds, rejoints par des milliers d'anonymes, encercle Microsoft, dont le propre logiciel Windows domine le marché des systèmes d'exploitations pour ordinateurs personnels et serveurs. Jusqu'à présent, cette entreprise avait su écarter de son chemin tous ses concurrents, mais cette fois-ci rien n'y fait : le tout puissant Microsoft ne se heurte pas à une entreprise classique, mais à des milliers de programmeurs bénévoles, répartis sur toute la planète et animés par le seul désir de proposer un logiciel complet, efficace et ouvert à tous.

Ce qu'il faut surtout noter, c'est que depuis la création de la FSF, le Bazar (par référence à l'ouvrage de Eric Raymond déjà cité) s'est étendu à la mesure de la montée en puissance de l'informatique. Les applications libres se comptent maintenant par milliers et il semble que tous les champs de l'activité humaine soient couverts. Si les logiciels ont naturellement inauguré ce mouvement, c'est grâce à leur intimité avec le numérique. Offert à la copie, le numérique permet un clonage parfait, par rapport aux techniques analogiques qui introduisent du bruit ou des pertes d'informations à chaque reproduction. De plus la diffusion en est rapide et facilitée, pour un coût proche de zéro. Or l'informatique est aujourd'hui omniprésente, les ordinateurs sont reliés en réseau grâce à l'Internet, et, progressivement, c'est tout le savoir qui se retrouve sous forme de 0 et de 1. En basculant du vinyl au disque compact, la musique a suivi ce chemin. Pour ensuite se marier à l'ordinateur, puis au Web et à la diffusion massive, par la magie d'un format, le MP3. La vidéo et les films suivent la même voie, avec le DVD vidéo, et l'augmentation des débits sur l'Internet. Tandis que la littérature, elle, existe

sous forme de fichiers numériques depuis longtemps ; L'ordinateur, d'outil de travail austère, se branche sur le quotidien, se pare d'extensions pour jouer de la musique, ou regarder des films. Tandis que s'annoncent déjà les livres à encre digitale, rechargeables à volonté. Les internautes samplent, triturent, diffusent, agrègent, améliorent ces flux de bits. Avec la même inconscience qu'ils photocopiaient quelques pages de leurs livres favoris pour les donner à leurs amis. Avec la même naïveté qu'ils découpaient des photos, puis les recollaient ensemble pour offrir un montage personnalisé à leur entourage. Avec la même bonne foi qu'ils enregistraient des extraits de musique à la radio, ajoutaient leur voix, pour composer une chanson de leur cru. Mais aujourd'hui, leur public potentiel, c'est la planète. Si au début du XVIII<sup>e</sup> siècle les censeurs français n'avaient à contrôler que 200 à 400 ouvrages par an, il suffit aujourd'hui de quelques minutes à n'importe qui pour mettre à disposition sur le Web ses œuvres originales, ses montages, collages, bidouillages, et ce dans tous les domaines du savoir. Et de quelques minutes supplémentaires pour qu'un autre internaute se saisisse de cette création, la personnalise, et la rediffuse par le même canal.

Nous avons une confrontation entre deux types d'entreprises. D'une part nous avons des sociétés commerciales, plus traditionnelles, dont l'activité et les revenus sont principalement basés sur la reconnaissance de leur propriété industrielle, elles vendent des outils qu'elles sont les seules à fabriquer et à réparer. D'autre part nous avons plusieurs entités qui participent de l'activité des Logiciels Libres : des sociétés commerciales (comme IBM par exemple, pour ne citer que la plus importante), des SSSL<sup>163</sup>, des Universités, des projets (comme Debian) et même des particuliers qui agissent à titre tout à fait privé comme chefs ou participants d'un projet. Parmi les arguments des sociétés

---

163. c'est-à-dire des *Sociétés de Services en Logiciels Libres*, nom dérivé de SSII : Sociétés de Services en Ingénierie Informatique, plus traditionnelles

basées sur la propriété industrielle, elles affirment qu'elles sont les seules à disposer des budgets nécessaires à développer et à soutenir l'innovation, en quelque sorte elles seraient les *leaders* et les acteurs du Logiciel Libre ne feraient que suivre, en pillant leur savoir faire.

C'est un argument qui a été souvent battu en brèche par les acteurs des Logiciels Libres, mais pourquoi ne pas citer Pierre Levy qui n'a *a priori* pas encore pris parti pour ou contre les uns ou les autres. Il nous désigne très clairement le problème de l'utilisation des Technologies de l'Information, quand elles sont uniquement sous le contrôle des entreprises qui les commercialisent, dans cette citation il donne son explication de la faillite des expériences « pédagogiques » d'introduction de l'informatique à l'école.

« Une conception totalement erronée de la technique et de ses prétendues « nécessités », auxquelles on a cru (ou fait croire) qu'il fallait « s'adapter », a empêché le gouvernement et la direction de l'Éducation Nationale d'imposer les fortes contraintes aux constructeurs de matériel et aux concepteurs de logiciels. Il n'ont pas été forcés d'inventer. (Lévy, 1990, page 9) »

## 1. L'innovation à travers la méthode scientifique

Les scientifiques parlent de reproductibilité pour confirmer les résultats de leurs expériences et valider leur découvertes. La reproductibilité rend les résultats scientifiques robustes. Un scientifique ne peut espérer justifier de toutes les conditions de test, ni nécessairement disposer de l'environnement de test qui permettrait de vérifier tous les aspects d'une hypothèse. En partageant les hypothèses et les résultats avec une communauté de pairs, le scientifique permet à de nombreuses paires d'yeux de voir ce qu'une seule pourrait manquer. Le partage ouvert des résultats scientifiques facilite

les découvertes. La méthode scientifique minimise la duplication des efforts parce que les pairs savent d'emblée s'ils travaillent sur des projets similaires. Le progrès ne stoppe pas simplement parce qu'un scientifique cesse de travailler sur un projet. Si les résultats ont de la valeur, d'autres scientifiques prendront le relais.

*« Les articles disponibles librement en ligne sont plus souvent cités. Pour un plus grand impact et accélérer les progrès scientifiques, les auteurs et les éditeurs devraient s'attacher à rendre l'accès à la recherche plus facile<sup>164</sup>. (Lawrence, 2001) »*

L'informatique, cependant, diffère fondamentalement de toutes les autres sciences. L'informatique a un seul moyen de permettre à des pairs de reproduire des résultats : partager le code source. Pour démontrer la validité d'un programme à quelqu'un, vous devez lui fournir les moyens de le compiler et de l'exécuter. Pour valider leur « découverte », c'est-à-dire leur création, les programmeurs doivent passer par une phase de tests, aussi appelée déboguage.

Là où les premiers parlent de découverte, les seconds voient création. Le mouvement Open Source est une extension de la méthode scientifique, car au cœur de l'industrie informatique se trouve la science informatique.

Dans le modèle de développement Open Source, le même principe est exprimé ainsi : « à condition de disposer de suffisamment de regards, tous les bogues restent superficiels ». En partageant le code source, les développeurs Open Source rendent leur logiciel plus robuste. Les programmes deviennent utilisés et testés dans une plus grande variété de contextes que ce qu'un seul programmeur pourrait générer, les bugs sont découverts alors qu'autrement

---

164. Articles freely available online are more highly cited. For greater impact and faster scientific progress, authors and publishers should aim to make research easy to access.

ils ne l'auraient pas été. Parce que le code source est fourni, les bugs peuvent souvent être supprimés, et pas seulement découverts, par quelqu'un qui, autrement, serait resté en dehors du processus de développement.

De façon similaire, dans le modèle de développement Open Source, le partage du code facilite la créativité. Les programmeurs travaillant sur des projets complémentaires peuvent appuyer leurs propres résultats sur ceux des autres, ou combiner les différentes ressources dans un seul projet. Un projet peut éveiller l'inspiration pour un autre projet qui n'aurait pas été conçu sans cela. Et un projet de valeur ne devient pas forcément orphelin lorsqu'un programmeur le quitte.

## 8.3 Logiciels Libres et recherche scientifique

### A. Introduction

Nous n'allons pas lancer ici une étude épistémologique qui nous dépasserait et dépasserait le cadre de notre recherche, en revanche nous souhaiterions ouvrir la réflexion sur la relation qui doit, ou ne doit pas, s'établir entre l'objet technique et le chercheur. Nous trouvons souvent chez les chercheurs, comme nous le trouvons chez les techniciens, les ingénieurs ou plus particulièrement les programmeurs, une approche commune de problèmes différents, spécifiques à leurs domaines, mais dont la résolution permet toujours de faire un petit peu plus avancer l'humanité, disons le bien commun pour employer un terme plus sobre.

*« Pour que le progrès technique existe, il faut que chaque époque puisse donner à celle qui la suit le fruit de son effort technique. (Simondon, 1969, page 70) »*

Cette approche se traduit entre autres par la transmission du savoir, qui peut se faire sous différentes formes : l'écriture de thèses, de mémoires ou d'ouvrages, par exemple dans le monde scientifique. Dans le monde technique, la transmission s'est longtemps faite par le mode d'enseignement, basé sur l'apprentissage, l'acquisition d'un tour de main, sa maîtrise puis son amélioration. Pour un programmeur en particulier, une des meilleures manières de maîtriser cet « art » est de lire le code écrit par d'autres et nous revenons ainsi au débat que nous avons ouvert avec les logiciels libres. Pourquoi le cadre légal le plus courant, encadrant le développement des logiciels commerciaux doit-il priver les programmeurs de leur première source de compréhension et d'apprentissage de leur métier : le code source des logiciels que d'autres ont écrit avant eux.

## **B. Une impasse scientifique**

Nous voudrions mettre en avant ce que nous pensons être une impasse scientifique qui n'est que très rarement évoquée. À notre sens, un travail de recherche qui utilise pour traiter le corpus de recherche, ou dont les travaux passent à travers un outil logiciel, pourrait être mis en défaut simplement du fait que ce logiciel n'est pas libre. Si, dans le cadre d'une méthodologie expérimentale de recherche, une hypothèse est validée par la computations de divers paramètres produisant un indicateur, et que le logiciel fonctionne comme une « boîte noire », que devons-nous réellement penser de la validité de cet indicateur ? Le logiciel répondant à un cahier des charges précis, nous pouvons au moins penser que *au moment* de l'expérimentation en laboratoire, les résultats fournis étaient fiables, toutefois, dans ce cadre précis, le chercheur a peu de moyens pour vérifier la fiabilité de son outil logiciel, si ce n'est le soumettre à une batterie de tests, un peu comme on étalonnait une machine

de mesure analogique. Il reste toutefois un problème grave : la reproductibilité de la recherche dans un autre laboratoire. Si le logiciel est fourni par une société locale, comment se le procurer dans une autre région, si c'est une société nationale, dans un autre pays ? Cela voudrait dire que les chercheurs ne devraient utiliser que des outils logiciels fournis par des multinationales pour mener à bien une expérimentation. Reste enfin le problème ultime, celui de la pérennité : si le fournisseur du logiciel disparaît, la recherche devient caduque.

### **C. Le domaine public en danger**

Il est particulièrement rassurant de constater que des hommes ont depuis longtemps régulièrement cherché à collaborer pour construire de grandes choses. Même si les lettres de cachets du Roi ou les secrets de fabrications servaient à protéger la maîtrise, ou plutôt la main-mise d'une personne ou d'une corporation, à l'opposé d'autres personnes travaillaient en commun à construire des cathédrales, des encyclopédies. Bâtir pour Dieu, c'était bâtir pour l'éternité, d'ailleurs les travaux duraient parfois à travers plusieurs générations d'ouvriers d'une ville qui trouvaient dans ce projet une manière de se fédérer, comme cela est d'ailleurs rapporté dans Wikipédia pour l'article « Cathédrale » :

*« Il est difficile aujourd'hui de donner une idée de l'empressement avec lequel les populations urbaines se mirent à élever des cathédrales. La foi avait certes son importance, mais il s'y joignait un instinct très juste d'unité et de constitution civile. »*

Sans doute l'idée de construire pour les générations futures était plus présente à l'esprit de gens qui mourraient tôt, et qui donc par la force des choses léguaient la continuité de leur œuvre à leurs enfants, leurs pairs qui devaient

continuer à faire progresser le projet, avec des ressources qui se renouvelaient et pouvaient remettre en question certains choix. Dans le même article nous relevons d'ailleurs :

*« Rares sont les cathédrales qui aient été finies telles qu'elles avaient été projetées ; et cela se comprend : la période pendant laquelle leur existence est pour ainsi dire un besoin impérieux, l'expression d'un désir irrésistible, est comprise entre les années 1180 et 1240. Soixante ans ! »*

Ce lien à travers les générations s'est perdu. Comment nos grand-parents auraient-ils pu prévoir les bouleversements qu'ils ont traversé, pour ne parler que ceux du mode de vie : tous les logements dotés d'eau courante, d'électricité et de gaz, télévision, téléphone, automobile, avions de ligne, ordinateurs, etc. De plus, la reconnaissance par les pairs n'est plus suffisante pour beaucoup de personnes de nos jours, chacun vise sa propre gloire personnelle, le monde s'individualise et qui chercherait à partager dans un monde où tout porte à croire qu'il n'y aura pas de quoi satisfaire tout le monde. Mais certains restent confiant dans un monde d'abondance et quand Gilbert Simondon mentionne une grande force qui unit les hommes dans un même projet, à travers les frontières et les métiers, il nous semble qu'il est en train de décrire ce qui n'existera que plusieurs décennies plus tard sous la forme des projets qui ont éclos dans le sillage de la *Free Software Foundation*, d'ailleurs jamais autant d'hommes et de femmes de pays différents, de langues différentes et de cultures différentes n'ont été réunis dans un projet tel que GNU/Linux, pour citer le plus célèbre, et donc sans doute le plus vaste.

*« C'est cette force qui était positive et créatrice, et réalisait un aussi remarquable groupement de chercheurs, de rédacteurs, de correspondants, en donnant une foi à cette équipe composée d'hommes qui collaboraient*

*sans être liés entre eux par des communautés sociales ou religieuses ;  
(Simondon, 1969, page 93) »*

Bien entendu, Gilbert Simondon ne pouvait pas prédire l'émergence des Logiciels Libres, ce qui constitue sans doute une autre preuve que cette communion d'effort à travers les siècles est une façon de travailler qui s'est avérée « naturelle » pour les personnes qui avaient l'ambition de rassembler et construire le savoir pour le transmettre, c'est en fait aux encyclopédistes qu'il fait référence. Ils ont joint leurs efforts pour la réalisation de l'*Encyclopédie !Diderot et d'Alembert* de Diderot et d'Alembert, comme aujourd'hui les hackers joignent leurs efforts pour construire un dispositif technique, comme les encyclopédistes de Wikipédia<sup>165</sup> renouvellent la compilation du savoir.

## 1. Propriété Intellectuelle contre domaine public

Comme nous venons de le voir, la confrontation autour de la propriété intellectuelle qui a démarré à propos du code des logiciels s'est petit à petit étendue à tous les domaines de la création humaine. Ce n'est maintenant plus simplement code source disponible contre code source secret, la bataille va bien au delà, puisqu'il s'agit à présent de défendre le domaine public contre l'avancée de la propriété intellectuelle, qui protège plus et plus longtemps.

Le corpus *CoLiSciences* provient d'ouvrages du domaine public, d'une part cela veut dire que nous pouvons les mettre en scène dans *CoLiSciences* de la manière qu'il nous convient, mais d'autre part nous avons un devoir moral de ne pas les utiliser à seule fin de profit. Bien entendu, un tel projet de numérisation coûte et il pourrait être envisageable de tirer un profit de leur mise en ligne pour couvrir ces coûts. Ce débat n'a pas sa place ici, en revanche

---

165. Cette encyclopédie Libre est présentée en détail section 3. page 78

il en existe un autre, parallèle, cette fois lié à la technologie de mise en scène. Si nous rendons obligatoire un outil particulier pour la lecture des ouvrages que nous avons numérisés, nous transportons la question de la propriété intellectuelle du contenu vers la propriété de l'outil pour accéder au contenu. Il nous faut donc utiliser un outil technique qui ne le rende pas prisonnier à nouveau de la propriété intellectuelle.

## 2. Le copyright utilisé contre la communauté

Les travaux de Claude Bernard, de Quatrefages, ... n'appartiennent qu'à eux. Si nous les compilons sur un CD-ROM que nous redistribuons, nous pourrions étiqueter cette nouvelle présentation comme une création personnelle et l'assujettir à un copyright. Est-ce juste de se réapproprier ces ouvrages ? Est-ce juste de ne pas être rétribué pour le travail de traduction sous le nouveau format, de mise en forme, le travail d'édition. Comment rendre le travail de l'éditeur indépendant des technologies qu'il utilise ? Question plus générale, que faire avec les éditions épuisées ? Pérennité contre propriété intellectuelle

## 3. L'ordinateur de confiance ... pour qui ?

TCPA<sup>166</sup> est un projet développé par Intel pour développer « *Une nouvelle plate-forme informatique pour le prochain siècle qui améliorera la confiance dans le monde PC* » comme l'annonce un slogan de cette société. La confiance est plutôt une attitude d'esprit, qui détermine le comportement des individus sur la base d'un sentiment plus que sur un raisonnement ou sur une recherche totale de preuves. Faire confiance, c'est d'une certaine façon, se déterminer spontanément, en faisant l'économie d'une analyse fouillée et rationnelle

---

166. TCPA signifie *Trusted Computing Platform Alliance*, soit Alliance pour une Informatique de Confiance en Français.

(Wikipedia, 2005b). Le dispositif qui est proposé permettra à l'utilisateur d'un ordinateur d'être dans une situation de confiance par rapport à ce qui apparaît sur sa machine, parce qu'il va lui déléguer complètement tout ce qui concerne l'analyse fouillée des évènements qui se produisent. Ce dispositif est présenté comme permettant de rendre la confiance perdue des utilisateurs dans une machine qui est devenue la porte ouverte de nombreux abus. En effet, nous sommes actuellement dans un monde où il ne faut surtout pas ouvrir les messages électroniques qui portent comme sujet « je t'aime », parce qu'ils se révèlent être le plus souvent des pourriels utilisant des sujets évocateurs et amicaux pour nous berner et mieux installer des virus sur nos machines personnelles.

Il est donc temps qu'un peu de confiance soit re-instillée, mais il n'est pas précisé la confiance de qui et il existe de nombreux acteurs du monde numérique qui doivent restaurer leur confiance. Les auteurs de contenus numériques (e-books, musique ou vidéo) qui voient leur travail copié sans jamais toucher un centime, les sociétés de logiciels qui voient elles aussi leurs produits copiés encore une fois sans qu'elles n'en retirent aucun bénéfice et de nombreux autres verraient la nouvelle économie numérique d'un bien meilleur œil si toutes ces possibilités de duplication incontrôlées n'existaient pas. Le TCPA devrait permettre de leur rendre confiance à eux aussi et c'est là que, sans pour autant être un pirate, les intérêts d'un utilisateur sont en opposition avec ceux des professionnels. Bien que le TCPA soit une puce de silicium servant à faire battre le cœur d'un ordinateur, l'enjeu n'est pas du tout technique mais sociétal et culturel. Il ne s'agit pas de comparer les performances entre une technologie et une autre comme savent le faire les ingénieurs au moyen de *benchmarks* mais de choisir entre un modèle de communication et un autre.

Sur cette nouvelle architecture, la société Microsoft prépare un nouveau logiciel, baptisé Palladium, qu'elle déclare vouloir incorporer dans les futures versions de Windows. Il s'installera sur des machines TCPA pour ajouter plusieurs fonctionnalités supplémentaires qui font question parce qu'en voulant renforcer la sécurité, elles se substituent à l'utilisateur de la machine et le dé-responsabilisent, pour lui donner confiance dans ce qui apparaît sur son écran, encore faut-il qu'il aie confiance dans les processus qui s'exécutent, quasiment à son insu, dans sa machine. Pour que utilisateurs et professionnels puissent avoir confiance les uns dans les autres, il faut en effet que ces processus puissent s'exécuter en dehors de tout contrôle humain direct. Ils vérifient absolument tout ce qui s'exécute sur la machine même ce pour lequel l'utilisateur n'a pas de problème de confiance. TCPA et Palladium fournissent une plate-forme informatique sur laquelle l'utilisateur ne peut plus toucher aux logiciels sans que cela ne soit contrôlé et communiqué à l'éditeur avec lequel les logiciels peuvent communiquer de manière sécurisée. Il y a de nombreuses utilisations qui rendent cette architecture très attractive pour les éditeurs :

- La location de films (DVD), de musique en ligne ou autre est enfin totalement sécurisée. Il est possible de payer à la séance, pendant un période, un week-end par exemple.
- Le piratage est impossible. Tout nouveau programme qui s'installe sera vérifié par Palladium, qui pourrait même l'effacer s'il ne trouve pas les droits nécessaires à son utilisation. Il serait donc aussi possible de louer un logiciel, si un mois le loyer n'est pas perçu, le logiciel est bloqué.
- Pour les documents sensibles, il est possible d'imaginer que les documents créés dans une entreprise soient marqués d'un « confidentiel »

électronique qui empêchera leur ouverture sur un ordinateur non reconnu.

Cette liste ne présente au premier abord aucun problème, il s'agit d'utiliser légalement son ordinateur personnel et personne ne devrait rien avoir à redire au fait que sa machine fasse ce type de contrôles. La question est que ces contrôles sont faits à l'insu du propriétaire de la machine : il ne fait pas partie du « réseau de confiance » ! Certain vont même jusqu'à dire que finalement c'est lui que la machine surveille. Cela pourrait n'être qu'un sentiment de malaise, faire peser un soupçon de culpabilité sur tous les utilisateurs d'un ordinateurs qui sont *a priori* pirates et qu'il faut donc surveiller coûte que coûte.

Mais cela pourrait devenir beaucoup plus grave, cette technologie mise entre de mauvaises mains peut conduire à de nombreuses dérives, en particulier à une censure automatique et aveugle qui interdirait tout type de document ne correspondant pas à certains critères ou qui contiendrait un contenu jugé impropre. Pour ceux qui ont installé un filtre parental sur leur ordinateur personnel et qui se sont vus ensuite refusé la lecture de leur quotidien en ligne parce qu'il contient de la violence, il est facile de comprendre que ce type de censure est très aléatoire et dangereux quand il est fait de manière automatique et systématique par un logiciel qui a le dernier mot devant l'humain qui l'utilise. Les autres pourront essayer d'imaginer dans ces conditions comment des journalistes vont pouvoir être alertés de mauvais agissements ? Message électronique d'un employé scrupuleux : bloqué par TCPA. Document : bloqué par TCPA. Dans ces conditions, il ne reste plus qu'à faire une photo de l'écran, mais quelle valeur aura celle-ci devant un tribunal ? Comme les documents pourraient même être dotés d'une date de péremption, ils auront de toutes façons sans doute automatiquement disparu quand la

justice voudra contrôler que ce qui apparaît sur la photo est un document bien réel et demandera à l'intéressé de dévoiler ses documents.

Ces technologies mettent en danger l'héritage inestimable que Gutenberg nous a légué. Les livres électroniques, une fois publiés sous ce format, seront vulnérables. Comme au temps des autodafés un tribunal pourrait ordonner qu'ils soient interdits et l'infrastructure TCPA rendrait cette interdiction incontournable. Du jour au lendemain, en quelques secondes un même ouvrage peut disparaître de tous les écrans d'ordinateurs, où qu'ils soient.

#### 4. Brevets Logiciels

Les brevets logiciels sont un autre terrain où les défenseurs du « domaine public » s'opposent aux défenseurs de la « propriété industrielle », ceci dit caricaturalement de manière à résumer une situation en fait plus complexe. Les logiciels ont traditionnellement été protégés par le droit d'auteur du fait de leur filiation à une filière culturelle : un logiciel est une suite d'algorithme, c'est-à-dire la formalisation mathématique d'un problème, traduit dans un langage de programmation pour pouvoir être exécuté par une machine. Il avait donc été établi que les logiciels étaient construits par des « idées » et ce concept n'étant pas, du moins en Europe, brevetable, la question de la brevetabilité des logiciels ne se posait pas. Mais ce sont aussi des industriels qui construisent des logiciels, pour cela ces industriels ont besoin de faire de la recherche, du développement, d'employer de grandes équipes. Tout cela a un coût et bien entendu tous les moyens pour pérenniser leurs investissements sont bons. Comme la limite entre technique et idée est parfois floue, une brèche s'est ouverte et l'Office Européen des Brevets (OEB) a accepté des brevets portant sur des techniques logicielles, alors qu'*a priori* les logiciels ne pouvaient pas bénéficier de ce type de protection. D'ailleurs, des brevets de

ce type existent aux États-Unis depuis le milieu des années 1980. L'annexe J présente plus en détails les risques associés à l'introduction de brevets de ce type.

## 8.4 Pourquoi des Logiciels Libres dans un projet de numérisation

La suite de la précédente réflexion, nous amène tout naturellement à préciser comment elle se pose dans le cadre plus spécifique de notre travail, plus précisément pour tout ce qui touche à la partie technique du projet de numérisation.

*« L'évolution des objets techniques ne peut devenir progrès que dans la mesure où ces objets techniques sont libres dans leur évolution et non nécessités dans le sens d'une hypertélie fatale. (Simondon, 1969, page 56) »*

Plusieurs approches sont disponibles pour mettre « du contenu » en ligne ; ce qui a motivé nos choix est principalement la pérennité et la fiabilité. Pour assurer la pérennité de nos choix, il est vite apparu que nous devions être indépendants des sociétés d'éditions de logiciel. Typiquement, le problème de tout système informatique est celui de la maintenance et du suivi. Si nous adoptons la solution d'un éditeur, il nous fallait ensuite continuer à suivre ses recommandations en termes de renouvellement de logiciel et risquer de se trouver dans le futur devant soit la disparition de l'éditeur, soit la disparition du logiciel du catalogue de l'éditeur, soit encore face à des coûts de maintenance subitement prohibitifs. Dans ce cas, il nous faudrait être capable de migrer notre solution vers une autre, mais choisir une solution « propriétaire », cela veut souvent dire être enfermé dans cette solution, et

courir le risque de ne pas pouvoir assurer la migration vers un système plus fiable ou moins coûteux.

## **A. Nos objectifs et contraintes**

### **1. Pérennité**

C'est en fait un peu de tout cela, plus un véritable engagement de développer une solution pérenne. Ces ouvrages, sous leur forme papier, étaient prêts à disparaître, pour certains de manière quasi définitive puisqu'ils pourraient être le dernier maillon d'une édition épuisée et, enfermés dans des archives, ils avaient de facto disparu pour le public. Dans notre démarche, nous avons déjà mentionné la part de sauvegarde du patrimoine qui n'est pas anodine. La numérisation a permis de transférer l'image, puis le texte des ouvrages, du support papier fragile vers un support électronique. Nous avons aussi déjà mentionné les différences du problème de la conservation du papier ou de l'électronique. Le papier se fragilise, mais se conserve assez longtemps. Pour l'électronique nous n'avons pas de vision à très long terme du problème, en revanche ce format offre certaines garanties du fait que la duplication ne pose maintenant plus de problème de manipulation d'ouvrages fragiles ou de perte d'information. La seule question qui reste en suspens est le choix du support de stockage et surtout celui du format, ainsi que des dispositifs de lecture associés à ce format, pour permettre une réelle pérennisation de notre travail. Ces textes centenaires méritent la meilleure technologie dont nous disposons aujourd'hui et surtout dont nous disposerons encore dans 10 ans, 50 ans, voire 100 ans.

C'est là tout le paradoxe des NTIC : comment allier technologie de pointe et technologie pérenne ? Si nous regardons les formats des documents Web,

en juste 10 ans, voire moins, de nombreux formats se sont succédé (Flash Macromedia, Cold Fusion, ...) et celui qui est le plus stable (HTML) est loin d'être le plus intéressant du point de vue de ses possibilités en termes d'interactivité et de médiatisation. Il semble au premier abord que pour assurer la pérennité du projet nous aurions dû nous interdire les technologies les plus modernes. Dans un domaine aussi précis que le nôtre, loin des batailles commerciales autour de logiciels « grand public » vendus à des millions d'exemplaires, les outils les plus novateurs sont souvent portés par de petites structures, en général fragiles. Si nous faisons un état des lieux de nos besoins, en termes de technologies, mais en essayant de garder du recul par rapport aux logiciels et en restant au niveau de fonctionnalités, ou des services que nous avons besoin de mettre à la disposition de notre projet, nous trouvons :

- Un serveur « Web ».
- Un gestionnaire de bases de données.
- Un format de stockage pour les textes numérisés.
- Un gestionnaire de contenu pour mettre à disposition des paratextes, glossaires et héberger le moteur hypertextuel.

Nous avons déjà abordé le sujet du format de stockage de données dans la première partie, il est donc inutile d'y revenir ici, si ce n'est pour rappeler que XML participe aussi de la mouvance « Open Source ». Le cas de XML est, il est vrai, un peu particulier puisque c'est une société commerciale, Microsoft en l'occurrence, qui a peut-être commencé à imposer son utilisation sur la toile. Dans leur approche XML, n'est pas tant un format de stockage que plutôt un format d'échange, c'est-à-dire que la vision Microsoft de l'utilisation d'XML est de permettre à des logiciels différents de sociétés différentes de parler un langage commun. XML serait dans ce cas une sorte d'espéranto des différents

serveurs de solutions « e-Business », chaque solution assurant le stockage des données sous un format propre (propriétaire serait plus juste). Ceci est vraiment très éloigné de nos préoccupations et il faut revenir aux travaux du W3C pour retrouver le standard ouvert, hérité de SGML, tel que nous l'avons décrit dans la première partie. Cette parenthèse pourrait paraître inutile et hors propos, mais elle témoigne de l'âpreté des tiraillements entre standards ouverts et solutions propriétaires, tiraillements dont il faut se préserver, autant que faire se peut.

Dans le chapitre traitant des ouvrages numériques, nous avons présenté *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* sous ses différents aspects, correspondants aux différents buts que nous souhaitons atteindre dans ce projet. La pérennisation d'un patrimoine culturel précis, sa mise à disposition d'un large public ont donc déjà été discutés dans ce cadre. Nous avons aussi évoqué les nouvelles contraintes apportées par la technologie : il est plus facile de lire un ouvrage papier du XIV<sup>ième</sup> siècle qu'une disquette des années 80. Il nous fallait donc éviter l'écueil de la mise à disposition du patrimoine visé par *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* au travers d'une technologie appelée à disparaître rapidement, disons à l'horizon d'une décennie.

Nous ne sommes bien entendu pas le seul projet à se poser ce type de question, les entreprises qui utilisent souvent intensivement des systèmes d'information savent combien il est important d'utiliser un système pérenne, mis à jour régulièrement, mais ces entreprises sont dans un cadre commercial traditionnel et elles jouent jeu égal avec les distributeurs de matériels et de logiciels. La pérennité, en transférant à l'utilisateur le droit de modifier les logiciels libres dont il a fait l'acquisition fait partie des problèmes qui peuvent être résolus par l'utilisation de logiciels libres.

Disposer du code source de tous les logiciels qui composent notre solution, même si en pratique cela semblerait difficile à mettre en œuvre, devrait nous permettre de les faire évoluer dans le temps en dehors de toute contrainte extérieure. Notre système *CoLiSciences* reposant sur les technologies « LAMP » (*Linux Apache MySql Php*), il est assez facilement portable, du moment qu'un serveur Web assorti du langage de script *Php* et d'une Base de Données *mySQL* est disponible. Effectivement, en fonction des besoins, nous avons utilisé ce système, en développement ou en production sur des plateformes *Red Hat*, *Mandrake*, *SUN Os* et même *Windows 2000* et *Windows XP* sans que cela ne pose de réel problème de migration

## 2. Fiabilité

Il est difficile de parler de fiabilité sans se lancer dans de longues descriptions et comparatifs techniques. Telle solution est réputée plus fiable parce qu'elle a obtenu telle labélisation, telle autre parce que des enquêtes auprès des utilisateurs montrent sa robustesse. Pour ce qui nous concerne, la solution *LAMP* mentionnée ci-dessus étant celle qui équipe une très grande partie de sites Web, comme le montre la figure 8.1 discutée au point suivant, elle semble répondre *de facto* au deuxième critère. De plus, la disponibilité du code source semblerait permettre à plus de personnes de contrôler ses éventuelles erreurs et donnerait un avantage aux solutions à bases de Logiciels Libres. En tout état de cause, le fait de disposer du code source rend le projet indépendant d'une tierce partie en cas de problème, puisque nous avons tous les éléments techniques à notre disposition, ce qui n'empêcherait pas toutefois de poser un problème de compétence, il faut donc savoir rester modeste et s'assurer que les logiciels utilisés répondent à un certain nombre de critères : fiabilité reconnue

et vérifiée, disponibilité d'un réseau de développeurs actifs et de moyens de communiquer avec eux.

### 3. Contraintes Techniques et Logicielles

Nous l'avons déjà présenté sous différents aspects, les logiciels libres couvrent une large gamme d'activités et ce sur de nombreux matériels différents, permettant de mettre à disposition une grande capacité de traitement, de manière relativement flexible. Mais c'est sans doute sur les outils développés spécifiquement pour Internet que les logiciels libres montrent le plus leur prédominance. Les serveurs Web sont le domaine où GNU/Linux s'est le mieux imposé du fait des liens étroits qui lient Internet à Unix et finalement à GNU/Linux par simple évolution entre architectures équivalentes. Depuis 1995 Netcraft, une société anglaise, tient à jour des statistiques sur le nombre de sites Web et les logiciels qu'ils utilisent. En juillet 2005 le graphique<sup>167</sup> représenté sur la figure 8.1 montre très clairement l'hégémonie du serveur Apache, développé sous une licence de Logiciel Libre et installé principalement sur des architectures de type GNU/Linux par rapport aux serveurs PWS ou IIS qui sont utilisés sur l'architecture MS Windows de Microsoft.

Les statistiques d'utilisation et leur historique disponible en ligne mettent aussi en évidence que la tendance a continué à s'affirmer puisque les serveurs Apache sont passés d'un peu plus de 60% des serveurs en 2002, jusqu'à atteindre près de 70% en 2004, niveau qu'ils continuent à chercher à atteindre en 2005, l'approchant de plus en plus près. En avril 2004 sur 49.750.568 sites interrogés, 67,2% soit 32.280.582 sites utilisent Apache et 21,02% soit 10.099.760 sites utilisent Microsoft IIS, en mai 2005 ces chiffres sont passés à 44.072.262 (69,37%) pour Apache et 13.049.346 (20,54%) pour IIS, en juillet

---

167. Ce graphique est mis à jour tous les mois et disponible en ligne à l'adresse [http://news.netcraft.com/archives/web\\_server\\_survey.html](http://news.netcraft.com/archives/web_server_survey.html)



FIG. 8.1: Les parts de marché Apache vs Microsoft

nous sommes à 47.030.635 (69,60%) contre 13.871.645 (20,53%) après une légère baisse en pourcentage par rapport à juin. En valeur absolue Apache « gagne » tout de même 1.857.740 nouveaux serveurs alors que IIS doit se contenter de 740.284. Cette enquête repose en juillet 2005 sur un total de 67.571.581 sites.

#### 4. Contraintes Budgétaires

Nous n’allons pas ici développer le TCO (Total Cost of Ownership), c’est-à-dire le coût d’appropriation d’une solution logicielle, si cher aux magazines spécialisés de l’industrie informatique. Il est bon toutefois de rappeler que le coût d’un logiciel n’est pas simplement son coût d’achat, il faut ajouter à cela le coût que cela peut générer sur de nouveaux achats de matériel, le coût de la formation du personnel ou de l’embauche de nouveaux spécialistes. Les solutions à base de Logiciels Libres ont déjà le premier avantage d’un faible coût, voire pas de coût des licences logicielles; à cela s’ajoute la disponibilité de personnes compétentes, la facilité de trouver des matériels capables de supporter ces architectures et même la possibilité de faire appel

à des sociétés tierces pour l'hébergement sans avoir de besoins spécifiques ou de personnalisations qui rendraient ce type de solution hors d'atteinte.

## B. Les composants logiciels du projet

### 1. Le logiciel système

**GNU/Linux** GNU/Linux est un système d'exploitation, ou logiciel de base.

*« C'est le logiciel ou plutôt l'ensemble des logiciels sur lesquels vont s'appuyer tous les autres programmes pour pouvoir s'exécuter sur un type de matériel particulier (Augier, 1997, page 21) »*

Il est possible de distinguer deux catégories de logiciels sous cette dénomination : la première est le noyau, c'est-à-dire la partie la plus proche du matériel et qui permet d'en donner une vue logique aux autres programmes. La deuxième est l'ensemble des outils, cette fois plus génériques et s'appuyant sur le noyau, qui permettent de voir les fichiers, d'en créer, de compiler des programmes, etc...

GNU/Linux possède de nombreuses particularités :

- C'est une variante d'Unix. Il possède la même architecture, un grand nombre de commandes en commun et la plupart des logiciels développés sur Unix peuvent être utilisés sur GNU/Linux.
- C'est un logiciel libre, le code source est disponible et modifiable.
- C'est la réunion de deux projets indépendants, GNU et Linux. Chacun isolé n'aurait pu aboutir à fournir un système utilisable avant plusieurs années.

**Le projet GNU** Nous avons déjà présenté le projet GNU dans l'introduction aux logiciels libres. La production de ce projet est un très

grand nombre d'outils systèmes et un noyau, appelé HURD qui n'est pas encore réellement opérationnel. En revanche les outils systèmes de GNU peuvent être utilisés avec un noyau Linux, qui est un autre projet de développement libre, à l'origine indépendant de GNU. La rencontre des deux a toutefois permis la création de systèmes GNU/Linux, parfaitement opérationnels, un peu à la manière d'un écosystème technique. Linux n'aurait pu être construit sans les outils GNU, GNU ne pourrait être utilisé sans le noyau Linux, difficile de comprendre comment les deux existent, comment ils ont pu se développer de manière indépendante, et chose encore plus surprenante, comment font-ils pour continuer à être développés de manière indépendante.

**Le noyau Linux** L'origine du noyau Linux est un message posté en 1991 par Linus Torvald, alors étudiant en Finlande, dans le groupe de discussion<sup>168</sup> `comp.os.minix`<sup>169</sup> :

```
Hello everybody out there using minix -  
I'm doing a (free) operating system (just a hobby,  
won't be big and professional like gnu)  
for 386(486) AT clones.
```

FIG. 8.2: Tout est parti d'un message de Linus

Un peu plus d'une dizaine d'années plus tard, le *big and professional* GNU utilise le noyau Linux pour s'exécuter, la « communauté Linux » compte 4 millions de programmeurs, des dizaines de sociétés commerciales dont l'activité fourniture de logiciels ou services est directement basée sur

---

168. Un groupe de discussion sur Internet est un forum ouvert à tous et traitant d'un sujet particulier, identifiable à la manière dont est formé le nom du forum.

169. En 1985 un professeur hollandais, nommé Andrew Tannenbaum, avait développé un système d'exploitation minimal, baptisé Minix, afin d'enseigner la programmation système à ses étudiants (Tannenbaum et Woodhull, 1987).

Linux et « Linux » lui même est décliné en dizaines de version différentes, appelées distributions. Celles-ci sont plus ou moins spécifiques à des secteurs d'activité, des pays ou des objectifs de performances différents.

Nous pouvons relever à travers quelques distributions cette grande diversité et les différents objectifs poursuivis :

- Debian<sup>170</sup> La version la plus respectueuse de la « philosophie » GNU et Logiciels Libres.
- Mandrake<sup>171</sup> Distribution Française, très orientée utilisateur final, bureautique.
- Red Hat<sup>172</sup> Distribution US, très orientée entreprise, le modèle et le projet commerciaux sont très aboutis.
- SuSE<sup>173</sup> Distribution Allemande, rachetée par la société Novell, plutôt positionnée sur le marché des serveurs.
- Red Flag<sup>174</sup> Distribution Chinoise, avec un double clin d'oeil dans le nom.
- Morphix-NLP<sup>175</sup> Cette distribution est l'œuvre d'un étudiant chinois, Zhang Le qui souhaitait faciliter la mise à disposition d'outils linguistiques basés sur la plate-forme Linux. Les deux particularités de cette distribution sont donc, d'une part le fait de pouvoir l'utiliser directement depuis le CD où elle est gravée, sans l'installer sur le disque dur de l'ordinateur et d'être orientée outils linguistiques (NLP *Natural Language Processing*, signifie traitement du langage naturel), elle permet donc à des non-informaticiens de tester différents outils

---

170. <http://www.debian.org/>

171. <http://frontal2.mandriva.com/>

172. <http://www.redhat.com/>

173. <http://www.suse.de/en/>

174. <http://www.redflag-linux.com/eindex.html>

175. <http://www.nlplab.cn/zhangle/morphix-nlp/>

sans avoir à les installer, donc sans avoir de compétences informatiques particulières.

**Apache** Ce projet de développement Libre d'un serveur Web est l'héritier des premiers développements de ce type au sein du *National Center for Supercomputing Applications*, à la création du Web. Nous avons déjà évoqué son très large taux d'utilisation, près de 70%, quand nous évoquons les problèmes liés aux brevets logiciels.

**MySQL** Serveur de bases de données libre. Utilisé pour stocker de nombreuses informations telles que le Glossaire, la Bibliographie, etc.

**PHP** Langage de scripts. C'est le langage utilisé pour coder les programmes qui animent et génèrent à la volée les pages affichées par *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*.

## 2. Les logiciels applicatifs

**phpNuke** C'est le portail (ou *Content Management System, CMS*) que nous avons utilisé et personnalisé pour héberger les modules développés spécifiquement pour animer le site *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*. Le moteur hypertextuel en particulier peut être considéré comme une brique qui s'inscrit dans l'architecture phpNuke, ce qui nous a évité de développer la plupart des fonction d'administration, de gestion de droits d'accès et autres problèmes qui sont depuis longtemps réglés pour nous concentrer sur la spécificité de *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*.

**CVS** Cet outil (*Concurrent Versioning system*) permet de gérer à distance les modifications faites par différentes personnes sur les mêmes documents. Cela permet de garder la trace des modifications et de savoir « qui a fait quoi ».

## C. Méthode de développement

La place de l'ingénieur est quelquefois difficile à trouver dans un projet patrimonial et culturel. Est-ce qu'il doit simplement faire fonctionner les outils comme le cheminot conduit la locomotive, sans pouvoir intervenir sur les aiguillages ? Est-ce qu'il est plutôt le pilote automobile qui doit trouver la bonne trajectoire ? Une citation de Gilles Simondon éclaire d'un jour nouveau ce rôle *a priori* technique, de plus il fait un parallèle entre le métier de l'ingénieur et celui de l'artisan qui prend une importance particulière dans le contexte de la programmation, qui est souvent un travail très artisanal. Un long polissage du code, essai après essai, permet de corriger petit à petit les imperfections pour atteindre au final un objet logiciel fini et agréable à voir comme à faire fonctionner.

*« L'artisan et l'ingénieur ne vivent pas seulement pour eux-mêmes ; témoins et agents de la relation entre la société humaine dans son ensemble et le monde des objets techniques dans son ensemble, ils ont une valeur exemplaire : c'est par eux que l'objet technique s'incorpore à la culture. (Simondon, 1969, page 85) »*

D'après cette citation, l'ingénieur n'est pas un acteur isolé mais une sorte de médiateur entre l'objet technique et la société humaine. Son rôle n'est donc plus uniquement de construire mais de communiquer, d'expliquer et d'adapter la technique aux besoins exprimés par d'autres qui n'ont pas accès à la compétence technique, mais qui doivent en faire usage dans le cadre de leur propre savoir faire. Ceci rejoint effectivement les préoccupations de nombreux techniciens, preuve en est la méthodologie de conception que nous avons utilisée pour le développement de *C<sub>o</sub>L<sup>i</sup>Sciences*.

## 1. Une méthode pour formaliser les interactions

Dans une équipe multi-compétences il faut arriver à développer l'écoute et trouver un langage commun qui permette d'additionner les compétences de chacun plutôt que de se contenter de leur intersection, qui elle s'amenuise avec la spécialisation et l'expertise de ses membres. Ceci est d'autant plus vrai pour tout ce qui est lié au développement de l'architecture technique du projet. Cette réflexion rejoint celle autour de la problématique de développement d'un système d'information dans laquelle de nombreux intervenants, aux compétences très diverses doivent travailler ensemble, les uns pour décrire leur besoin, d'autres pour les formaliser et enfin les derniers pour développer une solution basée sur cette formalisation.

Un certain nombre de méthodologies répondent à ce type d'attentes, elles reposent sur la notion d'agilité telle qu'elle est décrite dans la spécification des méthodes *Agile*<sup>176</sup>. Ces méthodes sont orientées clients, elles visent à répondre le plus rapidement à leurs attentes tout en proposant la meilleure solution. Il s'agit là d'une réponse aux difficultés rencontrées généralement par les projets de création et de développement de systèmes d'information. Ceux-ci sont généralement encadrés par des méthodologies lourdes et contraignantes, générant de nombreuses contraintes, souvent d'ordre administratif ou contractuel, ce qui a uniquement un impact négatif sur les ressources allouées au projet en particulier, ceci ayant pour effet de les éloigner de l'objectif premier du projet de produire une solution pour satisfaire un besoin client. Un rapport du Standish Group<sup>177</sup> (Standish Group, 1995) a étudié ce

---

176. Un site de référence se trouve à l'adresse : <http://www.agilemodeling.com/> mais de nombreuses autres références officielles existent aussi comme l'alliance Agile qui est présente en ligne à l'adresse : <http://www.agilealliance.com/> et aussi un manifeste très intéressant à l'adresse : <http://agilemanifesto.org/>

177. Cet organisme est représenté à l'adresse <http://www.standishgroup.com/>

phénomène et donne quelques chiffres pour mettre en évidence l'importance de bien cadrer les projets de développement de système d'information :

- Près d'un tiers des projets sont abandonnés avant la mise en place.
- Plus de la moitié dépassent leur budget.
- Moins de 2 sur 10 se terminent dans les temps et dans le budget alloué.

Parmi les nombreuses raisons invoquées, l'échec de ces projets est souvent le résultat d'un manque de communication entre les différentes parties prenantes, en particulier parce que le niveau d'implication de l'utilisateur final est régulièrement minime. Ce manque de communication engendre une confusion entre les besoins estimés et les besoins réels.

## 2. eXtreme Programming

L'eXtreme Programming<sup>178</sup> est une nouvelle méthodologie de programmation itérative et incrémentale, inventée par Kent Beck et Ward Cunningham officiellement en octobre 2000 (Beck, 2000). Elle est souvent mise en avant dans les projets de développement de plateformes Web, qui demandent une forte implication des utilisateurs mais aussi pour théoriser la façon d'avancer des communautés de programmeurs « Logiciels Libres ». La littérature technique y fait souvent référence sous l'acronyme « XP »<sup>179</sup> qui signifie donc **eXtreme Programming**.

Cette méthodologie s'applique tout particulièrement aux projets dont les frontières ne sont pas très précises et pour lesquels le cahier des charges évolue souvent avec la construction du projet même. C'est tout précisément le mode de fonctionnement d'un projet de recherche action, de nos premières

---

178. La méthode est décrite sur de nombreux sites, à commencer par l'officiel <http://www.extremeprogramming.org> en anglais et son équivalent en français qui se trouve à l'adresse <http://xp-france.net/>

179. Le système d'exploitation Windows XP de la société MicroSoft, ne tire pas son nom de la même « étymologie » mais du mot « eXPérience » ; qui copie qui ?

expérimentations nous construisons les étapes suivantes, chaque étape étant régulièrement remise en cause par les résultats des étapes qui l'ont précédé.

Dans ce type de projet, l'équipe de développement n'est pas uniquement constituée de programmeurs, toutes les parties prenantes font partie de cette méta-équipe : responsables comme utilisateurs, utilisateurs « savants », administrateurs et novices, tous participent par leurs interactions à l'avancement du projet de développement « XP ».

*« La nature humaine, chez le paysan, le berger, le mineur, le marin, se double ainsi d'une seconde nature qui est comme un pacte ancestral avec un élément ou une contrée. (Simondon, 1969, page 90) »*

C'est sans doute cette même seconde nature que nous retrouvons chez le hacker, le pacte n'est peut-être pas ancestral mais il cherche un équilibre, une symbiose, entre le code qu'il écrit et le code que d'autres écrivent.

L'eXtreme Programming est autant une réflexion sur le rôle de l'ingénieur et sa fonction qu'une méthodologie structurée détaillant les étapes nécessaires au bon déroulement d'un projet. Pour cette raison, cette méthode est construite autour de 4 valeurs fondamentales :

- « – *La communication : c'est le moyen fondamental d'éviter les erreurs. Le moyen à privilégier est la conversation directe, en face à face. Les moyens écrits ne sont que des supports et des moyens de mémorisation ;*
- *Le courage : il est nécessaire à tous les niveaux et de la part de tous les intervenants, notamment chez les développeurs (quand des changements surviennent à un stade avancé du projet, ou quand des défauts apparaissent) et chez le client (qui doit pouvoir prioriser ses demandes) ;*

- *Le retour d'information (feedback) : les itérations sont basées sur les retours d'informations du client, permettant une adéquation totale entre l'application finale et sa demande ;*
- *La simplicité : en Extreme Programming, la façon la plus simple d'arriver à un résultat est la meilleure. Prévoir préalablement des évolutions futures n'a pas de sens. Ce principe est résumé en une phrase : « Tu n'en auras pas besoin. » (en anglais « You ain't gonna need it. »). La meilleure manière de rendre une application extensible est de la rendre aussi simple (et donc simple à comprendre) et bien conçue que possible.*

*(Wikipedia, 2005c) »*

Ces valeurs sont déclinées en 13 pratiques, que nous nommerons sans les détailler pour ne pas rendre ce passage trop long et trop technique : Client sur le Site (*On-Site Customer*), Séance de Planification (*Planning game*), Intégration continue (*Continuous Integration*), Livraisons fréquentes (*Frequent Releases*), Rythme soutenable (*Forty-hour Week*), Tests de recette (*Acceptance Tests*), Tests unitaires (*Unit Tests*), Conception simple (*Simple Design*), Utilisation de métaphores (*Metaphor*), Remaniement du code continu (*Merciless Refactoring*), Convention de nommage (*Coding Standard*), Programmation en binôme (*Pair Programming*), Propriété collective du code (*Collective Code Ownership*).

Pour terminer de manière synthétique sur cette méthode en insistant sur son aspect itératif, en opposition aux méthodes traditionnelles par phases, le plus simple semble de présenter un schéma résumant le cycle de vie du projet de façon graphique. Sur la figure 8.3 les différents chemins sont mis en évidence, et nous pouvons voir comment s'articulent les itérations, entre phase de planification et tests de recette.

8.4. Pourquoi des Logiciels Libres dans un projet de numérisation

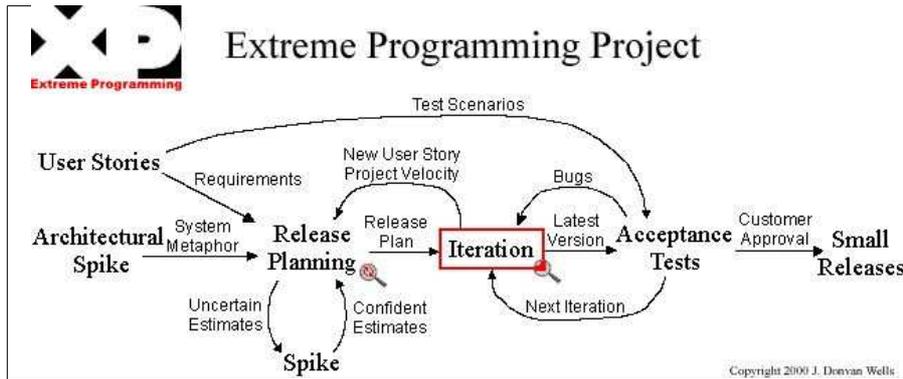


FIG. 8.3: Le projet eXtreme Programming

Les itérations sont elles-mêmes représentées plus précisément sur la figure 8.4 suivante.

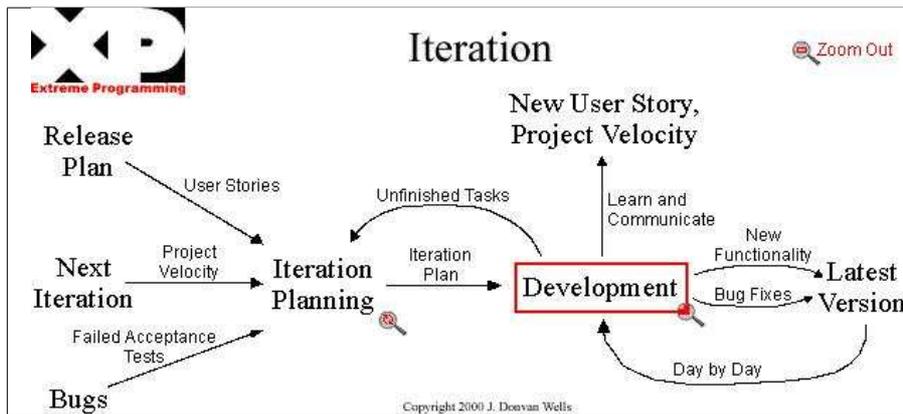


FIG. 8.4: Les itérations du projet eXtreme Programming



CHAPITRE 9

**LE PROCESSUS DE CONCRÉTISATION D'UNE  
BIBLIOTHÈQUE NUMÉRIQUE**

---

## Sommaire

---

<b>9.1</b>	<b>Introduction</b> . . . . .	<b>439</b>
<b>9.2</b>	<b>Bibliothèque</b> . . . . . <b>et</b> . . . . . <b>réseau...</b>	
	<b>... ou réseau de bibliothèque.</b> . . . . .	<b>440</b>
	A. Fédérer un réseau de bibliothèques . . . . .	442
	B. Quel lien avec les bibliothèques traditionnelles? . . . . .	447
<b>9.3</b>	<b>Bibliothèque</b> . . . . . <b>électronique...</b>	
	<b>... ou système d'information.</b> . . . . .	<b>448</b>
	A. Les outils . . . . .	448
	B. La relation outil et utilisation . . . . .	451
	C. Quelle bibliothèque? . . . . .	452
	D. Le Web Sémantique . . . . .	453
<b>9.4</b>	<b>Bibliothèque et société</b> . . . . .	<b>454</b>
	A. Les aspects économiques . . . . .	455
	B. Les aspects juridiques . . . . .	458

---

## 9.1 Introduction

Une des particularités, et sans doute pas des moindres, de la bibliothèque numérique est de pouvoir l'aborder sous la forme d'un objet technique.

*« L'objet technique existe donc comme type spécifique obtenu au terme d'une série convergente. Cette série va du mode abstrait au mode concret : elle tend vers un état qui ferait de l'être technique un système entièrement cohérent avec lui-même, entièrement unifié. (Simondon, 1969, page 23) »*

Depuis la composition des premiers textes sous format numérique, nous avons vu différents dispositifs mis au point pour les lire, pour les rassembler, puis pour les mettre à disposition de tous à travers le réseau Internet et son interface hypertextuelle, le Web. L'évolution pressentie du Web sémantique va vers effectivement l'unification de toutes ces ressources en rendant les systèmes communicants et capables de coopérer entre eux. Nous sommes toutefois encore très loin ce modèle et chacun des intervenants cherche pour l'instant à développer le standard qui fera de lui l'acteur principal, celui qui fixe les règles du jeu de l'édition numérique. A l'inverse du livre papier qui s'est développé après l'invention de l'imprimerie, nous pourrions dire que le livre électronique existe déjà, mais il cherche son Gutenberg qui industrialisera et unifiera les technologies. Nous sommes encore dans une phase de mise en place et de convergence.

## 9.2 Bibliothèque et réseau...

### ... ou réseau de bibliothèque.

Réfléchir sur la genèse du projet *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, sur les outils qu'il a fallu rassembler et ceux qu'il a fallu construire, devrait permettre de dégager les grands lignes de ce qu'un tel dispositif doit comporter comme éléments et comment leur mise en œuvre doit être pensée. Ce que Simondon décrivait, adapté à un monde plutôt mécanique, s'adapte à notre monde cybernétique de façon presque naturelle.

*« La machine qui est douée d'une haute technicité [...] suppose l'homme comme organisateur permanent, comme interprète vivant des machines les unes par rapport aux autres. (Simondon, 1969, page 11) »*

La bibliothèque numérique, « machine » ou système doué d'une haute technicité demande un grand travail d'organisation de tous les logiciels les uns par rapport aux autres. Il faut prendre garde aux limites endogènes et exogènes que ces dispositifs logiciels portent en eux comme nous le fait remarquer Bernard Stiegler dans (Stiegler, 1994). Nous pouvons par exemple citer comme limite endogène, qui posait véritablement question au début de notre projet au point de remettre en question toute l'architecture, le support des documents XML par PHP, le langage de script de notre moteur hypertextuel. Ce support en était à ses débuts, il n'était pas stabilisé et plusieurs voies étaient possibles sans que l'une d'entre elles ne soit vraiment fiabilisée. Nous essayions nos propres outils tandis que les développeurs des différents support XML de PHP faisaient eux aussi leur chemin. La convergence de nos projets a permis finalement de trouver le support qui nous convenait, au moment où nous en avons eu besoin. Cette limite est tombée au bon moment peut-être parce que nous avons choisi la bonne technologie, qui était en train d'être mise en place

en même temps que notre propre projet se mettait en place, mais comment tirer un enseignement de ce type de contingence, qui est tombée d'elle-même sans que nous n'ayons vraiment eu à intervenir ?

*« Dans un contexte d'innovation par exemple, on ne peut pas exprimer des besoins par rapport à des usages qu'on ne connaît pas encore. L'organisation doit être capable de produire les connaissances et de mobiliser l'intelligence de manière permanente, de façon à pouvoir formuler des réponses inédites à des problèmes nouveaux et imprévisibles. Dans ce contexte de production de connaissance face à des possibles indéfinissables a priori, la primauté revient à l'intelligence dans l'action et non aux connaissances génériques. (Prax, 1998, page 3) »*

Cherchons plus loin : quand les dispositifs de sécurité mis en place dans PHP ont été durcis, c'est toute l'architecture qui est remise en cause. D'une part le moteur phpNuke avait dû être refondu pour supporter ces nouveaux critères de sécurité et d'autre part, le niveau de personnalisation de phpNuke dans *C<sup>o</sup>L<sup>i</sup>sciences* rendait très coûteux de repartir sur une intégration dans la nouvelle version mise à jour de phpNuke.

*« Il y a d'une part le progrès comme développement des conséquences d'une invention technologique à l'intérieur d'un système technique stable, sans crise nécessaire, sans discontinuité brutale, ce que Gille appelle des « lignes technologiques », et d'autre part, le progrès comme déstabilisation du système technique, reconstitution autour d'un nouveau point d'équilibre, et naissance d'un nouveau système technique. Les nouveaux systèmes techniques naissent de l'apparition des limites des précédents, où le progrès est essentiellement discontinu. (Stiegler, 1994, page 47) »*

## A. Fédérer un réseau de bibliothèques

Il existe plusieurs projets transversaux de mise en ligne de grands corpus de textes. Nous pouvons citer en particulier GUTenberg et WikiBooks<sup>180</sup>.

Que faut-il réellement penser de ce type de projet ?

Quels sont les enjeux affirmés, et ceux moins avouables qui peuvent rester cachés derrière ce type de projet ?

### 1. Faut-il le faire ou laisser Google le faire pour nous ?

Le samedi 19 mars 2005 un article du journal Libération revenait sur l'annonce du moteur de recherche Google qui évaluait à 10 dollars seulement la mise en ligne d'un titre. Le chiffre paraît exagérément peu élevé, il faut bien entendu lire entre les lignes et interpréter : ces 10 dollars sont exclusivement utilisés pour la numérisation, il ne tiennent pas du tout compte du prix de l'ouvrage lui-même. L'astuce consiste à ne numériser que des ouvrages libres de droits d'auteur, mais alors cela veut dire qu'ils sont tombés dans le domaine public, donc vieux d'au moins 70 ans.

C'est sur cet aspect que revient plus précisément l'article :

*« Le principal obstacle à la numérisation de textes en ligne est le droit d'auteur. On ne connaît pas aujourd'hui la liste détaillée des livres proposés par les bibliothèques avec lesquelles Google a signé un partenariat, mais il projette de numériser les 8 millions de livres de la bibliothèque de l'université de Stanford, les 7 millions de volumes de celle du Michigan et 40 000 de celle de Harvard. La Bodleian Library de l'université d'Oxford, en Angleterre, est plus prudente puisqu'elle a annoncé ne confier à Google que les ouvrages publiés avant 1900, tombés dans le domaine public, de*

---

180. WikiBooks est un projet adossé à l'encyclopédie libre Wikipédia disponible à l'adresse <http://fr.wikibooks.org/wiki/Accueil>

---

*même que la New York Public Library. Aux Etats-Unis, les livres passent dans le domaine public soixante-dix ans après la mort de l'auteur. Les internautes n'auront probablement, dans un premier temps, qu'accès à des livres libres de droits. Il existe une autre perspective qui est de fournir pour les ouvrages encore sous droits uniquement des extraits, de manière à attirer les internautes et leur donner envie de lire l'ouvrage et donc de l'acheter. Une telle imbrication entre intérêts commerciaux et projet culturel ne peut qu'éveiller les soupçons, les ouvrages qui seront proposés sous cette forme seront-ils triés, filtrés et de quelle manière ? La pseudo démocratie qui pousse à mettre en ligne « ce que les gens demandent » a déjà modifié notre paysage audiovisuel en ne proposant que ce qui est le plus facile d'accès et qui pousse à consommer. »*

L'annonce originale faite par Google date du mois de décembre 2004 et présentait un projet de numérisation qui viserait à mettre des ouvrages des bibliothèques publiques et de plusieurs universités américaines. La proposition que Google a fait à ces bibliothèques est une sorte de marché gagnant-gagnant : Google paye la numérisation du fonds et offre en retour une copie numérique que la bibliothèque pourra utiliser comme bon lui semble. L'objectif sur lequel Google communique est de créer une « Bibliothèque Numérique Mondiale », ce qui offre pour eux plusieurs intérêts :

- Un effet d'annonce, les chiffres présentés font tourner la tête : 15.000.000 ouvrages, seulement 10\$ par ouvrage. Peu de sociétés sont capables d'une telle prouesse. D'autre part Google est depuis quelque temps<sup>181</sup> le *leader* incontesté des moteurs de recherche. Leur technologie leur laissant un peu de répit, il est temps pour eux de chercher à conquérir de nouveaux territoires. Leur domaine au sens large est le catalogage

---

181. Notion difficile à cadrer dans le milieu des Technologies de l'information, l'obsolescence guette les technologies après seulement quelques mois d'existence.

- de l'information, quand elle est sous forme numérique, il était presque naturel qu'après avoir fait la preuve de leur capacité à numériser *toute* la toile, il ne leur reste plus qu'à prouver qu'ils sont capables de numériser aussi les bibliothèques traditionnelles. Ce n'est finalement qu'encore une autre forme de poursuite du rêve de Ted Nelson avec Xanadu, de Paul Otlet avec le Mundaneum : construire une bibliothèque mondiale de tous les livres, de tous les savoirs<sup>182</sup>.
- Se positionner comme un acteur incontournable du Web. Il y a quelques années, pour beaucoup d'internautes, la « porte d'entrée » naturelle du Web était le portail d'information. Sur cette ou ces quelques pages, il était possible de trouver le lien de départ d'un voyage virtuel sur les sentiers de l'information. Mais le ou plutôt les chemins sont devenus trop complexes, trop long. Suivre des liens ne suffit plus, il faut maintenant voler, ou sauter, à travers le cyber-espace directement vers la page cherchée. C'est là que la technologie Google est venue à notre aide et que le moteur est devenu une espèce de « méta-portail », qui se reconfigure au gré des requêtes de recherche. C'est devenu la page de référence depuis laquelle tous les voyages sont lancés à travers la toile, devant Yahoo ! qui avait occupé cette place auparavant, et même devant la société Microsoft, qui n'a pas encore réussi à battre Google, ni sur le terrain des moteurs de recherche avec MSN Search, ni sur le terrain économique puisque les bruits de rapprochement et de rachat n'ont jamais abouti.

Bien entendu, les poids lourds d'Internet ne devraient pas tarder à se mettre sur les rangs pour essayer de tirer eux-aussi des bénéfices et de marquer leur territoire. Nous devrions donc voir Yahoo et Microsoft/MSN

---

182. Nous avons évoqué ces projets précédemment dans le cadre du chapitre traitant de l'hypertexte, page 67 et celui de la représentation des parcours de lecture, débutant page 145 dans le cadre de la topologie des sites Web.

---

réagir par des partenariats similaires, ce qui ne fera qu'engendrer confusion et multiplication des technologies en concurrence.

Pour l'instant les ambitions du projet Google sont immenses et ses chiffres laissent perplexes, il est prévu de mettre en ligne 15 millions d'ouvrages, soit plus de 4,5 milliards de pages, et ceci dans un délai relativement court de 6 ans. Le but de Google n'est pas qualitatif mais quantitatif. Le patron de Google n'a-t-il pas affirmé clairement son objectif : « *Notre business, c'est d'être exhaustif* », et donc de proposer le maximum de livres dans le maximum de langues et de laisser les utilisateurs et les chercheurs décider par eux-mêmes, par leurs choix, de la hiérarchie. Ce laissez faire et l'absence d'un projet culturel ambitieux ne peut qu'inquiéter. Derrière la mise à disposition des ouvrages en ligne, se cache peut-être une volonté plus terre à terre. Il s'agit de proposer en regard des ouvrages des liens vers des enseignes de vente à distance pour proposer les ouvrages à la vente ou à la location. Bien entendu, dans ces conditions se posera une fois de plus la question de la « ligne éditoriale » de Google. En effet, si un ouvrage est épuisé, quel intérêt économique y aurait-il à le mettre en ligne ?

BNM Bibliothèque Numérique Mondiale, 15.000.000 de livres numérisés par Google, avec à terme une volonté afficher de numériser... tout ce qui peut l'être, c'est-à-dire les ouvrages non-contemporains et libres de droit. Ce sont des ouvrages du domaine public, réservés à des spécialistes qui seuls seront capables d'en comprendre le contexte. Pour les autres lecteurs la consultation devrait être éclairée par des informations sur l'édition, sur la traduction, sur le contexte historique et politique. La numérisation « massive et totale » est un message trompeur qui passe très bien auprès du grand public, car il croit comprendre qu'il aura accès aux ouvrages dont il a besoin pour se cultiver, pour se former et pour se distraire. Ce qui n'est pas perçu est qu'en fait il

n'aura accès qu'à une sorte de bibliothèque musée, des ouvrages de plus de 70 ans, sans notice ni guide pour éviter qu'il ne se perde. Ce qui est d'autant plus paradoxal, puisque la société qui lui propose cela est justement celle qui jusqu'à présent le guidait à travers les millions de pages de la toile.

## 2. Grand projet institutionnel

Les projets du type de Google ne sont en fait pas nouveaux, et les politiques s'y sont essayés depuis déjà plusieurs années.

Nous avons vu dans la section 3.6 (page 144) que la BnF, à sa manière, avait déjà en 1998 lancé un grand programme de numérisation, ce qui correspondait à ses objectifs de conservation et de communication du patrimoine. Mais il s'agissait au départ uniquement de 100.000 documents numérisés, de plus ceux-ci n'étant accessibles qu'à travers le PLAO mis au point pour l'occasion, il n'était pas question de les mettre en ligne. Le chiffre est loin derrière ceux dont nous parlons aujourd'hui avec un projet tel que celui de Google, en revanche ce que la BnF promettait, qu'elle a tenu et qu'elle continue de tenir comme tous les intervenants de ce type, c'est la cohérence comme l'écrivait Jean Favier dans une lettre sur « *la numérisation des imprimés à la Bibliothèque nationale de France* », cité par Julie Bouchard :

*« Les textes numérisés ont été rassemblés afin de constituer pour les chercheurs un fonds de travail commun, un fonds de référence qui rassemble, dans les disciplines d'excellence de la bibliothèque, les textes et documents sources, ainsi que les ouvrages de synthèse et l'appareil critique nécessaires à leur étude, y compris des ouvrages de référence comme les dictionnaires. Comme n'importe quel fonds de bibliothèque, c'est une offre documentaire cohérente qui sélectionne les « meilleures » éditions en même temps que les éditions rares ou difficiles d'accès sur*

---

*un corpus d'auteurs et de textes, en français principalement. (Bouchard, Octobre 1996, page 24) »*

La riposte est d'abord venue de Jean-Noël Jeanneney, président de la BNF, à travers une tribune publiée dans *Le Monde* du 22 janvier 2005, qu'il a réitérée dans un ouvrage *Quand Google défie l'Europe, Plaidoyer pour un sursaut*. Le 27 avril 2005 la contre offensive européenne prenait corps avec l'engagement de 19 grandes bibliothèques publiques européennes pour « *une initiative commune visant à une numérisation large et organisée des œuvres appartenant au patrimoine de notre continent* ». Un message, que nous reprenons en annexe E, était envoyé dès le lendemain au Président du Conseil Européen à propos de la création d'une « bibliothèque numérique européenne ». Le 3 mai 2005, à Paris, lors des rencontres européennes de la culture, le président du conseil européen se prononçait pour un oui franc et massif en faveur de la création d'une bibliothèque européenne en ligne.

## **B. Quel lien avec les bibliothèques traditionnelles ?**

Les bibliothèques traditionnelles, comme la BNF par exemple, font naturellement partie des acteurs importants des projets de numérisation et de bibliothèque numérique. C'est en effet assez naturellement qu'elles ont commencé à réfléchir à la numérisation de leur fonds, grâce auquel elles peuvent rapidement proposer en ligne un contenu riche. Les projets qui n'ont pas un fond propre à numériser doivent généralement se tourner vers des ouvrages du domaine public, elles doivent alors proposer une lecture particulière, comme notre projet *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* ou une volonté de vulgarisation et de diffusion, comme le projet GUTenberg, ou une volonté de constitution d'une collection particulière, comme le projet HyperNietzsche<sup>183</sup>.

---

183. <http://www.hypernietzsche.org/>

Le dernier exemple est celui du projet Google, qui permet de mettre en évidence que les bibliothèques numériques ne pourront pas exister sans un solide partenariat avec les bibliothèques traditionnelles.

## 9.3 Bibliothèque électronique...

### ... ou système d'information.

#### A. Les outils

Nous l'avons dit, les choix techniques ne sont pas neutres, et cela pour plusieurs raisons, la première étant le plus souvent économique. Toutefois, nous pensons que la raison principale doit plus avoir à faire avec les ambitions culturelles du projet. Comme Simondon, nous cherchons à réconcilier technique et culture, pour cela considérant que les outils, comme objets techniques sont le fruit d'êtres vivants, nous devons maîtriser et comprendre la genèse, l'intention fabricatrice qui est à l'origine de la bibliothèque numérique, sous son aspect objet technique.

*« Découvrir un monde social et économique dans lequel l'utilisateur de l'objet technique soit non seulement le propriétaire de cette machine mais aussi l'homme qui l'a choisie et l'entretien. (Simondon, 1969, page 252) »*

Un certain nombre de technologies élémentaires pour ce type de projet peuvent être identifiées. La liste risquerait d'être longue sans un minimum de discipline, pour cela nous pouvons par exemple nous concentrer sur ce qui est vraiment commun à la majorité des projets de bibliothèque numérique sur le Web et donner cette courte liste :

**XML** a déjà été présenté dans la section 4.4 concernant la *Gestion électronique de documents*, page 84 du chapitre 4 et nous avons vu que c'était devenu la référence incontournable. Si le choix de XML coule pratiquement de source avec un projet de numérisation, il reste autour de cette technologie de nombreuses zones d'incertitude. D'une part il est un choix beaucoup plus critique qui est celui de la DTD dont nous discuterons dans le paragraphe suivant et d'autre part les outils utilisables pour la création et la modification des documents XML, comme ceux qui en permettent la consultation en ligne sont nombreux sans qu'émerge clairement une solution qui surpasse les autres.

**La DTD** est un choix stratégique important, de ce choix vont découler les outils disponibles ou pas pour transformer les ouvrages électroniques d'un format à un autre, les possibilités de partage et d'échange avec d'autres bibliothèques numériques.

**Apache** comme serveur ne fait pas de doute, c'est sur GNU/Linux qu'il semble le plus implémenté, mais ce choix ne doit pas être systématique, Open BSD est aussi une solution et qui de surcroît est présentée comme offrant encore plus de stabilité et de sécurité. En revanche l'utilisation de Windows est régulièrement déconseillée par les experts. Le langage de script qui animera le site et construira les pages dynamiquement ne semble pas déterminant, tout comme le gestionnaire de base de données.

Nous pourrions effectivement étendre cette liste au langage de script PHP, ou au système de gestion de base de données MySQL. Bien que ces outils soient très largement utilisés et retenus par les projets de publication sur le Web, ils peuvent être facilement remplacés par d'autres outils équivalents.

Pour continuer notre raisonnement autour de la construction de l'objet technique bibliothèque numérique, nous devons nous interroger sur ce qui

en constituera l'aboutissement. Ces briques assemblées pourraient donner naissance à un système qui ne parvient pas à trouver une cohérence interne, comme lors d'une greffe d'organe, le greffon devra être accepté et toléré par le reste de l'organisme.

« *L'objet technique concret, c'est-à-dire évolué, se rapproche du mode d'existence des objets naturels, il tend vers la cohérence interne, vers la fermeture du système des causes et des effets qui s'exercent circulairement à l'intérieur de son enceinte. (Simondon, 1969, page 46) »*

Toutes ces considérations devront être prises avec le recul qui devrait accompagner tout projet *a priori* technique, ces choix auront été faits en tenant compte des impératifs de pérennité, de conservation, d'accès ouvert à tous que nous avons déjà mentionné, en particulier pour notre projet *CoLiSciences*. C'est ce que nous rappelle Gilbert Simondon dans la citation suivante :

« *La spécialisation dite technique correspond le plus souvent à des préoccupations extérieures aux objets techniques proprement dits (relations avec le public, forme particulière de commerce), et non à une espèce de schèmes de fonctionnement compris dans les objets techniques. (Simondon, 1969, page 13) »*

Une bibliothèque électronique, sous forme de site Web ou autre (CD, DVD,...), n'est finalement, du point de vue technique rien de plus qu'un système d'information spécialisé dans l'archivage et la restitution de documents textuels. Si, techniquement, nous manquons peut-être un peu de recul dans le domaine spécifique des bibliothèques électroniques, nous avons en revanche la possibilité de trouver une nombreuse littérature sur les systèmes d'information, le déroulement de leurs projets de mise en œuvre, les erreurs à éviter et les points particuliers à vérifier. Sans entrer dans un luxe de détails, il est intéressant de noter que c'est l'usage, au sens de l'utilisation que l'on veut en faire et de

celle qui en sera faite finalement, du système qui a besoin d'être réfléchi, la scénarisation en quelque sorte des différentes boîtes noires le composant.

*« Tels qu'ils sont conçus actuellement, les systèmes d'information mémorisent rarement les modes opératoires, ils rassemblent surtout une masse souvent importante de rubriques et de symboles qui n'explicitent pas les motifs qui ont procédé à leur choix et à leur organisation. En conséquence, un système d'information tend à regrouper des données mises côte à côte sous la forme d'un stock, passif, pour lequel les utilisateurs ne disposent plus des clés qui leur permettront d'en faire un usage réellement signifiant puisque le pourquoi des choix, les enjeux qui les orientent... ont en partie ou radicalement disparu. [...] Dans ces systèmes d'information, le sens des informations (qui se trouve dans la capacité de les lier) est perdu au profit de leur juxtaposition. (Blasquez, 1992) »*

## **B. La relation outil et utilisation**

La bibliothèque n'existe que parce que des lecteurs viennent consulter des ouvrages. Avec le numérique, se pose la question des outils que le lecteur doit posséder pour accéder à la bibliothèque. Nous pensons que le PLAO<sup>184</sup> était l'exemple à ne pas suivre, parce que c'était un outil qui partait du postulat que l'utilisation à distance était impossible et ce type d'affirmation se révèle régulièrement erroné dans le cadre de l'utilisation des NTIC. Il interdisait l'accès aux ouvrages à travers Internet, et par là privait la bibliothèque de ses (cyber) lecteurs potentiels en leur interdisant l'accès aux ouvrages numériques. Nous pouvons considérer qu'il y a deux approches dans la mise

---

184. Poste de Lecture Assistée par Ordinateur, voir pour cela la section 3.6 sur les corpus électroniques à la page 144

en place d'un système de type « PLAO », une centrée sur l'utilisateur et l'autre sur les possibilités de l'outil :

**Utilisateur** Le système s'adapte à la configuration qu'il trouve sur le poste de l'utilisateur. Cela demande une réelle réflexion sur l'ergonomie du système et les fonctions indispensables. Il faudra ensuite rendre le système suffisamment souple pour adapter ces fonctions à tout type de poste utilisateur. Des fonction « étendues » pourraient être proposées à des utilisateurs dotés d'un poste particulier, mais ce genre de pratique ne fait généralement qu'engendrer des frustrations et des mécontentements.

**Outil** Le système ne fonctionne que sur un type de poste client particulier. Si l'accès à travers Internet est tout de même maintenu, cela se traduira par le téléchargement de *pluggins*<sup>185</sup> qui vont rendre le premier accès au site de la bibliothèque très fastidieux et très technique. De plus la pérennité de ces *pluggins* n'est pas du tout assurée, ni leur bon fonctionnement sur tout type de plate-forme, en particulier compte tenu de la rapide évolution des systèmes.

### C. Quelle bibliothèque ?

Nous nous posons toujours les mêmes questions que celles que se sont posés les concepteurs de la BnF, à 10 ans d'intervalle, comme le montre le mémoire de Julie Bouchard de 1995 (Bouchard, 1995). L'évolution technologique a simplement décalé ou repoussé certains questionnements, comme les possibilités de communications haut débit à travers Internet qui ont rendu aujourd'hui possible une large utilisation du mode images.

---

185. Ce sont de petits programmes téléchargeables qui fonctionnent à l'intérieur d'un explorateur Web, lui ajoutant des fonctions particulières qui ne sont pas prévues par la norme HTML du W3C.

## 1. Les typologies

Le dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* donne à lire des ouvrages numérisés sous trois formes :

1. Fac-similés, qui rappellent la forme originale de l'ouvrage.
2. Textes numérisés, texte électronique assorti de liens pour enrichir la lecture.
3. Lecture savante, les textes sont présentés après analyse.

Ces trois formes de présentation des ouvrages se traduisent en trois formes de typologies de parcours. Les fac-similés vont inciter à reproduire une lecture séquentielle telle que celle que propose un ouvrage papier. Les textes numérisés vont amener le lecteur à aller consulter le glossaire par exemple, c'est-à-dire à ouvrir des liens hypertextuels à l'extérieur du corpus proprement dit, voire à l'extérieur du site s'il fait référence à un dictionnaire en ligne, comme le TiLF<sup>186</sup>. La lecture savante pour terminer, propose de parcourir les ouvrages en suivant le fil conducteur d'une notion, donc en rupture totale avec la lecture séquentielle.

## D. Le Web Sémantique

Nous l'avons vu, (section 2. page 101), sous le terme Web sémantique, le dispositif mis en œuvre ne correspond que d'assez loin à l'élaboration d'un espace de travail en ligne où les chercheurs pourraient s'appuyer sur des outils dotés de sens. C'est dans cette direction que certaines recherches proposent de travailler, en étudiant un modèle de web socio-sémantique, fondé lui-même sur un Web Cognitivement sémantique.

---

186. Trésor Informatisé de la Langue Française, accessible depuis *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* par un double-clic sur n'importe quel mot.

*« Alors que le Web Sémantique est surtout destiné à l'inférence automatique d'agent logiciels portant sur des contenus stables et assez simples pour être décrits en logique formelle, le Web Cognitivement Sémantique, vise à soutenir les activités de recherche d'utilisateurs humains dans des corpus complexes et évolutifs. (Zacklad, 2005) »*

## 9.4 Bibliothèque et société

Dans l'étude des relations que la bibliothèque électronique peut entretenir dans un réseau de partenaires, que nous avons vu dans la section 9.2, à la page 195, nous avons cité et pris pour exemple le réseau que Google est train de mettre en place pour son projet de numérisation. Universités et bibliothèques sont appelés à mettre en commun leurs fonds pour les numériser et les mettre en ligne grâce à Google. Nous avons vu que ce projet a suscité de la part de plusieurs pays non-anglophones de nombreuses réactions. Pour l'Europe, multi-culturelle par sa construction, il est inquiétant de voir construit un réservoir de connaissance aussi important uniquement en langue anglaise, qui plus est par une société commerciale Nord-Américaine. C'est que la bibliothèque est un lieu où sont détenus tous les savoirs, son prestige peut aller au delà des frontières et s'étend à toute la civilisation qui a été capable de la construire, comme cela a été le cas pour la bibliothèque d'Alexandrie continue d'alimenter notre imaginaire collectif des siècles après sa disparition. Qu'une bibliothèque à vocation planétaire puisse être dirigée par une seule société, utilisant majoritairement une seule langue, quelle que soit la société et la langue, est une question qui mérite d'être posée et débattue. La capacité d'une entreprise, d'une institution, d'une ville, d'une région ou d'un pays à construire ce type de lieu est source de prestige. Le politique peut être

intéressé à utiliser cette image et sa valeur symbolique pour marquer un passage, un espace culturel ou une époque. En France, la construction de la Grande Bibliothèque, baptisée depuis « François Mitterrand » du nom du Président de la République qui en était l'instigateur, en est le témoin. Au delà des problèmes de conservation du patrimoine national que ce projet devait résoudre, le projet architectural et celui de numérisation ont constitué un tout qui a permis à l'homme politique de laisser une trace durable dans notre pays.

## A. Les aspects économiques

Le numérique bouleverse complètement la chaîne de production du livre. Pour un ouvrage papier, après l'investissement fait sur le premier exemplaire, pour le créer et en réaliser la mise en page, les coûts de fabrication des copies suivantes sont les mêmes pour le deuxième que pour le millième exemplaire. En revanche, avec le numérique le coup de copie est devenu complètement marginal.

Certains outils professionnels de PAO<sup>187</sup> sont très onéreux et complexes, ce qui a longtemps constitué un frein à leur utilisation par le grand public. Mais il existe maintenant de nombreuses alternatives à des prix très abordables, voire gratuits, permettant de produire par exemple un ouvrage type « Livre de poche ».

La contrepartie de ce gain est que le numérique autorise de nombreuses manipulations, en particulier la copie, qui rendent très difficile la rétribution des auteurs pour chaque livre en circulation d'une œuvre particulière.

L'industrie du livre n'est pas encore touchée aussi profondément que celle de la musique ou du cinéma, parce que le « ebook » dont nous avons parlé dans la section 3.4, page 73, n'est pas encore prêt à remplacer l'ouvrage papier

---

187. Publication Assistée par Ordinateur

en termes de confort de lecture, de simplicité d'emploi et de coût. Mais ce n'est pas prendre un gros risque que de prédire qu'un jour un dispositif rendant véritablement l'ouvrage papier obsolète sera mis au point. Il suffit de constater que les possibilités des téléphones portables rejoignent celles d'un ordinateur, que les agendas électroniques sont plus lisibles, plus puissants, que des connexions sans-fil à l'Internet sont présentes un peu partout dans les grandes villes, pour imaginer que tout cela pourrait converger vers la pérennisation du numérique en tant que dispositif portable de lecture et d'accès à l'information. Tout cela conduira peut-être d'abord la presse à utiliser ces nouveaux médias, en particulier la presse gratuite qui pourrait trouver ainsi un canal de distribution plus économique. La presse traditionnelle et le livre suivraient, si toutefois un modèle économique satisfaisant était mis en place.

Comme pour l'industrie de la musique et du film, celle du livre pourrait mettre en avant la question d'une juste rétribution des auteurs pour assurer sa propre survie. Cela pourrait paraître paradoxal, il suffit pour cela de réfléchir à la part qui revient à un auteur sur le prix d'un livre, généralement autour de 2 à 3% du prix de l'ouvrage seulement, pour mettre en évidence qui a besoin de qui. Bien entendu, l'auteur n'est pas le seul qui doit être payé, tous ceux qui ont participé à l'édition doivent l'être eux-aussi. À leurs salaires vient s'ajouter le papier, l'encre, l'amortissement du matériel d'imprimerie, les frais de logistique, stockage, transport. Dans une société où chaque jour une technologie numérique vient se substituer à une technologie plus ancienne, les auteurs pourraient se poser la question de savoir s'il est réellement utile de continuer à faire imprimer des ouvrages papiers, alors qu'ils pourraient éditer eux-mêmes leurs ouvrages numériques et ne plus avoir à supporter les frais de l'industrie du livre. Les lecteurs devraient se poser la même question avec encore plus d'acuité. Si les auteurs continuent pour l'instant à faire éditer

leurs œuvres sur papier c'est principalement parce que c'est le seul média qui leur assure un revenu. Ce marché gagnant-gagnant tiendra jusqu'à ce qu'une technologie numérique permettant la rétribution des auteurs soit mise au point, ou qu'un nouveau modèle économique soit développé.

Cette possibilité était déjà envisagée dans la bibliothèque numérique Xanadu de Ted Nelson, mais aucun dispositif ne permet encore d'en faire une réalité. Dans Xanadu, il est prévu que l'auteur d'un document soit payé au moyen d'un système de micro paiement, pris en charge par la bibliothèque, directement du lecteur à l'auteur. N'importe qui peut ajouter sa propre production et être à la fois lecteur et auteur, comme c'est le cas actuellement sur le Web. Si le document est consulté l'auteur est rétribué, sinon il ne l'est pas. Ce n'est plus un comité de rédaction ou de publication qui va devoir peser le pour et le contre, mesurer les risques à imprimer tel ou tel ouvrage, et finalement décider de la vie ou de la mort d'une œuvre. C'est directement le public. Cet aspect peut donner à réfléchir et être soumis à caution. En effet, dans un système où le plébiscite est utilisé pour payer les auteurs, il est plus « rentable » d'écrire des ouvrages légers et faciles. Mais, encore une fois nous sommes rattrapés par les paradoxes de notre société : n'est-ce pas déjà le problème avec l'industrie du livre qui repose de plus en plus sur des *Best Sellers* qui se taillent la part du lion et ne laissent que des miettes aux tirages plus modestes ?

« *Les causes économiques, en effet, ne sont pas pures. (Simondon, 1969, page 26) »*

Pour l'instant Xanadu reste une utopie, mais ce type de micro paiement peut être envisagé, les solutions techniques existent. Le DRM que nous avons déjà cité en est une parfaite illustration mais nous pourrions aussi imaginer un système fonctionnant sur le même principe que celui mis en place par

la SACEM pour rétribuer les compositeurs en fonction du passage de leur musique sur les ondes radio.

Le numérique va demander à l'industrie du livre de mettre en place un nouveau modèle économique. Ce n'est pas en fait le livre qui risque de disparaître, mais ce sont plutôt les imprimeurs qui ne sauront pas trouver ce modèle et avec eux toute l'industrie qui en découle : réseau de distribution, boutiques, etc. Le danger pour cette industrie est surtout que le modèle soit trouvé par un intervenant extérieur au livre, comme c'est le cas avec le iPod de la société Apple et les *majors* du disque.

« Les besoins se moulent sur l'objet technique industriel, qui acquiert ainsi le pouvoir de modeler une civilisation. (Simondon, 1969, page 24) »

## **B. Les aspects juridiques**

Les bouleversements que nous venons de décrire ont bien entendu un impact sur les aspects juridiques, parmi tout l'environnement légal qui avait été construit autour de l'industrie du livre pour protéger les investissements des éditeurs et des imprimeurs beaucoup de choses sont devenues inappropriées, voire obsolètes.

Notre système juridique de protection de la propriété intellectuelle a été bâti d'une part pour préserver les droits des auteurs, mais aussi pour permettre aux imprimeurs de supporter les lourdes charges que constituent la mise en route de l'édition d'un ouvrage et les risques encourus en cas de mévente. Sans coût de fabrication ni de stockage, ou presque, sans risque de mévente notre système juridique devra s'adapter à ce qui est offert par les technologies numériques. Ce nouveau modèle économique que nous mentionnons précédemment pourrait s'avérer soit illégal, soit risqué parce

que la protection légale dont il aurait besoin pour se développer fait encore défaut.



## **Conclusion et perspectives**



CHAPITRE 10

CONCLUSION

---

*« L'œuvre consacre l'échec de toute théorie. »*

Paul Klee

## Sommaire

---

<b>10.1 Conclusion Générale . . . . .</b>	<b>465</b>
A. Commentaires critiques sur la réalisation . . . . .	465
B. Notre modèle hypertextuel . . . . .	466
C. Contextes de publication sur le Web . . . . .	468
<b>10.2 Perspectives de recherche . . . . .</b>	<b>469</b>
A. Le futur du livre : Écouter un livre . . . . .	469
B. Bibliothèques numériques . . . . .	471
C. Le droit d'auteur numérique . . . . .	472
D. Production de document en collaboration . . . . .	473

---

## 10.1 Conclusion Générale

### A. Commentaires critiques sur la réalisation

Dans la première partie, nous posons la question de savoir ce qu'est *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences*, nous pourrions maintenant essayer de le voir comme un système cognitif, en cela qu'il répond aux questions posées par Francisco Varela pour identifier un tel système :

« Question 1 : Qu'est-ce que la cognition ?

Réponse : L'émergence d'états globaux dans un réseau de composants simples.

Question 2 : Comment cela fonctionne-t-il ?

Réponse : Des règles locales gèrent les opérations individuelles et des règles de changement gèrent les liens entre les éléments

Question 3 : Comment savoir qu'un système cognitif fonctionne de manière appropriée ?

Réponse : Quand les propriétés émergentes (et la structure résultante) sont identifiables à une faculté cognitive – une solution adéquate pour une tâche donnée. (Varela, 1996, page 77) »

C'est plus particulièrement dans la création de parcours que notre dispositif *C<sup>o</sup>L<sup>i</sup>S<sup>c</sup>iences* exprime son adéquation aux réponses données par Varela. Grâce au dispositif, nous pouvons étudier les chemins construits par des experts, essentiellement des enseignants de SVT (sciences de la vie et de la Terre), des historiens des sciences et des philosophes, correspondant en gros à des visites guidées, et décrire comment sont utilisés les outils proposés (par exemple les relations entre les notions aidant à construire une généalogie de ces notions en termes d'histoire des idées). Les observations des experts peuvent fournir des indications sur les usages possibles par des plus novices

et les chemins collectés enrichissent le corpus par l'adjonction de parcours sémantiques finalisés.

## **B. Notre modèle hypertextuel**

Pour répondre à notre objectif de définir les hypertextes et d'établir leur statut, nous avons mis en place un dispositif qui nous a permis de mettre en œuvre un système hypertextuel qui a évolué depuis le modèle à 3 couches de Dexter vers un modèle à 3 modules. Ce système théorique a été construit techniquement sur la base d'un moteur hypertextuel, ou plus précisément d'un automate, responsable de concrétiser la couche de présentation et celle de consultation. Cette réalisation permet d'une part de modéliser le fonctionnement du moteur hypertextuel en faisant appel aux théories habituellement utilisées avec les automates finis, comme nous l'avons montré avec les réseaux de Petri, et d'autre part d'élargir les capacités du système simplement. En effet, l'ajout d'une fonctionnalité à l'automate ne remet pas en question celles qui sont déjà en place et éprouvées. Pour rester dans la terminologie des réseaux de Petri, ceci correspond à la définition d'une nouvelle transition, ou à la redéfinition de l'action générée par une transition existante. De notre point de vue, cette modélisation d'un système hypertextuel apporte une nouvelle autonomie, indépendante du réseau, et dans nos perspectives de recherches futures il faudra penser à mettre en place ce même moteur sur un réseau différent pour vérifier la reproductibilité de ce modèle indépendamment du corpus utilisé.

Ceci nous amène à aborder une deuxième problématique qui consistait à développer de nouvelles modalités de constitution d'un corpus, et plus précisément à comprendre comment se faisait la transition d'un corpus textuel traditionnel vers un corpus hypertextuel. Avec la numérisation sous un

format XML nous avons défini la construction d'une base de connaissance, ne faisant pas appel aux traditionnelles bases de données relationnelles, par une méthodologie de déstructuration du texte qui enchaîne par une restructuration du corpus autour de concepts clefs, tels que les notions. Cette restructuration apporte une mise en relation des textes qui va au-delà des ouvrages, les notions pouvant être transversales au corpus, ce qui donne différentes formes d'appréhension globales du corpus : les vues « aériennes » de notion par exemple. Mais ce sont surtout les vues construites par notre nouveau dispositif de « lectochromie », en alliant le spatial au qualitatif, qui procurent le plus clairement de nouveaux modes de consultation et d'interrogation.

Cette transversalité se retrouve aussi dans notre construction de parcours de lecture, qui demandent aussi à être approfondis. Nous n'avons pour l'instant que la possibilité d'enregistrer des parcours, il faudrait développer le moyen de les visualiser d'une part et de les modifier d'autre part. Pour ce qui est de la visualisation, nous devons creuser plus avant comment construire graphiquement ces parcours, ceci pouvant prendre la forme de tableaux, d'arbres ou de cartes similaires à des cartes routières pour rester dans la métaphore du parcours. De plus, la visualisation devrait permettre de comparer facilement deux parcours différents, proposer des fonctions de zoom pour analyser précisément certaines étapes et en survoler d'autres, les mettre en confrontation avec le corpus qui apparaîtrait comme un territoire sur lequel le parcours s'effectue. Bien entendu, et c'était le deuxième point, ces parcours doivent être modifiables, le lecteur « ouvreuse de chemin » doit pouvoir leur apporter des commentaires pour guider les visiteurs qui suivront ses traces.

## C. Contextes de publication sur le Web

Nous ne reviendrons pas sur la question de savoir s'il y a rupture ou continuité dans le passage du livre imprimé à l'écran. De la même façon que l'imprimé donne lieu à des documents très divers, il en va de même avec le numérique. Il y aura sans doute continuité dans certains domaines, le livre imprimé n'est pas encore remplacé, loin s'en faut, par le e-Book et rupture dans d'autres, en particulier la presse et les magazines, ces imprimés qui ne sont pas fait pour être conservés mais « consommés » rapidement.

C'est donc plutôt sur la réflexion de la mise en ligne que nous désirons faire maintenant le point, sans pour autant s'engager dans une réflexion sur le futur du livre. Plusieurs approches sont possibles pour mettre du contenu en ligne et il semble qu'il soit possible de dégager des tendances fortes dans le domaine de la publication sur le Web.

**Le contexte technologique :** il est à la fois complexe et simple. Complexe parce qu'au premier abord une multitude d'outils s'offrent à nous, simple parce qu'au second abord il nous est apparu que le choix des Logiciels Libres plutôt que de logiciels propriétaires permettait d'éliminer un certain nombre de solutions, puis de choisir sereinement entre celles qui restaient, les outils étant à notre disposition pour réaliser toutes les études dont nous pourrions avoir besoin.

**Le contexte économique :** quand le plus gros moteur de recherche du moment décide de lancer une opération de numérisation de très grande envergure, cela ne peut pas nous laisser indifférent. Cela montre d'autant plus combien il est important que les chercheurs, en particulier au niveau Européen, continuent à proposer des solutions et des alternatives à ce type d'opération dont la finalité est avant tout commerciale.

**Le contexte politique :** est dans la continuation de la remarque précédente.

Après l'annonce de Google, nous avons entendu plusieurs déclarations d'intention de construction d'une bibliothèque électronique « plus universelle ».

La nouvelle tendance semble aller vers l'industrialisation de la numérisation des ouvrages papiers, tous les problèmes de droits d'auteurs ne sont pas entièrement résolus et Google connaît déjà des problèmes à ce sujet. Si un certain nombre de solutions sont envisageables, elles ne répondent certes pas à tous les cas de figure. Nous continuons à penser que les Logiciels Libres doivent prendre une plus grande part dans ce type de projet, en particulier dans un contexte multi-culturel de préservation de patrimoine.

## 10.2 Perspectives de recherche

En dehors de la continuité dans nos travaux sur l'hypertexte comme cela a été évoqué précédemment, de nouvelles perspectives de recherche pourraient s'organiser autour de deux axes principaux : d'une part l'étude des nouveaux usages de lecture et le futur du livre électronique, d'autre part comment vont être abordées les pratiques de travail collaboratif autorisées par l'utilisation de supports numériques et connectés pour les milieux comme la recherche par exemple.

### A. Le futur du livre : Écouter un livre

Un axe de recherche qui n'est pour l'instant abordé que comme substitut à l'écrit pour les non-voyants est celui du texte lu par un appareillage informatique, et à ce titre navigable comme n'importe quel texte écrit parce que l'interface orale s'interpose entre l'écrit et le « lecteur » non-voyant. Mais

en fait, le numérique prépare peut-être une nouvelle forme d'oralité, qui cette fois n'est plus transmise par les grillots<sup>188</sup>, mais par les lecteurs portables du type de l'iPod de Apple ou de simples clefs USB. Des « livres audios » ou *audiobooks* commencent à se répandre et il est même apparu un nouveau phénomène avec les radios en ligne : le *Pod Cast*. Il est en effet possible depuis quelques temps d'écouter une radio sur le Web<sup>189</sup>, la nouveauté c'est qu'il est maintenant possible de télécharger une émission de radio sur son baladeur numérique pour l'écouter tranquillement plus tard. Soit parce que l'on n'est pas disponible au moment de la diffusion, soit parce que l'on désire l'écouter plus attentivement, ou pendant une autre activité (jogging, trajet en voiture ou en transport en commun, ...).

Dans ce contexte où l'audio pourrait prendre plus de place, enregistré numériquement, donc plus facilement malléable, il est possible d'imaginer qu'après le *Pod Cast* qui permet d'enregistrer ses émissions, après l'*audiobook* qui permet d'écouter ses livres, le public va vouloir aller plus loin et participer à la création des contenus. L'hypertexte rend le texte plus dynamique et fait du lecteur un auteur, qui crée un nouvel ouvrage au fil de son parcours. Avec l'apparition de dispositifs audio pour remplacer l'écrit de nouveaux dispositifs hypertextes audio ou « hyperaudios » devront être mis au point pour faciliter la navigation à travers un discours enregistré.

Le livre n'est plus un objet inerte, c'est devenu un conteur, il parle et comme le conteur, il est capable de disgrégations dans le cours normal du texte.

---

188. Un grillot est un conteur en Afrique. Il est le détenteur de la tradition orale et de la mémoire du clan ou de la famille depuis les temps immémoriaux.

189. A première vue il semblait étrange que certains veuillent écouter une station de radio grâce un ordinateur coûtant un ou deux milliers d'euros, alors qu'un poste à transistor de quelques dizaines d'euros leur aurait permis la même chose, plus facilement. Mais c'est qu'avec le Web il n'existe plus de problème de zone de diffusion, il devient donc possible d'émettre et de recevoir de n'importe où dans le monde. Comme pour l'écrit, c'est la « zone de couverture » le critère qui l'emporte sur tous les autres.

## B. Bibliothèques numériques

Après la mise en place d'un dispositif de bibliothèque numérique tel que *CoLiSciences*, nous sommes en droit de nous poser la question de savoir ce que sera la bibliothèque du futur.

Compte tenu de nos objectifs, il va de soi que la numérisation doit être conduite avec intelligence, et pour cette raison elle sera sans doute longtemps partielle. Mais le rêve du Xanadu de Ted Nelson est chaque jour un peu plus à notre portée. Surtout devant le tour de force de Google qui commence à montrer ses premiers résultats, et ses premières limites.

Il découle deux attitudes de l'arrivée de Google dans ce domaine :

- Google va absorber toutes les Bibliothèques du monde. Il faut être avec Google et il n'y aurait qu'un seul système, mais alors comment en assurer le pluralisme ?
- Il faut être contre Google pour assurer le pluralisme, mais alors comment financer le projet : faire appel aux deniers publics ou réunir des partenaires commerciaux ? Quelle technologie utiliser ?

Pour nous, la Bibliothèque numérique devra permettre au lecteur d'accéder à l'intelligence du texte, sinon elle ne sera même pas un substitut de la bibliothèque comme institution.

A ce titre, *CoLiSciences* n'est pas une bibliothèque en ligne, c'est le bibliothécaire. Plus précisément c'est un dispositif qui permet de rendre les services de conseil et d'assistance qu'un bibliothécaire humain rend traditionnellement dans le cadre de ses fonctions.

### C. Le droit d'auteur numérique

Le numérique est un média en quête d'universalité, il est capable de contenir du texte, comme des images ou de la vidéo. Grâce aux technologies numériques les créateurs, comme les éditeurs et les distributeurs, ont maintenant à leur disposition des outils qui peuvent leur apporter beaucoup de souplesse et de facilité, pour la création et la diffusion des œuvres. Mais cette facilité est aussi une faiblesse, il devient très difficile de contrôler les duplications d'une œuvre numérique et de ce fait, de rétribuer les auteurs. Les technologies qui sont mises en place pour redonner au distributeur le contrôle de la diffusion d'une œuvre s'additionnent souvent de contraintes relativement lourdes pour le public. Par exemple, certains CD musicaux protégés contre la copie ne sont pas utilisables dans une radio de voiture, ou alors il n'est plus possible de prêter un livre numérique, ni de le revendre, ni de l'offrir. C'est ainsi que le droit du lecteur est en train d'être mis en avant et vient s'ajouter au droit d'auteur, par exemple avec les licences de diffusion du type *creative commons*. Les implications introduites par le support numérique sont à la fois légales et sociétales, elles nous font ressentir un besoin d'évolution.

Le dispositif légal du droit d'auteur et de la propriété intellectuelle a été mis en place dans un esprit d'universalité, tous les auteurs bénéficiant de la même protection, mais peut-il rester ainsi ? Un certain nombre de questions ne sont toujours pas posées, malgré leur trivialité. Est-il véritablement possible de protéger de la même manière un artiste numérique, dont les travaux peuvent être copiés à l'identique par le premier venu, que le serait un artiste traditionnel. Faut-il rétribuer de la même manière un orchestre symphonique et un musicien électronique qui compose et joue seul avec ses synthétiseurs ?

Ce nouveau droit du lecteur, assorti des économies engendrées par la mise en valeur d'un patrimoine sous forme électronique, devrait permettre une plus large diffusion de tout ce qui fait partie du domaine public. C'est sur cet aspect que se fondent beaucoup de projets de numérisation sans risquer d'enfreindre la loi sur le droit d'auteur : Google pour présenter une vaste bibliothèque numérique et attirer le public vers des ouvrages payants, la BNF pour présenter ses ouvrages les plus remarquables qui étaient réservés aux chercheurs, *CoLiSciences* pour valoriser un patrimoine qui n'était pas visible du public, etc.

Toutefois, le domaine public est aussi disponible pour d'autres usages. Il est possible de prendre un ouvrage dans le domaine public, de l'adapter et de lui ajouter une plus-value pour le présenter au public sous une forme qui permet de se le réapproprier<sup>190</sup>. Le traitement numérique pourrait être utilisé par certains pour cela, ce qui représente un risque éthique de détournement du domaine public. C'est ce que nous avons présenté dans l'annexe K.1, en montrant la complexité de la procédure de l'« achat » d'un ouvrage du domaine public, dont le coût est bien nul mais qui est pourtant protégé comme n'importe quel ouvrage commercial.

## D. Production de document en collaboration

Une direction qui n'a pas encore été suffisamment explorée est celle du travail collaboratif, à la fois pour les chercheurs qui travaillent sur *CoLiSciences*, et aussi pour les simples lecteurs qui pourraient partager des notes de lectures. L'exemple de l'encyclopédie Wikipédia mérite sans doute aussi d'être étudié

---

190. Par exemple, le « Blanche Neige » des frères Grimm appartient au domaine public, mais pas le film de Walt Disney du même nom et dont il est une adaptation. Aux termes de la loi votée en 1998 par Bill Clinton sous le nom de *Sonny Bono Act*, ce film et les personnages qui y apparaissent sont protégés pendant 95 ans après la sortie dans les salles, soit jusqu'en 2031.

de manière plus approfondie. Les dispositifs de type Wiki ouvrent de nouvelles voies au travail collaboratif, en permettant à un groupe d'individus qui ne se connaissent pas *a priori* de travailler ensembles. C'est même un dispositif qui permet d'aller bien au delà puisque ces individus sont rendus capables de produire un savoir ensembles. La capacité de l'équipe « *Hypertextes et textualité électronique* » à fédérer un certain nombre de projets de ce type, tant au niveau contenu, qu'au niveau technique, serait un formidable levier pour accélérer le processus de production et de mise en ligne de nouveaux ouvrages, la mise au point d'un dispositif technique plus sophistiqué, la construction et la validation de parcours de lecture faisant référence à des domaines différents,... Mais pour cela il faudrait que nous sortions du cadre traditionnel et institutionnel pour mettre en place un cadre de production collaborative qui permette à tous les intervenants de se développer dans le respect mutuel du travail des autres.

Un exemple de mise en place assez simple d'un modèle de travail collaboratif serait par exemple, de mettre en place ce type de procédure dans le cadre du contrôle d'un balisage sémantique fait de manière automatique par un dispositif de TALN, voire être une alternative à un balisage automatique.

Pour l'instant, la seule ébauche de production collaborative est liée à la lecture, ou plutôt à la construction de parcours de lecture qui sont visualisés sous la forme de lectochromes. Mais comme ceci se produit sans une participation active du lecteur, il n'est pas réellement possible de parler de production collaborative. Une piste intéressante consiste à continuer d'explorer la construction des parcours, tout en ajoutant une dimension collaborative qui autoriserait les lecteurs à commenter et à amender leurs propres parcours, voire ceux des autres lecteurs.

**Université de Nice - Sophia Antipolis**

*Faculté de Lettres, Arts et Sciences Humaines*

ÉCOLE DOCTORALE LETTRES, SCIENCES HUMAINES ET SOCIALES

**I3M**

INFORMATION, MILIEUX, MÉDIAS, MÉDIATIONS

## **Hypertextualité**

# **Construction du sens et représentation des connaissances**

Tome II

Bibliographie, Index, Glossaire et Annexes

## **THÈSE**

présentée et soutenue publiquement le 2 décembre 2005

*Pour l'obtention du grade de:*

**Docteur**

*Discipline :* **Sciences de l'information et de la communication**

par

Marc Augier

Directeur de thèse :

Monsieur le Professeur Norbert Hillaire

Jury:

Monsieur le Professeur Éric Bruillard (IUFM Caen)

Monsieur David Piotrowski (Chargé de recherche CREA-CNRS)

Monsieur le Professeur Pascal Vidal (CERAM Sophia Antipolis)

Monsieur Georges Vignaux (Directeur de recherche CNRS)

Monsieur le Professeur Norbert Hillaire (UNSA)

Mis en page avec la classe thloria.

# **Bibliographie, Index, Glossaire et Annexes**



# BIBLIOGRAPHIE

---

Cette bibliographie comporte 156 entrées.

## **Nota Bene :**

- *Datation des ouvrages et articles électroniques.* Les ressources trouvées en ligne sont généralement indiquées avec la date de dernière consultation dans le corps de la référence, avec mention de la date de publication quand elle était disponible.
- *Numérotation des pages des ouvrages et articles électroniques.* La référence au nombre de pages des références électroniques n'est pas pertinente, et n'est donc pas mentionnée, puisque l'impression dépend d'un grand nombre de paramètres. Le lecteur dispose de la possibilité de retrouver rapidement la citation concernée en utilisant la fonction recherche (CTRL + F) présente dans tous les navigateurs ou logiciels de visualisation. Cependant, pour les articles distribués autour de fichiers html distincts, nous faisons mention dans le corps de la citation, du nom du fichier duquel elle est extraite.
- *Consultation des ouvrages et articles électroniques.* Comme le Web est un environnement mouvant et difficilement contrôlable, certains documents pourraient disparaître ou être modifiés. Il pourra être utile de consulter le site <http://www.archive.org> qui archive un grand nombre de sites

Web. Par ailleurs, nous avons systématiquement conservé une copie de tous les travaux cités, que nous tenons à disposition de nos lecteurs et qui sera mise en ligne avec ce travail.

- *Ouvrages et articles parus en version imprimée et électronique.* Nous nous sommes efforcés, pour les documents ayant fait l'objet d'une publication papier et d'une publication web, de mentionner les deux références en entrée bibliographique. Quand le numéro de page n'apparaît pas dans le corps du texte de ce travail, c'est que nous avons uniquement consulté la version électronique du document.
- 

AGOSTINELLI, Serge.

*Comment penser la communication des connaissances*, chap. La construction d'un espace collectif de communication.

Paris : L'Harmattan, 1999.

AMÉLINEAU, C. et GIOVANNI, L.

« Utilisation pédagogique de l'outil logiciel hypertexte avec un public d'adultes illettrés en stage d'insertion sociale et professionnelle ».

Dans « Bruillard, E., Baldner, J.M., Baron, G.L. (dirs.) », p. 175-182. Paris, 1996.

AROT, Dominique.

*Les bibliothèques numériques (Document numérique, vol. 2, n°3-4)*, chap. Les bibliothèques françaises à l'âge du numérique.

Paris : Hermès, 1998.

---

ARSAC, Jacques.

*Implications philosophiques de la science contemporaine, tome 3 : Complexité, vie, conscience*, chap. l'informatique pose la question du sens.

Paris : PUF, 2003.

Sous la direction de Bernard d'Espagnat.

AUGIER, Marc.

*Les Logiciels.*

Paris : PUF, 1997.

AUGIER, Marc.

*Hypertextualisation et représentation des connaissances.*

Mémoire de DEA, UNSA, NICE, 2002.

AUGUSTIN.

*Les confessions de Saint Augustin.*

Cette traduction des Confessions est l'œuvre de M. Moreau, 1864.

En ligne à l'adresse : <http://www.abbaye-saint-benoit.ch/saints/augustin/confessions%/confessions.htm>.

AUSSENAC-GILLES, Nathalie et CONDAMINES, Anne.

« Documents électroniques et constitution de ressources terminologiques ou ontologiques ».

*Revue i3*, vol. 4, n° 1 : p. 75-94, 2004.

Publié ou en cours de publication, En ligne à l'adresse : [http://archivesic.ccsd.cnrs.fr/sic\\_00001016.html](http://archivesic.ccsd.cnrs.fr/sic_00001016.html).

AUTHIER, Michel et LÉVY, Pierre.

*Les arbres de connaissance.*

Paris : La Découverte, 1998.

BADDELEY, A.

*Working Memory.*

New York : Oxford University Press, 1986.

BALPE, Jean-Pierre, LELU, Alain, PAPY, Fabrice et SALEH, Imed.

*Techniques avancées pour l'hypertexte.*

Paris : Hermès, 1996.

BARON, Georges-Louis et BRUILLARD, Eric.

*Les technologies en éducation. Perspectives de recherche et questions vives.*

Paris : 31 janvier et 1<sup>er</sup> février 2002, INRP, MSH et IUFM de Basse-Normandie, 2002.

BARON, L., TAGUE-SUTCLIFF, J. et KINNUCAN, M.

« Labeled, Typed Links as Cues when Reading Hypertext Documents ».

*JASIS*, vol. 47, n° 12 : p. 896-908, 1996.

BEAUFILS, A.

« Appropriation de réseaux de navigation hypermédias par des élèves de collège : résultats d'une expérimentation et perspectives ».

*Actes des Troisièmes journées scientifiques Hypermédias et Apprentissages*, p. 225-236, 9-11 Mai 1996.

BECK, Kent.

*Extreme Programming Explained.*

Addison-Wesley Professional, 2000.

BERK et DEVLIN.

*Hypertext/ Hypermedia Handbook.*

New York : McGraw-Hill, 1991.

---

BERNERS, Lee, HENDLER, James et LASSILE, Ora.

« The Semantic Web ».

*Scientific American*, 17 mai 2001.

BERTRAND-GASTALDY, S. et DAVIDSON, C.

« Improved design of graphic displays in thesaurus through technology and ergonomics ».

*Journal of Documentation*, vol. 42, n° 4 : p. 225-251, 1986.

BLASQUEZ, James.

« Sciences sociales et intelligence artificielle ».

*Technologies Idéologies Pratiques*, vol. X, n° 2-4 : p. 29-40, 1992.

BLONDEAU, Olivier.

« Celui par qui le code est parlé. Pour une lecture expressive du phénomène hacker. », 2003.

Accédé en mai 2005, En ligne à l'adresse : [http://www.freescape.eu.org/biblio/article.php3?id\\_article=16%3](http://www.freescape.eu.org/biblio/article.php3?id_article=16%3).

BOUCHARD, Julie.

*Quand les puces auront mangé les livres.*

Mémoire de DEA, Conservatoire National des Arts et Métiers, Paris, 1995.

BOUCHARD, Julie.

« Des puces, des livres... et des hommes ».

*Futuribles*, vol. 213, Octobre 1996.

BOYDENS, Isabelle.

« La conservation numérique des données de gestion ».

*Document Numérique*, vol. 8, n° 2 : p. 13-22, 2004.

BROWN, P.J.

« Do we need maps to navigate round hypertext documents ? »

*Electronic publishing*, vol. 2, n° 2 : p. 91-100, 1989.

BRUILLARD, Éric.

*Les machines à enseigner.*

Paris : Éditions Hermès, 1997.

BRUILLARD, Éric.

*Hypertexte : de l'accès aux documents vers leur structuration.*

Paris : Colloque COLIS, PNER (Programme de Numérisation pour l'enseignement et la recherche), Maison des Sciences de l'Homme, 2002.

BRUILLARD, Éric, DE LA PASSARDIÈRE, Brigitte et BARON, Georges-Louis (eds.).

« Le livre électronique ».

*Sciences et Techniques Educatives*, vol. 5, 1998.

BUSH, Vanevar.

« As we may think ».

*The Atlantic Monthly*, 1945.

BUTOR, M.

« Les machines à lire ».

*Le Monde*, 11 décembre 1993.

CAMUS-VIGUÉ, Agnès et GAUDET, Françoise.

« Un colloque sur le web : une règle du jeu à inventer ? »

*Réseaux*, juin 2004, , n° 124, 2004.

Publié ou en cours de publication, En ligne à l'adresse : [http://archivesic.ccsd.cnrs.fr/sic\\_00001047.html](http://archivesic.ccsd.cnrs.fr/sic_00001047.html).

---

CHABIN, Marie-Anne.

« Document trace et document source. La technologie numérique change-t-elle la notion de document ? »

*Revue I3*, vol. 4, n° 1 : p. 141-158, 2004.

Publié ou en cours de publication, En ligne à l'adresse : [http://archivesic.ccsd.cnrs.fr/sic\\_00001020.html](http://archivesic.ccsd.cnrs.fr/sic_00001020.html).

CHARTIER, Roger.

« Pour une nouvelle économie du savoir, sous la direction de Ghislaine Chartron, Sylvie Fayet-Scribe, Brigitte Guyot et Jean-Max Noyer ».

*Solaris*, vol. 1 : p. 65-77, 1994.

Accédé le 8 juin 2005, En ligne à l'adresse : <http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d01/1char%tier.html>.

CHARTIER, Roger.

*Le livre en révolution.*

Paris : Textuel, 1997.

CLANCEY, W.J.

« Why today's computers don't learn the way people do ».

Dans « Future Directions in Artificial Intelligence », (P.A. Flach, R. A. (eds) Meersman, P. A. Flach et R. A. Meersman sous la direction de), p. 53-62. Amsterdam : Elsevier, 1991.

En ligne à l'adresse : <http://cogprints.ecs.soton.ac.uk/archive/00000455/00/115.htm>.

CLARKE, Arthur C.

*profil du futur : Un panorama de notre avenir.*

Paris : Planète, 1964.

CLÉMENT, Jean.

*Hypertextes et hypermédias : Réalisations, Outils, Méthodes*, chap. Du texte à l'hypertexte : vers une épistémologie de la discursivité hypertextuelle.

Paris : Hermès, 1995.

CLÉMENT, Jean.

« Des outils pour les lettres ».

*Dossiers de l'Ingénierie éducative*, vol. 32, 2000.

CLÉMENT, Jean.

« La littérature au risque du numérique ».

*Document Numérique*, vol. X : p. 1 à X, 2001.

COMTE-SPONVILLE, André.

*Dictionnaire philosophique*.

Paris : PUF, 2001.

CORDIER, Alain.

« Commission sur le livre numérique. Le livre numérique : rapport. »

Rap. tech., Ministère de la culture et de la communication, Paris, 1999.

En ligne à l'adresse : <http://www.culture.gouv.fr/culture/actualites/rapports/cordier/intro.htm>.

CROZAT, Stéphane et BACHIMONT, Bruno.

« Réinterroger les structures documentaires : de la numérisation à l'informatisation ».

*Revue I3*, vol. 4, n° 1 : p. 59-74, 2004.

Publié ou en cours de publication, En ligne à l'adresse : [http://archivesic.ccsd.cnrs.fr/sic\\_00001015.html](http://archivesic.ccsd.cnrs.fr/sic_00001015.html).

---

DAVID, René et ALLA, Hassane.

*Du Grafset aux réseaux de Petri.*

Paris : Hermès, 1997.

DEEPWEB.

« The deep Web : Surfacing Hidden Value ».

*The Journal of Electronic Publishing*, July 2001.

University of Michigan, En ligne à l'adresse : <http://www.press.umich.edu/jep/07-01/bergman.html>.

DEROSE, Steven J.

« XML and the TEI ».

*Computers and the humanities*, vol. 33, n° 1-2, 1999.

Selected papers from TEI 10 : Celebrating the Tenth Anniversary of the Text Encoding Initiative.

DUPOIRIER, Gérard.

*Les bibliothèques numériques.*

Paris : Hermès, 1999.

DUVIGNAUD, Jean.

« Débats ».

*Le Monde*, 18 janvier 1994.

ECO, Umberto.

*Lector in fabula. Le rôle du lecteur.*

Paris : Grasset, 1985.

EDUCAUSE.

« Edupage, 3 May 2000 », 2000.

<http://listserv.educause.edu/cgi-bin/wa.exe?A2=ind0005&L=>

edupage&T=0&F=&S=&P=159 (lu en ligne en juin 2005).

FARKAS, David K.

*Collaborative writing in industry : investigations in theory and practice*,  
chap. Collaborative Writing, Software Development, and the Universe of  
Collaborative Activity, p. 13-30.

New York : Baywood Publishing, 1991.

M.L. Lay & W.M. Karlis (Eds.).

FEBVRE, L. et MARTIN, H.-J.

*L'apparition du livre.*

Paris : Albin Michel, 1958.

GAINES et SHOW.

« Concept maps as hypermedia components ».

*International Journal Human - Computer Studies*, vol. 43 : p. 323-361, 1995.

GANASCIA, Jean-Gabriel.

« Le livre électronique : réflexion de prospective. »

*Publication en ligne du GIS Sciences de la cognition*, 2000.

En ligne à l'adresse : [http://www-apa.lip6.fr/GIS.COGNITION/  
livr1.html](http://www-apa.lip6.fr/GIS.COGNITION/livr1.html).

GANASCIA, Jean-Gabriel.

« Du néo-structuralisme supposé de l'hypertextualité ».

*Diogène*, vol. 16 : p. 9-24, 2001.

---

GIMONT, Jean-François.

*Le livre : du manuscrit à l'ère électronique.*

Liège (Belgique) : Ed. du CEFAL (Bibliothèque du bibliothécaire), 3e éd., 1998.

GOULEMOT, Jean Marie.

« De la lecture comme production de sens ».

*Pratiques de la lecture*, 1985.

GRANDBASTIEN, M.

« Quelques questions à propos de l'indexation et de la recherche de ressources pédagogiques sur le WEB ».

Dans Georges-Louis Baron et Éric Bruillard (eds.) « Les technologies en éducation : perspectives de recherche et questions vives », , vol. Actes du symposium international francophone, p. 220p. Paris : INRP, MSH et IUFM de Basse-Normandie, 31 janvier et 1er février 2002.

GUERPILLON, P.

« La mémoire du texte : antiquité et modernité ».

*Publication en ligne, Le PI@NOTYPE*, 2002.

En ligne à l'adresse : <http://www.pianotype.net/eBook/article2/index.html>.

HALASZ, Frank et SCHWARTZ, Mayer.

« The Dexter Hypertext Reference Model ».

*Proceedings of the Hypertext Standardization Workshop by the National Institute of Science and Technology (NIST)*, 1990.

HALASZ, Frank et SCHWARTZ, Mayer.

« The Dexter hypertext reference model ».

*Communications of the ACM*, vol. 37, n° 2 : p. 30–39, 1994.

En ligne à l'adresse : [citeseer.ist.psu.edu/halasz94dexter.html](http://citeseer.ist.psu.edu/halasz94dexter.html).

HILLMANN, Diane.

*Guide d'utilisation du Dublin Core (traduction de Guy Teasdale)*.

<http://dublincore.org/>, 2001.

En ligne à l'adresse : <http://www.bibl.ulaval.ca/DublinCore/usageguide-20000716fr.ht%#m>.

HUDRISIER, H. et HUDRISIER, G.

*Ordinateur et textes : une nouvelle culture?*, vol. 129, chap. Les Enjeux Culturels et Didactiques de la Lecture Assistée Par Ordinateur.  
Paris, 2000.

JACOBI, D. et BOQUILLON, M.

« Les représentations spatiales de concepts scientifiques : inventaire et diversité ».

*Didaskalia, Recherches sur la communication et l'apprentissage des sciences et des techniques*, vol. 5 : p. 11-24, 1994.

JACQUESON, A. et RIVER, A.

*Bibliothèques et documents numériques : concepts, composantes, techniques et enjeux*.

Paris : Editions du cercle de la librairie, 1999.

ISBN 2-7654-0716-9.

---

JAMMAMEJIAN, Aurélia.

« Adobe invente le livre chronodégradable ».

*Livres Hebdo*, n° 409, 19 janvier 2001.

J.B.GRIZE.

*Logique et Langage*.

Paris-Gap : Editions Ophrys, 1990.

JULLIA.

*Comment délinéariser le discours pédagogique, Le problème de l'architecture hypermédia comme faux problème.*

Mémoire de DEA, Céric, Université Montpellier III, 2001.

KANT, Emmanuel.

*Qu'est-ce qu'un livre ?*

PUF, 1995 (édition originale 1796).

LANGACKER, Ronald.

*Foundations of Cognitive Grammar*, vol. 1.

Stanford University Press, 1987.

LAUFER, Roger et SCAVETTA, Domenico.

*Texte, hypertexte, hypermédia.*

Coll. Que-sais-je ? Paris : PUF, 1992.

LAWRENCE, Steve.

« Online or Invisible ? »

*Nature*, vol. 411, n° 6837 : p. 521, 2001.

LE MOIGNE, Jean-Louis.

*La modélisation des systèmes complexes.*

Paris : Dunod, 1990.

LE MOIGNE, Jean-Louis.

*Le constructivisme - tome 2 : des épistémologies.*

ESF éditeur, 1994.

LEBERT, Marie.

« Le livre électronique : dossier ».

*e-Doc/Internet Actu*, vol. 20, mis en ligne le 26 octobre 2000.

En ligne à l'adresse : <http://www.internetactu.com/edoc/edoc20.html>.

LEBRAVE, Jean-Louis.

« Hypertextes-Mémoires-Écriture ».

*Genesis*, vol. 5, 1994.

LEBRAVE, Jean-Louis.

« Hypertexte et édition génétique : Flaubert ».

Dans « Banques de données et Hypertextes pour l'étude du roman »,  
Nathalie Ferrand (Ed.), 1997.

LEUF, Bo et CUNNINGHAM, Ward.

*The Wiki Way : Quick Collaboration on the Web.*

2001.

LÉVY, Pierre.

*Les technologies de l'intelligence. L'avenir de la pensée à l'ère informatique.*

Paris : La Découverte, 1990.

ISBN 2-02-013091-2.

LÉVY, Pierre.

*L'intelligence collective : pour une anthropologie du cyberspace.*

Paris : La Découverte, 1994.

---

MANGENOT, F.

*Les aides logicielles à l'écriture.*

Paris : Centre National de Documentation Pédagogique (CNDP), 1996.

MARSHALL, C.C. et ROGERS, R.A.

« Two Years before the Mist : Experiences with Aquanet ».

*Proc. 4th ACM Conference on Hypertext, ECHT'92, Milan*, p. 53-62, 1992.

MARSHALL C., Coombs J., Shipman F.

« VIKI : spatial hypertext supporting emergent structure ».

Dans « ECHT'94 », p. 13-23. 1994.

MARTIN, Henri-Jean.

*Les métamorphoses du livre. Entretiens avec Jean-Marc Chatelain et Christian Jacob.*

Albin Michel, 2004.

MASSELOT, Cyril.

*Accessibilité et qualité des systèmes d'information multimédia : transfert méthodologique et technologique.*

Thèse de doctorat, Université de Franche-Comté, U.F.R. Sciences du Langage, de l'Homme et de la Société, École Doctorale Langage, Espace, Temps, 2004.

MATURANA, Humberto R. et VARELA, Francisco J.

*L'arbre de la connaissance.*

Paris : Addison-Wesley, 1994.

MINSKY, Marvin Lee.

*The Society of Mind.*

New York : Simon and Schuster, 1987.

## Bibliographie

---

NANARD, M.

« L'interface des mondes réels et virtuels ».

*Cours n°10, INFORMATIQUE'93*, 1993.

NANARD, M.

*L'apport de travaux de recherche dans les hypertextes aux techniques éducatives.*

Paris : Bruillard, E., Baron, G.-L. et de la Passardière, B., (éds), 1994.

NANARD, M.

« Les hypertextes : au-delà des liens, la connaissance ».

*Sciences et Techniques Éducatives (STE)*, vol. 2, 1. : p. 31-59, 1995.

NIELSEN, Jakob.

*Hypertext and hypermedia.*

Boston : Academic Press, 1990.

NORMAND, Sylvie et BRUILLARD, Éric.

« Que révèlent les discours de futurs enseignants sur leur compréhension du fonctionnement des applications informatiques ».

*Point de vue, Sciences et Techniques éducatives*, vol. 8, n° 3-4 : p. 435-445, 2001.

OTLET, Paul.

*Traité de documentation, le livre sur le livre.*

Bruxelles : éditions Mundaneum, 1989.

PERELMAN, Ch.

*L'empire rhétorique, rhétorique et argumentation.*

Paris : Vrin, 1997.

PERROUSSEAUX, Yves.

*Manuel de typographie française élémentaire.*

Atelier Perrousseaux, 2000.

---

PETRI, Carl Adam.

*Kommunikation mit automaten.*

Darmstadt : Technical University, 1962.

PIAGET, J.

*Logique et connaissance scientifique.*

1967.

PIOTROWSKI, D.

« Lexicographie et informatique : autour de l'informatisation du TLF ».

Dans « Actes du Colloque de Nancy », Paris : Didier-Erudition, 1996.

PÉLISSIER, A. et TÊTE, A.

*Sciences cognitives textes fondateurs (1943-1950).*

PUF, 1995.

POMIAN, Krzysztof.

*Collectionneurs, amateurs et curieux. Paris, Venise : XVIe-XVIIIe siècle.*

Paris : Gallimard, Bibliothèque des histoires, 1987.

PRAX, Jean-Yves.

*La gestion électronique documentaire, Manager les flux d'information dans l'entreprise.*

InterEditions, 1998.

PUCCIA, Jean, SOULA, G., LE BŒUF, C. et BERNARD, M, J-L.and Fieschi.

« Appropriation des hypermedias : résultats et perspectives. Le cas de Top-Forum ».

Dans « Comment penser la communication des connaissances », , (S. Agostinelli sous la direction de). Paris : L'Harmattan, 1999.

QUENEAU, Raymond.

*Mille milliards de poèmes.*

Paris : Gallimard, 1961.

RASTIER, François.

*La sémantique interprétative.*

Paris : PUF, 1997.

RASTIER, François.

« Sémiotique et sciences de la culture ».

*Linx*, vol. 44-45 : p. 149-168, 2001.

RAYMOND, Eric.

« The Cathedral and the Bazaar ».

*First Monday*, vol. 3 N° 3, 1998.

Les modèles économiques basés sur le logiciel Open Source, En ligne à l'adresse : <http://www.linux-france.org/article/these/cathedrale-bazar/ca%thedrale-bazar.html>.

RAYMOND, Eric.

« Le chaudron magique ».

juin 1999, traduit en juillet.

Les modèles économiques basés sur le logiciel Open Source, En ligne à l'adresse : <http://www.linux-france.org/article/these/magic-cauldron/>.

REICHLER, C.

*L'interprétation des textes.*

Paris : Minuit, coll.« Arguments », 1989.

---

RENOULT.

« La bibliothèque numérique de la BnF ».

Dans « Les bibliothèques numériques », , ((éd.) Dupoirier, G. sous la direction de). Paris : Hermès, 1999.

RICHAUDEAU, François.

*Des neurones, des mots et des pixels.*

Atelier Perrousseau, 1999.

ROGER, D., LAVANDIER, J. et KOLMAYER, E.

« Navigation et Interfaces : Cartes conceptuelles et autres outils », 1998.

Vu en ligne Juin 2005, En ligne à l'adresse : <http://www.enssib.fr/Enssib/resdoc/DOCUMENTS/CERSI/kolmayer/k%olm-navig.bib.html>.

ROMARY, Laurent et HUDRISIER, Henri.

« TEI - Text encoding initiative », 2003.

Accédé en juin 2005, En ligne à l'adresse : <http://www.culture.gouv.fr/culture/dglf/riofil/tei.htm>.

ROSENTHAL, D.

*De la perception à l'appropriation d'un nouveau bien culturel : l'exemple du livre électronique.*

Mémoire de DEA, Paris I Panthéon Sorbonne, 1993.

ROSSI, Christian.

« De la diffusion à la conservation des documents numériques », 2005.

Working paper, En ligne à l'adresse : [http://archivesic.ccsd.cnrs.fr/sic\\_00001379.html](http://archivesic.ccsd.cnrs.fr/sic_00001379.html).

ROUET, Jean-François.

« Apprendre à lire un Hypertexte ».

*Cahiers de linguistique sociale*, p. 81-92, 1992.

ROUET, Jean-François.

« Hypermédias et individualisation des apprentissages : quels prérequis cognitifs ? »

*Le Français aujourd'hui*, vol. 129 : p. 9-18, 2000.

ROUET, Jean-François et TRICOT, A.

« Recherche d'informations dans les systèmes hypertextes : des représentations de la tâche à un modèle de l'activité cognitive ».

*Sciences et Techniques Éducatives (STE)*, vol. 2, 3 : p. 307-331, 1995.

ROUSSEAU, Jean-Jacques.

*Essai sur l'origine des langues.*

1781.

R.T.PEDAQUE.

« Document : forme, signe et medium, les reformulations du numérique. »

2003.

Article de travail., En ligne à l'adresse : [http://archivesic.ccsd.fr/sic\\_00000413.html](http://archivesic.ccsd.fr/sic_00000413.html).

RUGGIERO, Alessandro.

« Preservation of Digital Memory. Risks and Emergencies. Six Case Studies ».

Dans « The Future of Digital Memory and Cultural Heritage », Palazzo Vecchio, Florence, Italy, 16-17 October 2003 (48 pp.). : Istituto Centrale

---

per il Catalogo Unico delle Biblioteche Italiane e per le Informazioni Bibliografiche, 2003.

RUYER, R.

*La Gnose de Princeton.*

Paris : Fayard, 1975.

SABAH, Gérard.

*Ingénierie des Langues*, chap. Sens et traitements automatiques des langues (Chapitre 3), p. 77-107.

Hermès, 2000.

Jean-Marie Pierrel.

SALAÜN, Jean-Michel.

« Documents et numérique ».

*Contribution au rapport du Conseil d'Analyse Economique sur la Société de l'information.*, 2003.

Working paper, En ligne à l'adresse : [http://archivesic.ccsd.cnrs.fr/sic\\_00000831.html](http://archivesic.ccsd.cnrs.fr/sic_00000831.html).

SERRES, Michel et FAROUKI, Nayla et al.

*Le Trésor. Dictionnaire des Sciences.*

Paris : Flammarion, 1997.

SHANNON, C. E.

*Théorie mathématique de la communication*, chap. La théorie mathématique de la communication.

Retz-C.E.P.L., 1975.

W. Weaver, C. E. Shannon.

SHNEIDERMAN, Ben et KEARSLEY, Greg.

*Hypertext Hands-On !*

Reading, MA : AddisonWesley, 1989.

SIMONDON, Gilbert.

*Du mode d'existence des objets techniques.*

Paris : Aubier-Montaigne, 1969.

SIMONI, J.-L. et FLUHR, Ch.

*Accès à l'information à travers les graphes de termes.*

Communication à ISKO-France, Lille, 13-14 Oct., 1997.

STANDISH GROUP.

1995.

En ligne à l'adresse : [http://spinroot.com/spin/Doc/course/Standish\\_Survey.htm](http://spinroot.com/spin/Doc/course/Standish_Survey.htm).

STIEGLER, Bernard.

*La technique et le temps, 1. La faute d'Épiméthée.*

Paris : Galilée, Cité des Sciences et de l'Industrie, 1994.

STOCKINGER, Peter.

*Les nouveaux produits d'information. Conception et sémiotique du document.*

Paris : Hermes Science Publications, 1999.

STOCKINGER, Pierre.

*Traitement et contrôle de l'information, procédures sémiotiques et textuelles.*

Paris : Hermès, 2001.

TANNENBAUM, Andrew et WOODHULL, Albert.

*Operating Systems : Design and Implementation.*

Prentice Hall, 1987.

---

TEASDALE, G.

« L'hypertexte, historique et applications en bibliothéconomie ».

*Cursus*, vol. 1, n° 1, 1995.

Périodique électronique.

TRANSFERT.NET.

« Et voici le Linux book électronique. Le livre électronique : dossier de Transfert.net ».

*Transfert.net*, 2000.

En ligne à l'adresse : [http://www.transfert.net/fr/dossiers/dossier.cfm?idx\\_dossier=%9](http://www.transfert.net/fr/dossiers/dossier.cfm?idx_dossier=%9).

TRICOT, A. et BASTIEN, C.

« La conception d'hypermédias pour l'apprentissage : structurer des connaissances rationnellement ou fonctionnellement ? »

Dans « Actes des Troisièmes journées scientifiques Hypermédias et Apprentissages », p. 57-72. CREPS de Chatenay-Malabry, 9-11 Mai 1996.

E. Bruillard, J.M. Baldner, G.L. Baron.

TURING, A. M.

« Computing Machinery and Intelligence ».

*Mind*, vol. 49 : p. 433-460, 1950.

En ligne à l'adresse : <http://cogprints.ecs.soton.ac.uk/archive/00000499/00/turing.h%tml>.

VAN CUYCK, Alain.

« Les pratiques de lecture face au numérique : un fait social total ».

Dans « Première conférence internationale francophone en sciences de l'information et de la communication (CIFSIC), Dixième colloque bi-

latéral Franco-roumain, 28 juin-2 juillet 2003, Université de Bucarest,  
Communication faite dans le cadre de l'atelier A5 : do », 2003.

Soumis à évaluation, En ligne à l'adresse : [http://archivesic.ccsd.cnrs.fr/sic\\_00001385.html](http://archivesic.ccsd.cnrs.fr/sic_00001385.html).

VAN HERWIJNEN, Eric.

*SGML Pratique.*

International Thomson Publishing France, 1995.

VANDERDORPE, Christian.

*Du papyrus à l'hypertexte : essai sur les mutations du texte et de la lecture.*

Paris : La Découverte (Cahiers libres), 1999.

VARELA, F.

*Invitation aux sciences cognitives.*

Paris : Le Seuil, 1996.

VERGNAUD, Gérard.

*Apprentissages et didactiques, où en est-on ?*

Paris : Hachette, 1994.

VIGNAUX, Georges.

*Le Discours, acteur du monde. Énonciation, argumentation et cognition.*

Paris-Gap : Éditions Ophrys, 1988.

VIGNAUX, Georges.

« L'hypothèse du livre électronique ».

*Cahiers de médiologie*, vol. Lux , des Lumières aux traditions, n° 10 : p. 65-75,  
2000.

---

VIGNAUX, Georges.

« Le projet CoLiSciences », 2001.

Document de présentation, En ligne à l'adresse : <http://www.colisciences.net>.

VIGNAUX, Georges.

« Le programme COLISCIENCES et son futur : Parcours de lecture et navigations dans un site hypertextuel », 2002.

Programme Numérisation pour l'Enseignement et la Recherche (PNER).

VIGNAUX, Georges et KHADIYATOULAH, Fall.

« Genèse et construction des représentations : les discours sur l'informatisation ».

*Protée*, vol. 18, no 2(printemps 1990) : p. 33-44, 1990.

WIENER, Norbert.

*Cybernetic or control and communication in the animal and the machine.*

New York : John Wiley, 1948.

WIKIPEDIA.

« Benoît XVI — Wikipedia, L'encyclopédie Libre ».

2005a.

Accédé le 28-mai-2005, En ligne à l'adresse : [http://fr.wikipedia.org/wiki/Benoît\\_XVI](http://fr.wikipedia.org/wiki/Benoît_XVI).

WIKIPEDIA.

« Confiance — Wikipedia, L'encyclopédie Libre ».

2005b.

Accédé le 2-juillet-2005, En ligne à l'adresse : <http://fr.wikipedia.org/wiki/Confiance>.

WIKIPEDIA.

« Extreme Programming — Wikipedia, L'encyclopédie Libre ».

2005c.

Accédé le 10-octobre-2005, En ligne à l'adresse : [http://fr.wikipedia.org/wiki/Extreme\\_Programming](http://fr.wikipedia.org/wiki/Extreme_Programming).

WIKIPEDIA.

« Replies to common objections — Wikipedia, L'encyclopédie Libre ».

2005d.

Accédé le 28-mai-2005, En ligne à l'adresse : [http://en.wikipedia.org/wiki/Wikipedia:Replies\\_to\\_common\\_obje%ctions](http://en.wikipedia.org/wiki/Wikipedia:Replies_to_common_obje%ctions).

WIKIPEDIA.

« Wikipedia as a press source — Wikipedia, L'encyclopédie Libre ».

2005e.

Accédé le 28-mai-2005, En ligne à l'adresse : [http://en.wikipedia.org/wiki/Wikipedia:Wikipedia\\_as\\_a\\_press\\_s%ource](http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_as_a_press_s%ource).

WILLIAMS, Sam.

*Free as in Freedom, Richard Stallman's Crusade for Free Software.*

O'Reilly, 2002.

WITTGENSTEIN, L.

*Le tractatus Logico-Philosophicus.*

Paris : Gallimard, 1961.

---

WOLF, G.

« the curse of Xanadu ».

*Wired Magazine*, , n° 3(06), 1995.

En ligne à l'adresse : <http://www.wired.com/wired/archive/3.06/features/xanadu.html>.

XML.

*Extensible Markup Language (XML) 1.0 (Second Edition)*.

World Wide Web Consortium (W3C), 2004.

En ligne à l'adresse : <http://www.w3.org/TR/REC-xml>.

ZACKLAD, Manuel.

« Vers le Web Socio Sémantique : introduction aux ontologies sémiotiques ».

Dans « Ingénierie des Connaissance 2005 », 2005.

Soumis à évaluation, En ligne à l'adresse : [http://archivesic.ccsd.cnrs.fr/sic\\_00001347.html](http://archivesic.ccsd.cnrs.fr/sic_00001347.html).



# LES INDEX

---

# INDEX DES SITES WEB CITÉS

---

## **00h00**

<http://www.00h00.com/>, 86

## **ABU**

<http://abu.cnam.fr/>, 150, 554

## **ActiveWorlds**

<http://www.activeworlds.com/>, 340

## **Adobe**

<http://www.adobe.fr/products/main.html>, 95

## **adae.gouv.fr**

[http://www.adae.gouv.fr/article.php3?id\\_article=172](http://www.adae.gouv.fr/article.php3?id_article=172),

## **Agent Land**

<http://www.agentland.fr/>, 14

## **Agile**

<http://agilemanifesto.org/>, 431

<http://www.agilealliance.com/>, 431

<http://www.agilemodeling.com/>, 431

**AHDS**

<http://ahds.ac.uk/>, 143

**Aleph**

<http://aleph.ens.fr/>, 157

**Amazon**

<http://www.amazon.com>, 88

<http://www.amazon.fr>, 22

**ARTFL**

<http://www.lib.uchicago.edu/efts/ARTFL/databases/TLF/>, 554

<http://www.lib.uchicago.edu/efts/ARTFL/>, 161

**At folio**

<http://www.atfolio.net/>, 89

**Athena**

<http://un2sg4.unige.ch/athena/html/athome.html>, 161

<http://un2sg4.unige.ch/athena/html/francaut.html>, 161

**Audible**

<http://www.audible.com>, 87

**Barnes & Nobles**

<http://www.barnesandnoble.com/subjects/ebooks/ebooks.asp>,  
88

**BiblioML**

<http://www.biblioml.org/>, 218

**BookKenStock**

<http://www.bookenstock.com/>, 88

**BM Lisieux**

<http://www.bmlisieux.com/>, 157

<http://www.bmlisieux.com>, 554

**BNF**

<http://www.bnf.fr/>, 151

<http://www.bnf.fr>, 554

**British Library**

<http://www.bl.uk/treasures/gutenberg/homepage.html>, 370

**Burioni.it**

<http://www.burioni.it/news/novita/cdbasile.htm>, 162

**CETH**

<http://www.ceth.rutgers.edu/>, 159

**CNAM**

<http://cnum.cnam.fr/>, 150

**Cedars**

<http://www.leeds.ac.uk/cedars/>, 143

**Centre National du Livre**

<http://www.centrenationaldulivre.fr/doc/chiffres-cles.htm>,  
80

**Clairefontaine**

<http://www.clairefontaine-paperpc.com/>, 73

<http://www.clairefontaine.com>, 73

***CoLiSciences***

<http://www.colisciences.net/admin.php>, 576

<http://www.colisciences.net>, 6, 158, 317, 554

**Cornell**

<http://www.clir.org/pubs/abstract/pub93abst.html>, 143

**Cyberthèses**

<http://www.cybertheses.org/>, 158

<http://www.cybertheses.org>, 554

**Cytale**

<http://www.cytale.com/>, 89

**Debian**

<http://www.debian.org/>, 428

**Direction du livre et de la lecture**

<http://www.culture.fr/culture/guides/dll/fiche10chiffres2000-2001.html>, 80

<http://www.culture.gouv.fr/culture/sedocum/dll-cd.htm>, 80

**DocBook**

<http://www.docbook.org/>, 218

**Douglas Engelbart**

<http://sloan.stanford.edu/MouseSite/1968Demo.html>, 586

<http://www.bootstrap.org/>, 586

<http://www.histech.rwth-aachen.de/www/quellen/engelbart/ahi62index.html>, 586

<http://www.iath.virginia.edu/elab/hf10035.html>, 586

<http://www.superkids.com/aweb/pages/features/mouse/mouse.html>, 586

**Dublin Core**

<http://dublincore.org/>, 219

**ETC**

<http://etext.lib.virginia.edu/>, 159

**Editronics Education**

<http://www.editronics-edu.fr/>, 92

**Élysée**

[http://elysee.fr/elysee/francais/interventions/lettres\\_et\\_messages/2005/avril/message\\_a\\_propos\\_de\\_la\\_creation\\_d\\_une\\_bibliotheque\\_numerique\\_europeenne.29633.html](http://elysee.fr/elysee/francais/interventions/lettres_et_messages/2005/avril/message_a_propos_de_la_creation_d_une_bibliotheque_numerique_europeenne.29633.html), 568

**Everybook**

<http://www.everybook.net/>, 87

**eXtreme Programming**

<http://www.extremeprogramming.org>, 432

**FSF**

<http://www.fsf.org/philosophy/free-sw.fr.html>, 378

<http://www.fsf.org>, 378

**Fing.org**

<http://www.fing.org/>, 91

**Frantext**

<http://www.atilf.fr/frantext.htm>, 155, 554

**Gallica**

<http://gallica.bnf.fr/>, 155

<http://gallica.bnf.fr>, 554

**Gemstar eBook**

<http://www.gemstarebook.com/>, 86

**Grand Dictionnaire Universel**

<http://www.chass.utoronto.ca/epc/langueXIX/gdu/>, 161

**Gutenberg**

<http://promo.net/pg/>, 159

**Havas**

<http://www.havas.fr/>, 91

**HyperNietzsche**

<http://www.hypernietzsche.org/>, 391, 447

**HyperWave**

<http://www.hyperwave.com/>, 199

**IMSS**

[http://galileo.imss.firenze.it/memorie\\_digitali/](http://galileo.imss.firenze.it/memorie_digitali/), 397

**iTunes**

<http://www.apple.com/fr/itunes/>, 373

**KeeBook**

<http://www.keebook.com/>, 340

**Keebook**

<http://www.keeboo.com/>, 90

**Kenyon.edu**

[http://www2.kenyon.edu/people/adamsal/gui/bush\\\_engelbart.  
htm](http://www2.kenyon.edu/people/adamsal/gui/bush\_engelbart.htm), 587

**LED**

<http://www.numilog.com/>, 90

**Lamarck**

<http://www.lamarck.net/>, 150

<http://www.lamarck.net>, 554

**Lergonome.org**

[http://www.lergonome.org/pages/detail\\_articles.php?indice=  
14](http://www.lergonome.org/pages/detail_articles.php?indice=14), 328

**LexoTor**

<http://www.chass.utoronto.ca/epc/langueXIX/lexotor/>, 157

**Librius**

<http://www.librius.com/>, 88

**Licence Libre**

Artistic

<http://www.linux-france.org/article/these/licence/artistic/fr-artistic.html>, 389

#### BSD

<http://www.debian.org/misc/bsd.license>, 387

#### Creative Commons

<http://www.creativecommons.org/>, 390

#### Free Music Public License

<http://www.ethymonics.co.uk>, 390

<http://www.musique-libre.com>, 391

#### GNU Free Documentation License

<http://www.gnu.org/copyleft/fdl.html>, 391

#### GPL-fr

<http://www.linux-france.org/article/these/gpl.html>, 387

#### GPL

<http://www.fsf.org/copyleft/gpl.html>, 387

#### LGPL-fr

<http://www.linux-france.org/article/these/licence/lgpl/>,  
387

#### LGPL

<http://www.fsf.org/copyleft/lesser.html>, 387

#### Licence HyperNietzsche

<http://www.hypernietzsche.org/doc/puf/chevet2/>, 391

#### Licence Ludique Générale

<http://jeuxlibres.free.fr/llg.htm>, 391

#### Licence Publique Multimédia

<http://www.videontv.org/licence/>, 391

#### Licence Publique de Traduction

<http://www.in3activa.org/doc/fr/LPT-FR.html>, 391

Licence de Libre Diffusion de Documents

<http://pauillac.inria.fr/~lang/licence/lldd.html>, 391

Licence de libre diffusion Mnémosyne

[http://www.contenuouvert.org/article.php3?id\\_article=11](http://www.contenuouvert.org/article.php3?id_article=11), 391

Licence du GESI

<http://www.gesi.asso.fr/documentlibre/CharteGe.pdf>, 392

OpenContent License

<http://opencontent.org/opl.shtml>, 390

W3C

<http://www.w3.org/Consortium/Legal/copyright-software-19980720>,  
388

<http://www.w3.org/Consortium/Legal/ipr-notice-20000612>,  
388

artlibre.org

<http://artlibre.org/licence/lal.html>, 389

dsl

<http://dsl.org/>, 390

eff

[http://www.eff.org/IP/Open\\_licenses/20010421\\_eff\\_oal\\_1.0.html](http://www.eff.org/IP/Open_licenses/20010421_eff_oal_1.0.html), 391

perl

<http://www.perl.com/pub/a/language/misc/Artistic.html>,  
389

**Linux-france.org**

## INDEX DES SITES WEB CITÉS

---

<http://www.linux-france.org/article/these/cathedrale-bazar/cathedrale-bazar.html>, 392

### **Mandrake**

<http://frontal2.mandriva.com/>, 428

### **MediaLab (MIT)**

<http://www.media.mit.edu/micromedia/elecpaper.html>, 100

### **Memex**

<http://www.cs.brown.edu/memex/>, 586

[http://www.kerryr.net/pioneers/memex\\\_pic.htm](http://www.kerryr.net/pioneers/memex\_pic.htm), 586

### **Microcosm**

<http://www.multicosm.com/>, 200

### **Morphix-NLP**

<http://www.nlplab.cn/zhangle/morphix-nlp/>, 428

### **Mundaneum**

<http://www.mundaneum.be/>, 168

### **NEDLIB**

<http://www.kb.nl/coop/nedlib/index.html>, 144

### **Netcraft**

[http://news.netcraft.com/archives/web\\_server\\_survey.html](http://news.netcraft.com/archives/web_server_survey.html),  
424

<http://www.netcraft.com>, 603

### **Numilog**

<http://www.numilog.com/>, 614

<http://www.numilog.fr>, 624

### **OAIS**

<http://ssdoo.gsfc.nasa.gov/nost/isoas/us/overview.html>,  
143

**OASIS**

<http://www.oasis-open.org/>, 218

**OLF**

<http://www.granddictionnaire.com>, 85

**OTA**

<http://ota.ahds.ac.uk/>, 160

**OWL**

<http://www.w3.org/TR/owl-guide/>, 251

**Open source Initiative**

<http://www.opensource.org/>, 383

**OpenSource.org**

<http://www.opensource.org/licenses/index.html>, 387

**PANDORA**

<http://pandora.nla.gov.au/index.html>, 144

**Persée**

<http://www.persee.fr>, 554

<http://www.persee.org>, 157

**Portland Pattern Repository**

<http://c2.com/cgi/wiki?WelcomeVisitors>, 188

**Poynter.org**

<http://www.poynter.org/eyetrack2000/index.htm>, 328

**Proposicion.org.ar**

<http://proposicion.org.ar/doc/referencias/index.html>, 603

**RCA**

<http://www.rca.com/>, 86

**Rameau**

<http://rameau.bnf.fr/>, 208

**Red Flag**

<http://www.redflag-linux.com/eindex.html>, 428

**Red Hat**

<http://www.redhat.com/>, 428

**Revel**

<http://revel.unice.fr/>, 158

**Revues.org**

<http://www.revues.org>, 339

**Rocket Editions**

<http://www.powells.com/rocketeditions.html>, 88

**SCD**

<http://www.scd.univ-lille3.fr/>, 211

**SWPAT**

<http://swpat.ffii.org/patents/samples/index.en.html>, 608

**Scintilla**

<http://www.scintilla.org/>, 310

**Standish Group**

<http://www.standishgroup.com/>, 431

**Storyspace**

<http://www.eastgate.com/Storyspace.html>, 200

**SuSE**

<http://www.suse.de/en/>, 428

**Sun Microsystems**

<http://www.sun.com/980713/webwriting/wftw9.html>, 328

**TEI (Text Encoding Initiative)**

<http://www.tei-c.org/>, 217

**text-e**

<http://www.text-e.org/>, 124

### **TLFi**

<http://www.atilf.fr/tlfi.htm>, 156, 554

### **Ted Nelson**

[http://iberia.vassar.edu/~mijoyce/Ted\\\_sed.html](http://iberia.vassar.edu/~mijoyce/Ted\_sed.html), 587

<http://ted.hyperland.com/whatsay/>, 166

<http://www.iath.virginia.edu/elab/hf10155.html>, 586

<http://www.sfc.keio.ac.jp/~ted/>, 587

### **Universal Library**

<http://www.ulib.org/>, 159

### **Vannevar Bush**

<http://www.iath.virginia.edu/elab/hf10034.html>, 586

<http://www.isg.sfu.ca/~duchier/misc/vbush/>, 586

<http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm>, 586

### **W3C**

<http://www.w3.org>, 213

### **Wikibooks**

<http://fr.wikibooks.org/wiki/Accueil>, 162, 442

<http://fr.wikibooks.org/>, 149

### **Wikipédia**

[http://en.wikipedia.org/wiki/Wikipedia:Size\\_comparisons](http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons),  
193

<http://fr.wikipedia.org/wiki/Accueil>, 117

<http://fr.wikipedia.org/wiki/>, 379

<http://fr.wikipedia.org/>, 190

## *INDEX DES SITES WEB CITÉS*

---

[http://meta.wikimedia.org/wiki/Complete\\_list\\_of\\_language\\_Wikipedias\\_available](http://meta.wikimedia.org/wiki/Complete_list_of_language_Wikipedias_available), 192

<http://www.wikipedia.org>, 148

### **World Wide Web**

<http://www.w3.org/WWW/>, 200

### **XML**

<http://www.w3.org/XML/>, 213

### **XMetal**

<http://www.xmetal.com/>, 303

### **Xanadu**

<http://www.xanadu.com.au/xanadu/>, 200

<http://xanadu.com/>, 587

### **Xerox PARC**

<http://www2.parc.com/hsl/projects/gyricon/>, 100

### **Yahoo**

<http://www.yahoo.com>, 339 387

# INDEX

---

- Agile, 431
- Archivage, 143, 204, 450
- Automate fini, 31, 37, 285, 286
- Bibliothèque Numérique, 47, 107,  
118, 136, 150, 153, 160, 162,  
227, 257, 267, 294, 299, 306,  
333, 335, 338–340, 439, 440,  
443, 445, 447–449, 614
- Brevet logiciel, 418, 429, 602
- Catalogage, 62, 204, 205, 212, 443
- Classification, 58, 184, 278
- Classification Décimale de Dewey,  
    205, 208, 209, 211
- Classification Décimale Universelle,  
    168, 211
- Co<sup>L</sup>i<sup>S</sup>ciences*, 6, 12, 21–24, 50, 225,  
242, 243, 248, 257, 261, 264,  
281, 286, 295, 298, 299, 342,  
363, 375, 396, 422, 423, 465,  
608–610
- Architecture, 302
- Corpus, 228, 413
- Collection, 7, 8, 36, 57, 68, 69, 71,  
139, 141–143, 147, 150–155,  
161, 169, 189, 207, 218, 244,  
253, 257, 258, 447
- Creative Commons, 148, 390
- Cybernétique, 27, 440
- DocBook, 218
- DRM, 97, 371, 457, 614, 618
- DTD, 213, 303, 305, 308
- Dublin Core, 214, 219, 251
- Encyclopédie, 75, 112, 117, 118, 138,  
145, 153, 170, 192–195, 197,

- 203, 208, 209, 243, 257, 294,
- 338, 344, 379, 411, 413, 442
- Columbia Encyclopedia, 193
- Diderot et d'Alembert, 161, 169,
- 193
- Encyclopædia Britannica, 193
- Grolier, 193
- Microsoft Encarta, 193
- Nationalencyklopedin, 193
- Wikipédia, 148, 190
- Éphémérialisation, 393
- eXtreme Programming, 432, 433, 435,
- 510
- Gestion Électronique de Documents,
- 204
- Métadonnées, 118, 119
- Gestion des Connaissances, 204
- HTML, 159, 165, 202, 203, 283, 300,
- 304, 421, 609
- Hypertexte, 165, 170, 183, 199, 258,
- 282, 293, 299, 327
- Ancêtres, 169
- Corpus Hypertextuel, 327
- Fondements théoriques, 176
- Interface Hypertextuelle, 165
- Modèle de Dexter, 178
- Modèle à deux couches, 177
- Modèle à trois modules, 179
- Mundaneum, voir à cette entrée
- Parcours, 282
- Système Hypertextuel, 24, 55
- Indexation, 204
- Intelligence Artificielle, 28
- Lectochrome, 228, 343–347, 353, 361
- Lectochromie, 342, 347, 357
- Livre électronique, 129
- Cartable numérique, 90
- Cartable électronique, 90, 107
- Formats, 136
- Normes, 93
- Numérisation, 11, 24, 82, 138,
- 146, 149, 151, 283, 299, 304–
- 306, 310, 328, 413, 420
- Logiciels Libres, 50, 117, 270, 369,
- 375, 376, 378, 383, 386, 396,
- 419, 423, 425, 428, 432, 468,
- 602–605, 607, 608
- Ceux utilisés dans *Co<sup>L</sup>i<sup>S</sup>ciences*
- Apache, 397, 423, 429
- GNU/Linux, 397, 423
- mySQL, 397, 423, 429
- Php, 423, 429
- phpNuke, 297
- Innovation, 407

- 
- Open Source, 307, 375, 383, 385,  
386, 398, 402, 408, 421
- Métadonnées, 103, 118, 121, 128,  
143, 144, 214, 219, 235, 251,  
278
- Mundaneum, 65, 168, 211, 339, 444
- Normes
- Livres électroniques, 95
- Ontologie, 32
- Parcours, 7, 13, 18, 21, 22, 24, 50, 55,  
56, 62, 75, 103, 110, 113, 118,  
128, 169, 174, 175, 182, 186,  
187, 225, 228, 233–235, 242,  
260, 282, 289, 291, 292, 294,  
296, 326–328, 331, 342, 343,  
345, 346, 349, 362–364, 520
- Butinage, 13, 18, 245, 293, 364
  - Désorientation, 247
  - Lecture, 258
  - Navigation, 5, 8, 13, 14, 18, 21,  
83, 120, 127, 128, 170, 174,  
244, 245, 250, 260, 263, 291,  
331, 333, 342
- PDF, 95
- Rameau, 205, 208, 209
- Réseau de Petri, 31, 286
- SGML, 96, 159, 213, 214, 305, 422
- SVG, 96
- Système expert, 32
- Sémiotique, 27, 47, 133
- TCPA, 371, 414
- Palladium, 414
- TEI (Text Encoding Initiative), 116,  
159, 214, 217, 308
- Web, 129
- Web Sémantique, 13, 14, 20, 127,  
173, 247, 248, 453
- Langage, 251
  - RDF, 248
  - XML, 248, 250
  - Ontologie, 248
- Wiki, 125
- Wikipédia, 117, 148, 190, 379, 413
- Wikibooks, 149, 162
- World Wide Web, 5, 12, 13, 21, 60,  
69, 90, 92, 105, 118, 127, 159,  
165, 166, 200–203, 246, 247,  
283, 294, 305, 306, 328, 420,  
421, 424, 429, 602, 609
- Web profond, 203
- Wysiwyg, 60, 61, 105

*INDEX*

---

Xanadu, 13, 166, 200, 338, 444, 587

XML, 12, 22, 96, 138, 202, 213, 214,  
242, 273, 275–278, 280, 303–  
305, 421

# GLOSSAIRE

---

**Actes de langage** : « Je te baptise », « Je le jure », ces paroles n'ont pas pour but de décrire le monde, de transmettre un message, mais constituent en elles-mêmes une action, un « acte ». D'où la notion d'acte de langage due à John L. Austin. Un acte de langage a une valeur illocutoire, c'est-à-dire qu'il établit un type de relation de dialogue (jurer, questionner, ordonner, etc.); il peut avoir des effets perlocutoires, c'est-à-dire des conséquences pratiques (convaincre, intimider, etc.). Il y a des actes indirects (constater 'il est tard' pour demander de partir),.

**Allophone** : En phonologie, un allophone est l'une des réalisations sonores possibles d'un phonème. Au sein d'une même langue, les allophones ne constituent pas des unités pertinentes que le système de la paire minimale permettrait d'opposer. Par exemple, en français, le /r/ roulé ou non roulé constituent des allophones d'un phonème unique.,

**Analyse de conversation** : Champ de recherche qui a pris son essor dans les années 70 et qui renvoie à une multitude d'approches comme l'ethnographie de la conversation, l'ethnométhodologie, la sociolinguistique,

etc. L'analyse de conversation étudie les conversations en situation réelle. Elle montre que le langage courant est loin de correspondre aux règles de la grammaire formelle, qu'il existe beaucoup de différences dans l'expression selon les milieux sociaux et les situations, que le sens des mots dépend beaucoup du contexte, des intonations et des expressions faciales qui les accompagnent, que la conversation comporte beaucoup d'implicite (et donc une culture commune entre interlocuteurs), que la conversation est fortement ritualisée par des tours de parole, etc.,

**Austin John L. (1911-1960)** : Philosophe du langage, chef de file de l'école d'Oxford et initiateur de la théorie des actes de langage''' à travers son étude des «performatifs». Son œuvre de référence, publiée en 1962, s'intitule *How to do things with words* (*Quand dire c'est faire*, traduit en 1970),.

**Backdoor** : (« porte de service ») pratique consistant à livrer un système tout en laissant une possibilité cachée d'entrée, généralement un nom d'utilisateur et un mot de passe dont l'existence n'est pas révélée au propriétaire du système. Le fournisseur peut alors pénétrer dans le système comme bon lui semble.,

**Bande passante** : Mesure de la puissance de transfert de données d'un réseau (capacité par unité de temps),.

**Barthes, Roland (1915-1980)** : Professeur de «sémiologie littéraire» au Collège de France, où il occupa la chaire de sémiologie de 1977 à 1980. Il fut l'un des principaux animateurs de l'aventure structuraliste française. Il reste avant tout un très fin analyste de la culture contemporaine dans *Le Système de la mode*, Seuil (1967) et *Mythologies*, Seuil (1957),.

---

**Benveniste, Emile (1902-1976)** : Linguiste français, professeur au Collège de France de 1937 à 1970. E. Benveniste s'est fait connaître comme spécialiste des langues indo-européennes, puis a participé au développement de la linguistique structurale à partir de 1939. Son travail sur les déictiques (pronoms, démonstratifs, etc.) réintroduit le sujet dans le discours, et fait de lui un des fondateurs des études sur l'énonciation. Il est notamment l'auteur de : *Le Vocabulaire des institutions indo-européennes ; Problèmes de linguistique générale.*,

**BSD** : Version du système d'exploitation UNIX qui est à l'origine de beaucoup d'autres versions. GNU/Linux est compatible avec BSD. Actuellement, il existe plusieurs versions de BSD en logiciels libres : <http://www.netbsd.org>, <http://www.openbsd.org>, ou <http://www.freebsd.org/>,

**Bureau** : Espace de travail graphique permettant d'afficher simultanément plusieurs applications, c'est une métaphore imagée d'un bureau réel sur lequel des piles de documents peuvent être déposées.,

**Catalogage** : Pratique consistant à établir la description physique des documents afin d'établir leur notice bibliographique.,

**Cercle de Vienne** : Le cercle de Vienne a rassemblé dans les années 30, autour du philosophe allemand Moritz Schlick (1882-1936), des philosophes (Ludwig Wittgenstein, Rudolf Carnap) et des logiciens (Kurt Gödel, Frank Otto Neurath, Hans Reichenbach). Karl Popper était aussi un proche du groupe. Le Cercle produit en 1929 un manifeste, « la conception scientifique du monde », qui dénonce la philosophie spéculative et la métaphysique qui prétendent connaître le monde mais ne sont que des jeux de l'esprit. Le manifeste prend position en faveur de la science et de la raison.,

**Cheval de Troie** : Programme qui semble exécuter une tâche anodine, mais effectue en sous-main des opérations plus sensibles ; par exemple, un programme pourrait jouer un petit film pour vous amuser mais en même temps envoyer votre liste d'adresse Outlook à des tiers.,

**Chomsky, Noam (né en 1928)** : Professeur de linguistique au MIT (Massachusetts Institute of Technology) depuis 1956, il est le fondateur du courant de la « grammaire générative et transformationnelle », qui se distingue par la recherche des structures innées (par opposition à « acquis ») du langage naturel. Dans sa version première, la « grammaire » de Chomsky ne se préoccupait pas du sens des énoncés. En 1972, il a produit une théorie élargie à la sémantique. La grammaire générative a connu plusieurs formulations successives.,

**Client** : Dans une architecture client/serveur, c'est le nom donné au programme recevant ou demandant les ressources fournies par un serveur.,

**Code source** : Les logiciels sont une suite d'instructions en binaire, c'est-à-dire une longue suite de uns et de zéros, compréhensible pour l'ordinateur, mais incompréhensible pour un humain. Pour rendre plus facile la programmation, on utilise des langages de programmation, compréhensibles pour un être humain un peu habitué. Le texte écrit en langage de programmation, le code source, est lu par un programme spécial, le compilateur, qui traduit le code source en instructions binaires pour l'ordinateur. Sans le code source, il est impossible de corriger ou d'améliorer les programmes ; c'est la raison pour laquelle l'esprit des logiciels libres préconise la publication systématique du code source avec tous les programmes.,

---

**Cognitif/cognitivism** : Au sens large, le terme « cognitif » désigne tout ce qui concerne les fonctions mentales. Dans un sens plus restreint, la perspective « cognitive » renvoie à une théorie psychologique qui considère les fonctions mentales comme des dispositifs de traitement de l'information.,

**Computation, computationnisme** : Computation provient de l'anglais *computer* qui signifie « ordinateur », mais aussi calculateur. Le computationnisme est un paradigme des sciences cognitives qui envisage la pensée sur le modèle de l'ordinateur et comme un programme informatique. Le cerveau est considéré comme une machine à traiter de l'information et la pensée est réductible à une suite d'opérations mathématiques et logiques simples, qui se succèdent selon un ordre déterminé. On parle aussi de modèle « comuto-représentationnel » ou « comuto-symbolique » ou parfois encore de « symbolisme ».,

**Connotation** : C'est le *sens figuré*. Les connotations sont les significations qui viennent s'ajouter au sens ordinaire d'un mot, l'ensemble des références implicites auxquels elles renvoient. La connotation introduit un double langage au sein d'un langage de base. Généralement, les termes scientifiques ne possèdent pas de connotation parce qu'ils renvoient à la réalité qu'ils décrivent.,

**Cracker** : Informaticien qui utilise ses compétences dans des buts malhonnêtes et destructeurs, que cela soit par intérêt personnel ou par nihilisme. Ne pas confondre avec hacker, comme le font malheureusement souvent les médias.,

**Dénotation** : C'est le *sens dénoté*, *sens propre* ou *sens premier*. La dénotation est la première définition que donne le dictionnaire d'un lemme, cette

signification invariante et non subjective est celle sur laquelle tout le monde s'accordera.,

**Descripteur** : Expression ou mot choisis dans un thésaurus pour représenter le contenu d'un document.,

**Descripteur** : Voir Vedette.,

**Disque dur** : Périphérique de stockage de masse ; les ordinateurs modernes contiennent typiquement au moins un disque dur sur lequel est écrit le système d'exploitation. C'est également là que sont habituellement stockées les données de l'utilisateur. ,

**Document primaire** : Document complet et original.,

**Document secondaire** : Document issu d'un traitement documentaire (par exemple : bibliographie),,

**DRM** : *Digital Rights Management*, technologie de contrôle et de protection des droits d'utilisation de contenus numériques.,

**Énonciation** : « Tu vois ce que je veux dire ? », « J'aimerais que tu ailles... »  
Le discours porte diverses marques de la présence du locuteur et de la situation de parole (autoréférence) et permet d'établir des degrés variables d'adhésion ou de distance par rapport aux propos proférés. L'énonciation est donc un phénomène complexe qui distingue la communication d'une simple transmission de signaux codés.,

**Entrée** : Voir Vedette-matière.,

**Extranet** : Réseau privé utilisant la technologie Internet et accessible à travers celui-ci, mais protégé à l'entrée par un système de type nom d'utilisateur/mot de passe.,

**Forme** : Nature, présentation matérielle et support du document par ex. : Dictionnaires, Ouvrages illustrés, Cédéroms.,

**Forme exclue** : Voir Terme exclu.,

---

**Forme retenue** : Voir Vedette.,

**Freeware** : Un freeware ou «gratuiciel» est un logiciel gratuit. Il n'existe aucune garantie sur le fait que ce logiciel restera gratuit, ni sur les raisons de sa gratuité, ni sur son innocuité. Il n'y a aucune relation entre freeware et logiciel libre.,

**FTP** : File Transfer Protocol : protocole pour le transfert de fichiers à travers un réseau. Non sécurisé.,

**Geek** : Passionné de technique, particulièrement informatique. C'est l'état d'esprit du hacker, sans en avoir nécessairement toutes les compétences techniques.,

**GNU** : Acronyme (récursif) de GNU's Not UNIX. Projet créé par Richard Stallman pour développer un système d'exploitation libre, c'est maintenant un grand projet de développement de logiciels libres.,

**GNU/Linux** : Système d'exploitation réunissant le noyau Linux et les logiciels issus du projet GNU.,

**GPL** : *General Public License* Accord de licence qui protège l'utilisateur et l'auteur d'un logiciel. La GNU garantit que les logiciels libres ne peuvent être détournés et resteront libres.,

**Grammaire** : La grammaire analyse la construction des phrases. On distingue la syntaxe (étude des règles d'accords, de structuration de la phrase), et la morphologie, qui s'intéresse à la forme des mots (déclinaison, conjugaison).,,

**Grammaire générative** : La grammaire générative et transformationnelle a été fondée par le linguiste américain Noam Chomsky.,

**Grice, Paul H. (1913-1988)** : Philosophe-linguiste pragmatique américain de Harvard, il s'est fait connaître à partir de 1957 pour ses analyses sur l'intentionnalité dans le discours. Il a travaillé sur les règles de la

communication et ce qu'impliquent leur non observance. Un recueil de ses articles a été publié en 1989 (*Studies in the Way of Word*),.

**Groupe de discussion** : Voir Usenet.,

**Hacker** : En anglais un *hack* est un cheval à louer, à l'origine le hacker est donc une «bête de somme» de la programmation. Son labeur acharné a fait de lui un virtuose et sa curiosité l'a amené à pénétrer des systèmes auxquels il n'avait pas accès. Mais ce n'est pas un «pirate», pour lequel le terme correct est *cracker* et son comportement n'est jamais délictueux. C'est un passionné pour qui la programmation est un jeu. Les systèmes modernes doivent beaucoup aux hackers : Unix puis Linux, Internet et le WWW par exemple.,

**Herméneutique** : A l'origine, science de l'interprétation des textes philosophiques ou des textes sacrés et qui vise à en dégager la ou les significations profondes, le sens caché. L'herméneutique s'est étendue aujourd'hui à d'autres disciplines comme la psychanalyse, la sémiotique pour lesquelles l'interprétation consiste à décrypter les phénomènes observables de la réalité comme des signes d'une réalité cachée. Dans *Le Conflit des interprétations*, Seuil, Paris, (1969) le philosophe Paul Ricoeur a montré qu'un même ensemble de faits est susceptible d'être soumis à une pluralité d'interprétations toutes cohérentes. Se pose alors la question de la validité et du contrôle des interprétations. Dans *Les limites de l'interprétation*, Grasset, Paris, (1992) le sémiologue Umberto Eco explique que la spéculation interprétative a des limites.,

**HTML** : HyperText Markup Language ou « Langage Hypertexte à balises » : langage standard pour la présentation de données sur Internet.,

**HTTP** : HyperText Transport Protocole : protocole superposé à TCP/IP pour le transfert de données HTML essentiellement.,

---

**Hymes, Dell (né en 1927)** : Anthropologue et linguiste américain, il a été professeur à Harvard, Berkeley, puis en Pennsylvanie et enfin en Virginie. Il est surtout connu pour avoir proposé et développé une « ethnographie de la parole » qui influença la pratique des linguistes de la communication. Il a également impulsé un courant d'« anthropologie de la communication », représenté depuis 1972 par la revue *Language in Society*.

**Implicite** : Ensemble des phénomènes par lesquels l'énonciation exprime davantage ou autre chose que le sens immédiat de l'énoncé. Pour la pragmatique, outre les présupposés, l'implicite concerne ce qu'on laisse entendre (involontairement), ce qu'on donne à entendre (qu'on suggère à dessein), ce qu'on fait entendre (qu'on oblige à déduire par une violation apparente des principes de conversation). Les interactionnistes élargissent l'implicite à l'ensemble des signes qu'échangent inconsciemment les interlocuteurs.

**Incunables** : Un incunable (du latin *incunabulum*, « berceau », à comprendre comme « enfance de l'imprimerie ») est un livre imprimé avant 1501, juste avant la naissance de la typographie. Sa mise en page imite généralement celle des livres manuscrits.

**Indexation** : Description du contenu d'un document à l'aide d'un langage documentaire pour faciliter la mémorisation de l'information et permettre la recherche.

**Indexation** : Opération destinée à représenter par les éléments d'un langage documentaire ou naturel les résultats de l'analyse d'un document, en vue d'en faciliter la recherche.

**Inférence** : Opération logique qui permet de passer d'une proposition à une autre par déduction, induction, généralisation. On désigne aussi par là

toute opération par laquelle on admet une proposition dont la vérité n'est pas connue directement, en vertu de sa liaison avec d'autres propositions déjà tenues pour vraies. Par extension, on nomme ainsi la proposition inférée.,

**Instruction** : Ordre compris par le processeur qui va l'exécuter. Une instruction peut être un calcul logique, mathématique ou un accès à la mémoire.,

**Interactions verbales** : L'étude des interactions verbales fait aujourd'hui partie des sciences du langage à plusieurs titres. Au niveau le plus global elle distingue des types d'interaction selon qu'ils sont : complémentaires, plus ou moins formels, plus ou moins compétitifs, En fait, toutes ces situations exigent un minimum de coopération et de conventions et peuvent être assimilées à des contrats de communication. Les analyses peuvent faire appel à tous les outils de l'analyse du discours (pragmatique, énonciation, argumentation, actes de langage...). Elles peuvent porter spécifiquement sur les aspects interactifs de ces discours (tours de parole, reprises, reformulations) ou bien sur des procédés généraux de construction du sens (présupposition, implicitation, modalisation, modulation).,

**Internet** : « Réseau de réseaux » ; système qui interconnecte des réseaux locaux, permettant ainsi l'échange de données dans le monde entier.,

**Inventaire** : Registre qui répertorie tous les documents de la bibliothèque.,

**Inversion** : Opération qui permet de placer en premier élément d'une vedette le terme le plus significatif.,

**Jakobson, Roman (1896-1982)** : Linguiste d'origine russe, installé en Tchécoslovaquie en 1921, puis aux Etats-Unis en 1941, il publie son œuvre principale *Essais de linguistique générale* en 1963. Il est un des fondateurs de la

---

phonologie moderne, mais s'intéresse aussi à la poésie et à la fonction de communication du langage. Il collabore avec les fondateurs de la théorie de l'information, Shannon et Weaver, et aura une grande influence sur le mouvement structuraliste en France.,

**Labov, William (né en 1927)** : Professeur de linguistique à New York, il est un des fondateurs de la sociolinguistique. Son étude de 1962, sur l'île de Martha's Vineyard, montrait comment un changement linguistique pouvait être directement lié à un changement social. Il a mené ensuite de grands programmes de dialectologie urbaine. Son œuvre de référence, *La Sociolinguistique*, a été traduite en 1975.,

**LAN** : Local Area Network : réseau local (de petite envergure).,

**Langage d'indexation** : Langage documentaire utilisé pour fournir, dans un catalogue, une bibliographie ou un index, un accès cohérent au sujet des documents. Il est composé d'un vocabulaire contrôlé de termes représentant des concepts et des entités, et d'une structure sémantique permettant de relier ces termes entre eux. Il peut également comporter une syntaxe pour combiner ces termes dans des chaînes.,

**Langage documentaire** : Langage contrôlé utilisé pour codifier le contenu d'un document. Il en existe deux types : les langages à structure hiérarchique (classification), qui utilisent des indices symboliques et les langages à structure combinatoire (thésaurus ou listes d'autorité), qui utilisent les mots du langage naturel.,

**Langue/parole** : Distinction fondatrice de la linguistique générale. La langue, système conventionnel abstrait partagé par tous les locuteurs, s'oppose à la parole, réalisation pratique d'un individu dans une situation.,

**Lemme** : Le lemme (ou encore lexie) est l'unité autonome constituante du lexique d'une langue. Dans le vocabulaire courant, on parlera plus souvent de mot, notion qui, cependant, manque de clarté. On construit des énoncés avec des lemmes, les lemmes sont faits de morphèmes.,

**Lexique** : Ensemble des mots, des unités formant le vocabulaire d'une langue. La lexicologie étudie le vocabulaire et la formation des mots.,

**Libre** : Par opposition à propriétaire ou fermé : logiciels dont le code source est disponible et auxquels il est autorisé d'apporter des modifications.

**Linux** : Implémentation libre d'un noyau (kernel, c'est-à-dire le coeur d'un système d'exploitation) UNIX, développé à partir de 1991 par Linus Torvalds, et par la suite par d'innombrables personnes. Par extension, le système d'exploitation et parfois même l'ensemble du système d'exploitation et des principaux logiciels qui tournent dessus. Pour cette dernière signification, on parle aussi de GNU/linux, pour souligner l'importance des logiciels GNU qui font de Linux un système utilisable. ,

**Liste d'autorité-matière** : Liste des vedettes et des termes exclus reliés entre eux par des relations sémantiques, élaborée et développée au fur et à mesure des besoins des bibliothèques qui l'utilisent. Elle est généralement encyclopédique.,

**Logiciel** : Suite d'instructions qu'un ordinateur comprend et utilise pour effectuer une tâche. Les logiciels sont tout d'abord écrits dans un langage compréhensible par un être humain. Ce texte, nommé code source, est ensuite traduit en un autre langage, cette fois compréhensible par l'ordinateur. Le résultat de cette traduction s'appelle code exécutable.

---

**Logiciel Libre** : Logiciel dont le code source est disponible et modifiable selon les termes d'un accord de licence de type GPL,

**Logiciel libre** : L'expression «Logiciel libre» fait référence à la liberté pour les utilisateurs d'exécuter, de copier, de distribuer, d'étudier, de modifier et d'améliorer le logiciel.,

**Mémoire vive** : Aussi appelée « RAM » (Random Access Memory) ou « mémoire centrale » ; mémoire d'accès rapide, généralement volatile, c'est-à-dire dont le contenu est perdu à la mise hors tension. Les programmes sont chargés en mémoire vive pour exécution.,

**Mac OS X** : Système d'exploitation fourni en standard par Apple. Il est basé sur un noyau BSD (« Darwin », qui est open source). Depuis sa sortie, de très nombreuses applications libres ont été portées pour Mac OS X, de sorte qu'il est possible de se procurer gratuitement une grande quantité de programmes de qualité très faciles à utiliser. Le passage de Apple à UNIX montre bien à quel point ce système est incontournable, et l'adoption de la philosophie open source pour une partie du code est une illustration des avantages que ce mode de développement apporte à tous, utilisateurs comme développeurs. Mac OS X peut être pré-installé sur des Macs vendus – le prix du Mac est alors majoré par le prix de l'OS X.,

**Morphème** : Plus petit segment de signe porteur de signification. Le mot « valise » constitue à lui seul un morphème, car il ne peut pas être décomposé en éléments plus simples ayant une signification. Mais dans le mot « nageuse », par exemple, « nage » et « euse » sont deux morphèmes qui permettent de distinguer « nageuse » des deux mots voisins « nageur » et « rageuse ». De même que le phonème,

le morphème est une entité abstraite susceptible de se réaliser de plusieurs manières dans la chaîne parlée.,

**Morphologie** : Etude de la formation des mots et des variations de forme qu'ils subissent dans la phrase.,

**MS-Windows** : Système d'exploitation de Microsoft pour les PC compatibles IBM. Microsoft Windows est le système d'exploitation le plus utilisé sur PC, principalement parce qu'il est pré-installé sur la plupart des PCs vendus – le prix du PC est alors majoré par le prix de Windows.,

**Multiples** : Les vedettes sont dites «multiples» lorsqu'elles peuvent être complétées d'une façon (nature et forme) définie par RAMEAU. (Voir Cas des multiples).,

**Multitâche** : Se dit d'un système d'exploitation capable de gérer plusieurs processus à la fois.,

**Neurolinguistique** : Etude de l'organisation du langage à partir des troubles occasionnés par des lésions corticales.,

**Non-descripteur** : Voir Terme exclu.,

**Note d'application** : Définit la signification précise d'une vedette et/ou en précise l'emploi.,

**Note d'orientation** : Voir Renvoi général d'orientation.,

**Notice d'autorité-matière** : Permet de gérer les points d'accès aux notices bibliographiques. Constituée d'une vedette (ou forme retenue), de terme(s) exclu(s), et de relations sémantiques. Contient parfois des notes d'application, pour préciser et justifier la forme retenue et son emploi.,

**Noyau** : Élément fondamental d'un système d'exploitation. Il fait la liaison entre le matériel et le reste du système.,

---

**Open Source** : Les termes «Free Software» («logiciel libre») et «Open Source» décrivent tous deux plus ou moins la même catégorie de logiciels, mais correspondent à des conceptions différentes du logiciel et des valeurs qui lui sont associées. Le projet GNU continue d'utiliser le terme Free Software pour exprimer l'idée que la liberté est plus importante que la seule technique.,

**OS** : Abréviation de l'anglais Operating System ; voir Système d'exploitation.,

**Périphérique** : Tout système informatique (i.e. ordinateur) se compose d'un processeur (il exécute les instructions), de mémoires (RAM ou ROM, elles stoquent les instructions et les données) et de périphériques. Ces derniers sont l'interface de l'ordinateur avec le reste du monde. Ils peuvent être très divers, comme un clavier, une souris et un moniteur, mais aussi un capteur de température et un moteur.,

**PC** : Personal Computer ; tout ordinateur personnel, par opposition aux Mainframes, les grands ordinateurs centraux. Depuis quelques années, le terme de « PC » s'applique aux ordinateurs personnels compatibles IBM (c'est ainsi qu'on entend le terme « PC » utilisé pour exclure les Macintosh, qui sont pourtant aussi des ordinateurs personnels).,

**Philosophie du langage** : ou philosophie analytique. Ecole philosophique anglo-saxonne née en Angleterre dans les années 30. Proche de la logique formelle et de la linguistique, la philosophie analytique rejette la prétention à connaître le monde ou à trouver une vérité. Elle s'intéresse plutôt aux énoncés du langage. Les énoncés « analytiques » sont des énoncés logiques qui portent sur le langage, à la différence des énoncés synthétiques qui portent sur les faits. Pour le courant analytique, l'analyse des propositions linguistiques permet d'augmenter la connaissance en clarifiant le sens. Le rôle de la philosophie est donc

d'élucider le sens du langage. Le courant anglais de la philosophie analytique (ou philosophie du langage ordinaire) est représenté par l'école d'Oxford : John L. Austin (1911-1960), Gilbert Ryle (1900-1976), Alfred J. Ayer (1910-1989).

**Phonétique** : La phonétique est une branche de la linguistique qui étudie les sons de la parole (ex : les accents, les formes de prononciation des mots...). Alors que la phonétique étudie des sons (ou « phones »), la phonologie, elle, étudie comment ils sont agencés dans une langue pour former des énoncés. La transcription phonétique s'écrit entre crochets droits.,

**Phonème** : Un phonème est la plus petite unité discrète ou distinctive (c'est-à-dire permettant de distinguer des mots les uns des autres) que l'on puisse isoler par segmentation dans la chaîne parlée. Un phonème est en réalité une entité abstraite, qui peut correspondre à plusieurs sons.,

**Phonologie** : La phonologie est une branche de la linguistique qui étudie comment s'organisent les sons d'une langue afin de former des énoncés. Il ne faut pas la confondre avec la phonétique qui, elle, s'intéresse aux sons eux-mêmes, indépendamment de leur fonctionnement les uns avec les autres. La transcription phonologique se place entre barres obliques : /ra/ est la transcription du mot français rat.,

**Pirate** : Voir « cracker ».,

**Pourriel** : Le terme anglais est *spam*. Il s'agit d'un message électronique commercial non sollicité, c'est-à-dire message électronique vous proposant d'acheter un produit provenant d'une société pour laquelle vous n'avez jamais fait de démarche.,

**Précision** : Voir Qualificatif.,

---

**Pragmatique** : Elle étudie le langage, et particulièrement la parole, en tant qu'il est un instrument d'action sur autrui. Cette discipline envisage les rapports des signes avec ceux qui les emploient en situation (elle se distingue de la syntaxe et de la sémantique). Elle analyse la communication comme une action dans le monde social : elle s'intéresse aux actes de langage et aux principes de conversation. Le philosophe anglais J.-L. Austin est la principale figure de la pragmatique.,

**Processus** : Instance de logiciel, au sens d'entité chargée en mémoire vive et en cours d'exécution.,

**Programme** : Voir logiciel.,

**Psycholinguistique** : Elle s'intéresse aux différents liens entre le langage et les autres fonctions psychiques : mémoire, raisonnement, perception...,

**Qualificatif** : Élément de la vedette (et parfois du terme exclu) qui aide à son identification, en cas d'homonymie ou de polysémie. Placé entre parenthèses à la fin de la vedette concernée.,

**Quasi-synonymes** : Mots de sens voisin, voire ayant entre eux certains rapports d'inclusion, traités comme les synonymes dans le langage d'indexation.,

**Récuratif** : Voir récursif.,

**Réseau** : Tout mode de liaison entre ordinateurs (par exemple Internet) par lequel transitent données, commandes et services.,

**RAM** : Random Access Memory voir Mémoire vive.,

**Rejet** : Voir Inversion.,

**Relation associative** : Relation qui lie deux vedettes dont le sens est proche. Cette relation n'est ni une relation hiérarchique ni une relation de synonymie Les termes liés par cette relation sont appelés termes associés.,

**Relation d'équivalence** : Relation entre une vedette et un terme exclu.,

**Relation hiérarchique** : Relation qui lie deux vedettes dont l'une exprime une notion plus générale (terme générique) et l'autre une notion plus spécifique (terme spécifique).,

**Relation sémantique** : Relation entre vedettes déterminée par leur signification.,

**Renvoi d'équivalence** : Réponse à une demande formulée par un terme exclu, signalant le synonyme ou le quasi-synonyme retenu.,

**Renvoi d'orientation** : Voir Renvoi général d'orientation.,

**Renvoi général d'équivalence** : A partir d'un terme exclu , renvoi qui oriente vers un ensemble de vedettes ou différentes solutions pour remplacer ce terme exclu.,

**Renvoi général d'orientation** : A partir d'une vedette, renvoi qui oriente vers un ensemble de termes utilisables pour l'indexation.,

**Rhétorique** : C'est l'art de convaincre. C'est une discipline qui est liée à la parole de l'avocat, de l'homme politique... La rhétorique ancienne concernait aussi bien les fondements d'une pensée convaincante que les aspects physiques de l'expression (gestes, mimiques...). C'est dans la société classique française que la rhétorique s'est peu à peu cantonnée au seul ornement du discours. La rhétorique classique a disparu de l'enseignement à la fin du XIX<sup>e</sup> siècle. On appelle « nouvelle rhétorique » le courant de recherche initié par Chaïm Perelman (1912-1984). La nouvelle rhétorique vise à énoncer des règles générales de l'argumentation qui ne tiennent pas uniquement compte de la forme du discours, mais aussi des différents types de réception de l'auditoire.,

---

**Sémantique** : Domaine de la linguistique qui étudie le sens, la signification des mots et plus généralement le contenu des signes et leur évolution. Comme adjectif, s'applique à tout ce qui concerne l'attribution d'une signification à un signe ou à un énoncé. Cette notion a été introduite dans la littérature linguistique par Michel Bréal et traduite en anglais en 1897. Le mot provient du grec *Semainein* (« vouloir dire, signifier ») et Bréal insista sur le *sens* au niveau verbal. Lady Welby, sa contemporaine introduisit une théorie des *Significs*, évaluation plus organismique du « sens » de Bréal.,

**Sémiologie (ou sémiotique)** : La sémiologie n'est pas une partie de la linguistique, mais au contraire une discipline englobante. La « science des signes » regroupe non seulement les signes du langage écrit ou parlé, mais aussi les signes visuels (images, symboles). Cette discipline se préoccupe surtout d'analyser la multiplicité des sens contenue dans un même signe. Son fondateur est l'Américain C.S. Peirce (1839-1914),

**Saussure, Ferdinand de (1857-1913)** : Philologue et linguiste suisse, il enseigna à l'École Pratique des Hautes Études à Paris, puis à l'université de Genève. Son *Cours de linguistique générale* (1916) est souvent considéré comme l'acte fondateur de la linguistique comme science. Sa théorie du signe, sa conception de la langue comme système sont les fondements du structuralisme en linguistique. Son œuvre la plus importante (publiée de son vivant) est le *Mémoire sur le système primitif des voyelles dans les langues indo-européennes*, essai de linguistique historique écrit à l'âge de vingt ans. C'est son *Cours de linguistique générale* (1916), reconstitué après sa mort par Charles Bally et Albert Séchehaye à partir des notes de ses élèves, qui le rend célèbre et lui vaut parfois le titre de « père du structuralisme ».,

**Searle John R. (né en 1932)** : Philosophe, enseigne à l'Université de Californie depuis 1959. Il se situe dans le prolongement de l'école d'Oxford et est connu pour avoir systématisé la notion d'acte de langage. A ce titre, il est un auteur important pour la linguistique pragmatique.,

**Serveur** : Système destiné à fournir une ressource. Par exemple un serveur de fichiers stocke les données pouvant être utilisées par de nombreux clients.,

**Shannon, Claude E. et Weaver, Warren (1894-1978)** : Ce sont les fondateurs de la théorie de l'information. Tous deux mathématiciens, ils se sont fait connaître en 1949 par leur Théorie mathématique de la communication.,

**Shareware** : Les Sharewares ou « partagiciels » proposent des démonstrations, bridées ou non, de programmes propriétaires payants, dont le code source n'est pas distribué.,

**Signe linguistique** : Un signe linguistique est une entité formée par la réunion d'un signifié (un concept) et d'un signifiant (une forme sonore ou image acoustique).,,

**Signifié** : Le signifié renvoie au concept, à l'objet ou à l'idée représentée par un signifiant.,

**Signifiant** : Le signifiant est la représentation d'un signe : pour la langue parlée c'est le son produit, pour la langue écrite c'est le graphisme du mot.,

**Sociolecte** : Toute particularité de langage qui peut être considérée comme appartenant à un groupe social ou une catégorie culturelle. Elle peut se placer à trois niveaux différents : phonétique, syntactique, ou lexical. La notion de sociolecte ne signifie pas que leur emploi est réservé aux membres d'un groupe, mais qu'il fait référence à ce groupe.,

---

**Sociolinguistique** : « C'est qu'est-ce que je dis », « Je n'en disconviens pas »...

on ne parle pas de la même façon dans un quartier populaire et dans les couloirs des grandes écoles. L'expression linguistique varie selon les milieux sociaux . Tel est le but de la sociolinguistique : analyser les rapports entre le langage et la société, et s'intéresser aux formes d'expression, différentes selon les groupes sociaux. S'il semble indéniable que les faits sociaux agissent sur le vocabulaire, il n'a pas été prouvé qu'il existe des corrélations directes entre un système phonologique, la morphologie ou la syntaxe et la structure d'une société ou d'une civilisation.,

**Sous-vedette** : Section (sous-élément) d'une vedette, subordonnée à l'élément placé en tête. Ne pas confondre avec la subdivision. On trouve des sous-vedettes principalement dans les vedettes de collectivité, par ex. : France. Conseil d'Etat.,

**Spam** : Voir « pourriel ».,

**Strawson, Peter Frederick (né en 1919)** : Professeur à Oxford, il est un des auteurs majeurs de la philosophie dite du langage ordinaire qui, dans le prolongement de John L. Austin, fonde la pragmatique. Ses travaux fondamentaux portent sur la notion de référence et celle de vérité (*Essays in Philosophical Analysis*, 1952, *Individuals*, 1959). A partir de 1967, il s'intéresse à la philosophie de l'esprit.,

**Structuralisme** : Théorie appliquée à la langue puis étendue à l'ensemble des pratiques sociales. Elle consiste à organiser les faits en un système qui peut se décrire explicitement de façon cohérente, interne et autonome sans référer à autre chose qu'à lui-même. Le système de la linguistique structurale définit la valeur des signes par les séries de leurs oppositions, dans un ensemble pleinement cohérent à un moment

donné (synchronie). Par exemple, un texte doit pouvoir être décrit sans référence sociale ou historique autre que celle qu'il met lui-même en évidence et sans recours aux intentions supposées de son auteur.,

**Subdivision** : Vedette placée après la tête de vedette. Délimite la tête de vedette en apportant des précisions : de sujet : subdivision de sujet de lieu : subdivision géographique de temps : subdivision chronologique de forme du document : subdivision de forme.,

**Subdivision affranchie** : Subdivision applicable à une ou plusieurs catégories de vedettes.,

**Subdivision spécifique** : Subdivision propre à une ou quelques têtes de vedette.,

**Sujet** : Contenu intellectuel du document.,

**Synonymes** : Mots de formes différentes, mais de même sens ou de sens très voisin, rejetés au rang de termes exclus.,

**Syntagme** : Ferdinand de Saussure désigne par syntagme toute combinaison d'unités linguistiques (lettres, mots...) qui se suivent et sont liées entre elles. Par exemple, « le chien » est le syntagme nominal dans la phrase : « le chien aboie ». Ce syntagme, fait de deux éléments, se compose du déterminant « le » et du nom « chien ».,

**Syntaxe** : Etude des règles par lesquelles les mots se combinent et s'agencent entre eux pour former une phrase cohérente. Plus généralement, elle se donne pour but de décrire les relations existantes entre les unités linguistiques et des fonctions qui leurs sont attachés. Alors que la grammaire décrit les règles d'usage, la syntaxe recherche la logique sous-jacente.,

**Syntaxe** : Règles de construction des vedettes-matière.,

---

**Système d'exploitation** : Le système d'exploitation a deux fonctions principales : fournir un niveau d'abstraction pour communiquer avec le matériel, et gérer les ressources. Les systèmes d'exploitation les plus connus sont Windows et MacOS, mais il y en a d'autres comme la famille Unix, dont GNU/Linux fait partie. Un système d'exploitation se compose de deux parties : Le noyau en est la partie centrale, véritable chef d'orchestre de l'ordinateur. Des bibliothèques et des logiciels, complètent le noyau.,

**Tête de vedette** : Premier élément d'une vedette-matière construite. Elle exprime l'essentiel du sujet.,

**TCP/IP** : Protocole Transmission Control Protocole / Internet Protocole ; norme de transmission de données à travers le Réseau : un fichier sera découpé en paquets qui contiennent l'ordre dans lequel ils doivent être réassemblés, leur destination, le moment de leur envoi et l'expéditeur ; les paquets sont envoyés à travers le Réseau et prennent chacun une route qui peut être différente.,

**TCPA** : *Trusted Computing Personal Architecture.*,

**Terme associé** : Voir Relation associative.,

**Terme exclu** : Mot ou groupe de mots à ne pas utiliser pour l'indexation. Fait l'objet d'un renvoi d'équivalence ou d'un renvoi général d'équivalence.,

**Terme générique** : Voir Relation hiérarchique.,

**Terme retenu** : Voir Vedette.,

**Terme spécifique** : Voir Relation hiérarchique.,

**Thesaurus** : De manière générale, il s'agit d'un dictionnaire de termes normalisés et préférentiels, organisés dans des relations sémantiques, utilisés lors de l'indexation des documents.

Dans le langage de description Rameau, c'est une liste organisée de descripteurs et de non-descripteurs, qui obéissent à des règles terminologiques propres et qui sont reliés entre eux par des relations sémantiques.

**Traitement automatique du langage naturel - TALN** : Le langage naturel, c'est le langage humain (par rapport aux langages informatiques). Le TALN désigne donc l'ensemble des techniques visant à décoder et faire reproduire par un ordinateur le langage humain : synthèse vocale, lecture et traduction automatiques, reconnaissance de l'écriture...,

**UNIX** : Famille de puissants systèmes d'exploitations répondant à certains standards. Les versions les plus connues sont UNIX System V, Solaris de Sun, BSD (la Berkeley Standard Distribution), et bien entendu GNU/Linux.,

**Usenet** : Usenet est un espace de discussion, que l'on a pris l'habitude maintenant d'appeler forum, sur lequel n'importe qui, connecté à Internet, peut participer.,

**Vedette** : Forme retenue de la notice d'autorité-matière. Choisie parmi un ensemble de termes équivalents, elle représente une notion ou une entité.,

**Vedette d'autorité** : Voir Vedette.,

**Vedette multiple** : Voir Multiples.,

**Vedette-matière** : Terme polysémique qui désigne tantôt les vedettes contenues dans la liste d'autorité-matière, tantôt la vedette (ou la combinaison de vedettes) établie par le catalogueur pour servir de point d'accès à une notice bibliographique. Pour lever cette ambiguïté, on

---

emploie les termes de : - vedettes, pour celles qui figurent dans la liste d'autorité - vedettes-matière, pour celles qui figurent dans le catalogue bibliographique.,

**Vedette-matière bibliographique** : Voir Vedette-matière.,

**Vedette-matière construite** : Tête de vedette suivie d'une ou plusieurs subdivisions, agencées selon la syntaxe. Figure dans la liste d'autorité et/ou le fichier bibliographique.,

**Vedette-matière d'autorité** : Voir Vedette.,

**Virus** : Programme destiné à effectuer des tâches indésirables pour l'utilisateur. Cela peut aller de la simple écriture d'un message d'avertissement à la transmission de données confidentielles (mots de passe), la destruction de données ou même la mise hors service de certaines parties matérielles de l'ordinateur. Les virus qui utilisent un réseau pour se propager sont appelés « vers ».,

**WAN** : Wide Area Network : réseau de grande envergure ,

**Wittgenstein, Ludwig (1889-1951)** : La pensée de L. Wittgenstein, exprimée dans le *Tractatus logico-philosophicus* (1921), a fortement influencé le courant du positivisme logique du Cercle de Vienne. Notre connaissance du réel est tributaire du langage. Or, le langage repose sur des enchaînements logiques qui sont cohérents entre eux mais ne reflètent que fortuitement la structure du monde. La plupart des discours que nous tenons sur le monde sont, en règle générale, non vérifiables.,

**WWW** : « World Wide Web » (Toile d'araignée Mondiale) : système de présentation de données reposant sur le standard HTTP et HTML, et mis au point au début des années 90 au CERN pour la présentation des données scientifiques. Le WWW s'est depuis ouvert aux usages publics et commerciaux, popularisant ainsi l'usage d'Internet.,

**X** : Abréviation de X Window System : protocole d'interface graphique permettant à un client de se connecter à travers le réseau à un serveur,

# LES ANNEXES

---



ANNEXE A

## **LA GED ET LA CHAÎNE DE NUMÉRISATION**

---

G rard Dupouitier  
**Technologie  
 de la GED**  
 l' dition  lectronique

  Herm s, Paris, 1994  
 Editions Herm s  
 14, rue Lantier  
 75017 Paris  
 ISBN 2-86601-432-4

**Description fonctionnelle et exemples de Ged**

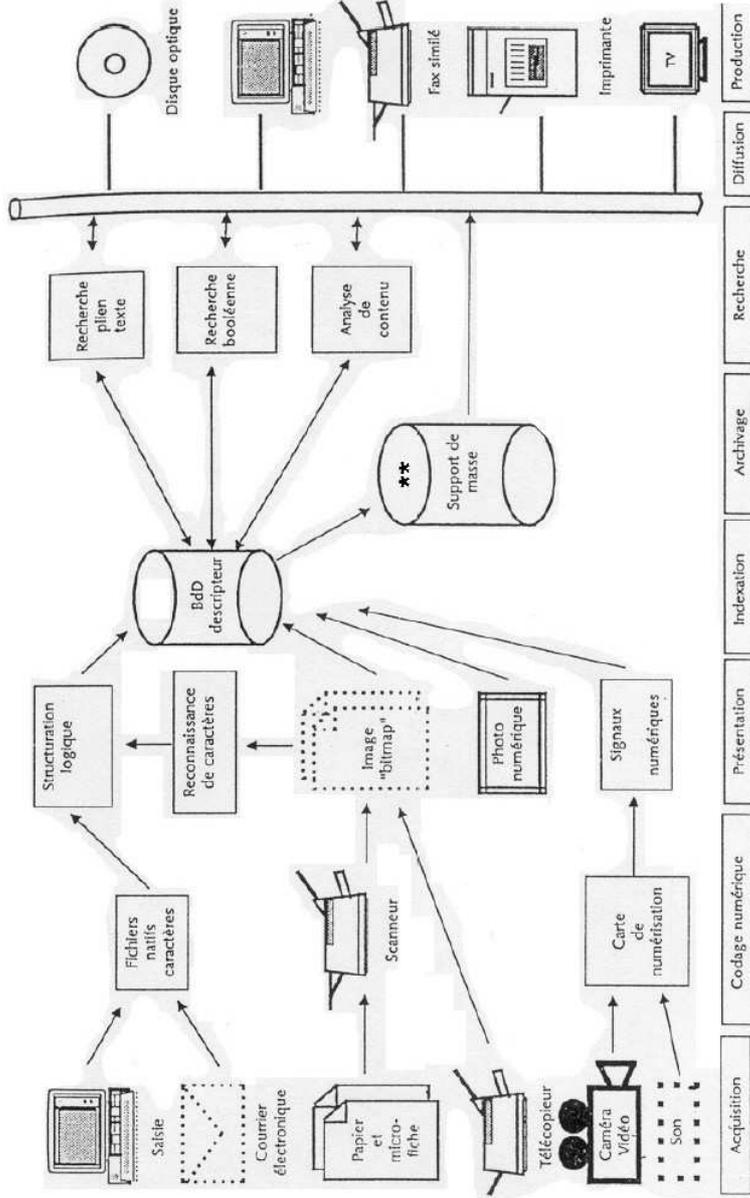


Figure 2.1. Cartographie des techniques selon les  tapes fonctionnelles majeures des processus de GED

\*\* Etape non conserv e si publication directe sur le web

--> passage direct au support de masse

ANNEXE B

## LES PROJETS DE NUMÉRISATION

---

Projet	Pays/Langue(s)	Création	Site
ABU Association des Bibliophiles Universels	F/F	1993	<a href="http://abu.cnam.fr/">http://abu.cnam.fr/</a>
Lamarck	F/F	2001 revu en 2003	<a href="http://www.lamarck.net">http://www.lamarck.net</a>
Gallica	F/F	1998	<a href="http://gallica.bnf.fr">http://gallica.bnf.fr</a>
BNF	F/F	1996	<a href="http://www.bnf.fr">http://www.bnf.fr</a>
ARTFL  <i>Le Trésor de la Langue Française</i> Frantext	US-F/F  US-F/F US-F/F	1999  2001 2003	<a href="http://www.lib.uchicago.edu/efts/ARTFL/databases/TLF/">http://www.lib.uchicago.edu/efts/ARTFL/databases/TLF/</a>  <a href="http://www.atilf.fr/frantext.htm">http://www.atilf.fr/frantext.htm</a>
TLFi	F/F	2001	<a href="http://www.atilf.fr/tlfi.htm">http://www.atilf.fr/tlfi.htm</a>
Bibliothèque électronique de Lisieux	F/F	1996	<a href="http://www.bmlisieux.com">http://www.bmlisieux.com</a>
Persée	F/F	2003	<a href="http://www.persee.fr">http://www.persee.fr</a>
Cyberthèses	F/F	2000	<a href="http://www.cybertheses.org">http://www.cybertheses.org</a>
<i>CoLisciences</i>	F/F	2002	<a href="http://www.colisciences.net">http://www.colisciences.net</a>

ANNEXE C

**LE PROJET *CoLiSciences***

---

Le texte qui suit provient d'un article paru dans l'édition du 21 avril 2004 du journal « *Le Monde* ». Il présente de manière concise et synthétique le projet *Colisciences* et ses objectifs.

### **Comment les lecteurs naviguent dans un recueil de textes informatisés**

En s'attachant à la création d'illusions efficaces, les concepteurs de réalité virtuelle tentent de comprendre comment le cerveau humain s'approprié le monde qui l'entoure. D'autres chercheurs explorent le processus inverse et étudient comment une image mentale est élaborée par le cerveau à partir d'un univers déjà virtuel par essence.

C'est le cas de Georges Vignaux. Logicien, cognitiviste, linguiste et directeur du laboratoire « communication et politique du CNRS », il effectue cette recherche à partir du texte informatisé. « *Nous avons, explique-t-il, constitué un corpus d'environ 5 500 pages rassemblant les écrits des biologistes et des naturalistes de la seconde moitié du XIX<sup>e</sup> siècle, tels que Jean-Baptiste Lamarck, Georges Cuvier, Etienne Geoffroy Saint-Hilaire, Louis Pasteur ou Claude Bernard.* »

Grâce aux techniques bien connues sur Internet des liens hypertextes et des interrogations par requêtes, les utilisateurs de Colisciences ([www.colisciences.net](http://www.colisciences.net)) vont partir à la découverte du contenu de cette base de textes. Au départ, ils ignorent tout de sa taille, du classement et de l'organisation des liens. Contrairement au livre, dont le sommaire et l'index fournissent des indices sur la quantité d'informations qu'il contient et sur son organisation, une base de données de textes reste essentiellement masquée. C'est son exploration qui, progressivement, crée une image de son contenu dans le cerveau de l'utilisateur.

#### **UNE EXPÉRIENCE COGNITIVE**

---

En collaboration avec le laboratoire des usages de la Cité des sciences de La Villette à Paris, Georges Vignaux va étudier ce mécanisme d'appropriation des textes de Colisciences. *Il s'agit d'une recherche fondamentale sur le processus de la lecture et la construction du sens*, indique-t-il. Pour la mener, les postes de consultation vont espionner leurs visiteurs, enregistrant chaque requête et chaque clic de souris sur un lien hypertexte. Un oculomètre va même suivre le regard des utilisateurs sur l'écran. De quoi fournir une radiographie de l'usage de la base de données, dont les apports devraient se révéler passionnants. Car, au-delà de l'application à la découverte du corpus de Colisciences, c'est bien à une expérience cognitive inédite que s'attellent Claude Vignaux et son équipe.

Dans son livre *Du signe au virtuel* (Editions du Seuil, 2003), il définit le changement induit par les nouvelles technologies dans notre rapport au monde. « *C'est essentiellement la disparition d'un certain rapport direct à l'outil, lequel va devenir de plus en plus médiatisé. Dans la plupart des situations sociales, professionnelles et pratiques, les activités humaines cessent d'être concrètement visibles : elles deviennent virtuelles* », écrit-il.

Mais ces travaux sur la compréhension des modes d'appropriation sont difficiles à gérer dans un contexte de production incontrôlée d'information. La solution ? « *Créer une niche, un monde à nous* », répond Georges Vignaux. Et pour cela la maîtrise des processus cognitifs jouera un rôle essentiel.



ANNEXE D

**LES *referers* CoLiS ciences**

---

La figure D.1 permet d'avoir un aperçu de la page d'administration du site *CoLiSciences* ouverte sur la liste des « référants ». Cette liste n'est pas triée, il faut donc l'exporter pour pouvoir la manipuler et en extraire une liste résumée ou triée suivant un critère particulier.

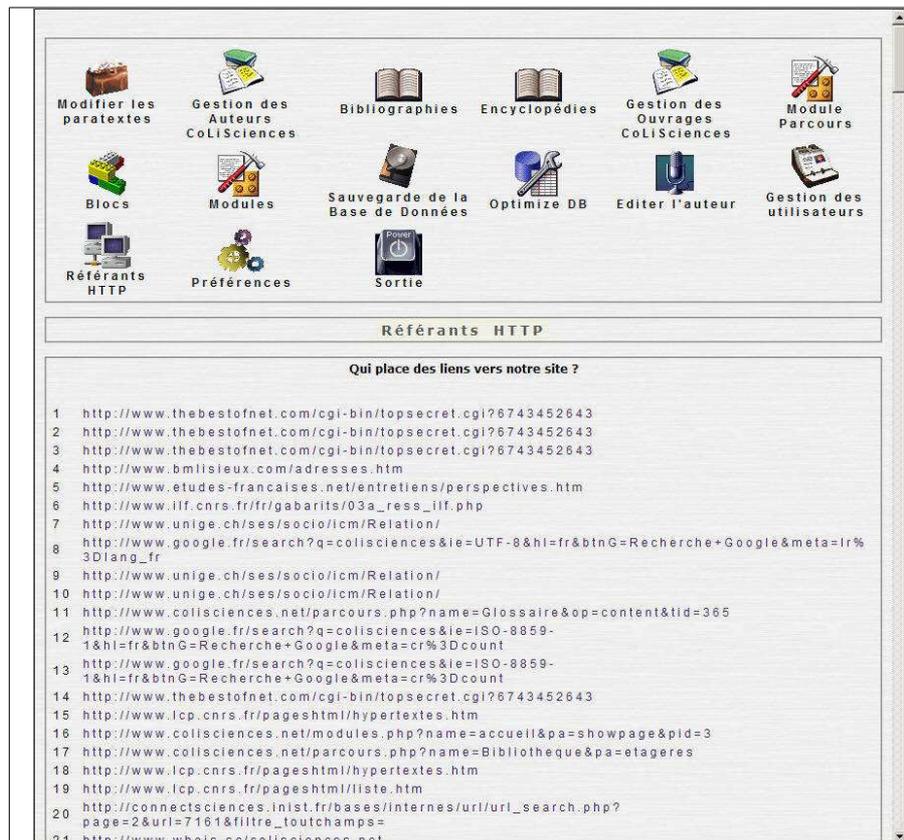


FIG. D.1: La page des référants dans l'administration *CoLiSciences*

C'est ce type de liste, triée par fréquence d'apparition de l'adresse qui est présenté à la suite. Le site depuis lequel *CoLiSciences* a été appelé est d'abord celui de la Bibliothèque de Lisieux, suivi par le moteur de recherche *Google*.

- 36 <http://www.bmlisieux.com/adresses.htm>
- 33 <http://www.google.fr/search?sourceid=navclient&hl=fr&ie=UTF-8&q=colisciences>
- 28 <http://www.whois.sc/colisciences.org>

---

16 [http://www.thebestofnet.com/cgi-bin/topsecret.  
cgi?6743452643](http://www.thebestofnet.com/cgi-bin/topsecret.cgi?6743452643)

14 <http://www.lcp.cnrs.fr/pageshtml/liste.htm>

10 [http://www.science.gouv.fr/index.php?action=  
page&target=select&rubrique=5](http://www.science.gouv.fr/index.php?action=page&target=select&rubrique=5)

10 <http://www.whois.sc/colisciences.com>

9 [http://www.ilf.cnrs.fr/fr/gabarits/03a\\\_ress\  
\\\_ilf.php](http://www.ilf.cnrs.fr/fr/gabarits/03a\_ress\_ilf.php)

9 <http://www.miragorobot.com/scripts/mrinfo.asp>

8 <http://www.liensutiles.org/sciences.htm>

6 <http://www.netcraft.com/survey/>

6 [http://www.science.gouv.fr/index.php?action=  
page&target=select&rubrique=4](http://www.science.gouv.fr/index.php?action=page&target=select&rubrique=4)

6 <http://www.unige.ch/ses/socio/icm/Relation/>

5 [http://www.docforum.tm.fr/journe\\\_rdv/je281102.  
htm](http://www.docforum.tm.fr/journe\_rdv/je281102.htm)

4 [http://www.science.gouv.fr/index.php?action=  
page&target=select&rubrique=25](http://www.science.gouv.fr/index.php?action=page&target=select&rubrique=25)

3 [http://lcp.damesme.cnrs.fr/pageshtml/  
hypertextes.htm](http://lcp.damesme.cnrs.fr/pageshtml/hypertextes.htm)

3 <http://lcp.damesme.cnrs.fr/pageshtml/liste.htm>

3            <http://liris.cnrs.fr/~veglin/atelierSDN/Vignaux.html>

3            <http://localhost/dglfweb//theme.php3?cle=informatique>

3            <http://perso.wanadoo.fr/labyrinthe/liens.html>

3            <http://www.google.com/search?q=colisciences\&ie=UTF-8\&oe=UTF-8>

3            <http://www.ird.fr/sais/cgi/TrSaS?procedure=portail.AfficherObjets\&action=GrSELECTION\&nomObjet=sites\&>

3            <http://www.lcp.cnrs.fr/pageshtml/hypertextes.htm>

3            [http://www.mylinea.com/tic/sciences\\\_et\\\_mathematiques/historique/](http://www.mylinea.com/tic/sciences\_et\_mathematiques/historique/)

3            <http://www.science.gouv.fr/dispatch.php?lien=509>

2            <http://bmlisieux.com/adresses.htm>

2            <http://colisciences.univ-paris13.fr/>

2            [http://connectsciences.inist.fr/bases/internes/url/url\\\_consult.php?decor=1\&page=3\&domain=4\&sdomain=](http://connectsciences.inist.fr/bases/internes/url/url\_consult.php?decor=1\&page=3\&domain=4\&sdomain=)

2            <http://www.antibot.net>

2            <http://www.etudes-francaises.net/entretiens/perspectives.htm>

---

2           [http://www.radiofrance.fr/chaines/  
france-culture2/emissions/science\\\_frictions/  
fiche.php?diffusion\\\_id](http://www.radiofrance.fr/chaines/france-culture2/emissions/science\_frictions/fiche.php?diffusion\_id)

2           <http://www.silworld.com.tw/home>

1           [http://143.126.28.29/dglfweb/theme.php3?cle=  
informatique](http://143.126.28.29/dglfweb/theme.php3?cle=informatique)

1           <http://194.254.163.88:8080/colislab>

1           <http://194.254.163.88:8080/colislab/>

1           [http://65.54.184.250/cgi-bin/linkrd?\\\_lang=  
FR\&lah=b5ca8d827cc0c42b0c738679c1ae4d82\&lat=  
1091200993\&hm\\\_](http://65.54.184.250/cgi-bin/linkrd?\_lang=FR\&lah=b5ca8d827cc0c42b0c738679c1ae4d82\&lat=1091200993\&hm\_)

1           [http://album.revues.org/categorie.php?id=79\  
&step=1](http://album.revues.org/categorie.php?id=79\&step=1)

1           [http://colisciences.univ-paris13.fr:8080/dglflf/  
News\\\_Item.2004-09-01.9587530106/newsitem\\\_view?  
portal](http://colisciences.univ-paris13.fr:8080/dglflf/News\_Item.2004-09-01.9587530106/newsitem\_view?portal)

1           [http://fr.srd.yahoo.com/S=9159818/K=  
colisciences/v=2/l=WS1/R=1/\\*-http://www.  
colisciences.net/](http://fr.srd.yahoo.com/S=9159818/K=colisciences/v=2/l=WS1/R=1/*-http://www.colisciences.net/)

1           <http://jmvanel.free.fr/bookmarks.html>

1           [http://listes.cdeacf.ca/mhonarc/techno/msg00079.  
html](http://listes.cdeacf.ca/mhonarc/techno/msg00079.html)

1           <http://localhost:8080/colislab>

1           <http://search.msn.fr/results.aspx?FORM=SMCRT&q=colisciences>

1           [http://search.netscape.com/ns/boomframe.jsp?query=colisciences&page=1&offset=0&result\\\_url=redir%3Fsr](http://search.netscape.com/ns/boomframe.jsp?query=colisciences&page=1&offset=0&result\_url=redir%3Fsr)

1           [http://webmail4.wanadoo.fr/wanadoo/inbox\\\_read\\\_main.html?IDMSG=106&](http://webmail4.wanadoo.fr/wanadoo/inbox\_read\_main.html?IDMSG=106&)

1           <http://www.bmlisieux.com/colloque/colloque.htm>

1           [http://www.ceram.edu/recherche/R3\\\_public/R3-1\\\_fr/R3-1-4\\\_chercheur/R3-1-4-2\\\_axe/R3-1-4-2-2\\\_CV/F3-1-4-](http://www.ceram.edu/recherche/R3\_public/R3-1\_fr/R3-1-4\_chercheur/R3-1-4-2\_axe/R3-1-4-2-2\_CV/F3-1-4-)

1           <http://www.etudes-francaises.net/colloques/lisieux2002/colis.htm>

1           <http://www.fabula.org/vlib/affiche.php?id=575>

1           <http://www.fabula.org/vlib/affiche.php?id=575>

1           <http://www.fabula.org/vlib/categorie.php?id=52&ordre=titre&type=\&public=\&all=1>

1           <http://www.fabula.org/vlib/categorie.php?id=52&ordre=titre&type=\&public=\&step=1>

1           <http://www.fastcrawl.com>

1           <http://www.faxo.com>

1           <http://www.fsttrk.org/guest.html>

---

1 <http://www.google.be/search?hl=fr\&ie=UTF-8\&q=colisciences\&spell=1>

1 <http://www.google.lu/search?hl=de\&ie=UTF-8\&q=www.colisciences.net\&btnG=Suche\&meta=>

1 <http://www.inter7.com>

1 <http://www.lcp.cnrs.fr/pageshtml/hypertextes.htm>

1 <http://www.lutecium.fr/stp/affiche-vignaux.html>

1 <http://www.novaris.com.au/phpbb2>

1 <http://www.science.gouv.fr/index.php?action=form\&target=ressourceweb>

1 <http://www.science.gouv.fr/index.php?action=page\&target=select\&select\&rubrique=2>

1 <http://www.sciences.gouv.fr/index.php?action=page\&target=select\&rubrique=25>

1 <http://www.strikersacademy.com/strikeracademy.html>

1 <http://www.support.naym.ru>

1 <http://www.whois.sc/colisciences.com>

1 <http://www.whois.sc/colisciences.com>

1 <http://www.whois.sc/colisciences.com>

1 <http://www.whois.sc/colisciences.com>

1 <http://www.temario.net/TEMA-SCIENCES.HTM>



ANNEXE E

**MESSAGE AU PRÉSIDENT DU CONSEIL EUROPÉEN**

---

Message<sup>191</sup> adressé à M. Jean-Claude JUNCKER, Président en exercice du Conseil européen et à M. José Manuel DURAO BARROSO, Président de la Commission européenne, par M. Jacques CHIRAC, Président de la République, M. Aleksander KWASNIEWSKI, Président de la République de Pologne, M. Gerhard SCHROEDER Chancelier de la République Fédérale d'Allemagne, M. Silvio BERLUSCONI, Président du Conseil italien, M. José Luis Rodriguez ZAPATERO Président du Gouvernement espagnol et M. Ferenc GYURCSANY, Premier ministre de la République de Hongrie.

Paris, Varsovie, Berlin, Rome, Madrid, Budapest, le 28 avril 2005

Monsieur le Président,

Le patrimoine des bibliothèques européennes est d'une richesse et d'une diversité sans égales. Il exprime l'universalisme d'un continent qui, tout au long de son histoire, a dialogué avec le reste du monde. Pourtant, s'il n'est pas numérisé et rendu accessible en ligne, ce patrimoine pourrait, demain, ne pas occuper toute sa place dans la future géographie des savoirs.

C'est la raison pour laquelle nous souhaitons prendre appui sur les actions de numérisation déjà engagées par nombre de bibliothèques européennes pour les mettre en réseau et constituer ainsi ce qu'on pourrait appeler une bibliothèque numérique européenne, c'est-à-dire une action concertée de mise à disposition large et organisée de notre patrimoine culturel et scientifique sur les réseaux informatiques mondiaux.

Il s'agit de coordonner et de soutenir des initiatives aujourd'hui dispersées afin de les mettre en cohérence et de les amplifier. L'Union pourrait à la fois fournir le

---

191. Disponible sur le site de la Présidence de la République Française, à l'adresse [http://elysee.fr/elysee/francais/interventions/lettres\\_et\\_messages/2005/avril/message\\_a\\_propos\\_de\\_la\\_creation\\_d\\_une\\_bibliotheque\\_numerique\\_europeenne.29633.html](http://elysee.fr/elysee/francais/interventions/lettres_et_messages/2005/avril/message_a_propos_de_la_creation_d_une_bibliotheque_numerique_europeenne.29633.html)

---

cadre d'une concertation entre les institutions concernées, et apporter sa contribution à la solution des problèmes à surmonter : sélection des fonds numérisés de manière à éviter les redondances, mise au point des outils de recherche et de présentation les plus performants, actions de recherche-développement dans le domaine des techniques de numérisation à mettre en œuvre, réflexion à engager sur les attentes et les besoins des utilisateurs, etc.

Afin de progresser dans l'examen de la faisabilité et des modalités de mise en œuvre d'un tel projet, il nous semblerait opportun que celui-ci puisse, à brève échéance, faire l'objet d'un débat entre les ministres de la Culture d'une part, de la Recherche d'autre part, à la lumière d'une première communication de la Commission et des témoignages et propositions que pourraient soumettre les Etats dont les institutions compétentes sont d'ores et déjà engagées dans cette direction.

Nous transmettons copie de ce courrier à tous les membres du Conseil européen. Veuillez agréer, Monsieur le Président, l'assurance de notre haute considération.

Signé :

Jacques CHIRAC, Président de la République française,

Aleksander KWASNIEWSKI, Président de la République de Pologne,

Gerhard SCHROEDER Chancelier de la République Fédérale d'Allemagne,

Silvio BERLUSCONI, Président du Conseil italien,

José Luis Rodriguez ZAPATERO Président du Gouvernement espagnol,

Ferenc GYURCSANY, Premier ministre de la République de Hongrie.



ANNEXE F

**LES *statistiques* <sup>C</sup>*o*<sup>L</sup>*s* *ciences***

---

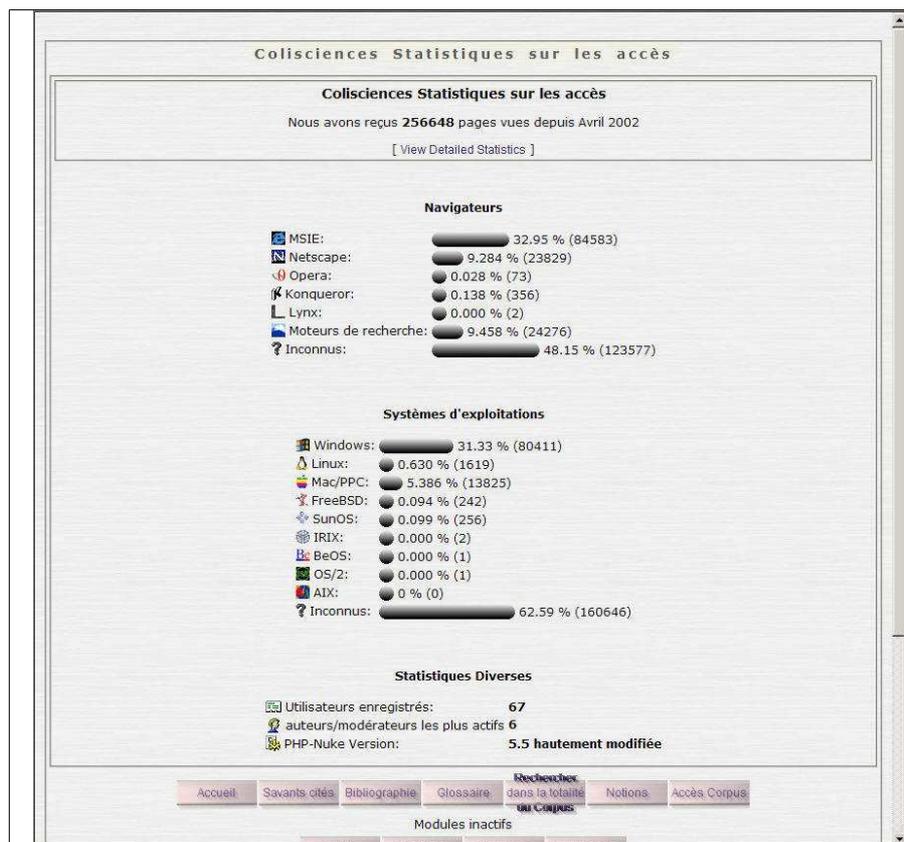


FIG. F.1: Les statistiques de connexion

ANNEXE G

**MANUEL DE L'ADMINISTRATEUR *CoLiS*ciences**

---

## G.1 Manuel de l'administrateur *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*

### A. Introduction en forme d'avertissement

Ce manuel a été réalisé avec la première version du site *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, le graphisme des copies d'écran est différent de celui qui est utilisé aujourd'hui. Les fonctionnalités sont strictement identiques même si au final les textes ne sont pas présentés aux mêmes endroits et de la même manière. Toutefois, il reste un dernier point à clarifier pour mieux comprendre comment le dispositif a évolué. La structure du site a elle aussi été modifiée, l'accès aux ouvrages par exemple ne se fait plus depuis la page d'accueil, ce remaniement de la structure peut laisser croire de plus profond changements, mais en fait il n'en est rien, les fonctions sont les mêmes et ce n'est que l'architecture hypertextuelle des liens entre eux qui peut donner cette impression d'un changement plus profond.

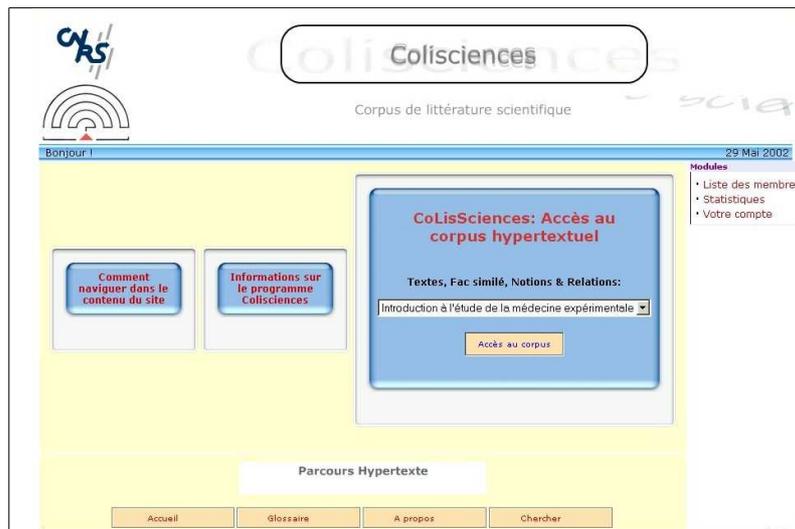
### B. Modification de la page d'accueil

La page d'accueil de la première version était divisée en 3 zones : une zone à gauche, une au milieu et une à droite que nous pouvons reconnaître sur la figure G.1 et retrouver les fonctionnalités suivantes présentées de gauche à droite :

- Comment naviguer dans le contenu du site.
- Informations sur le programme *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*.
- *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* : Accès au corpus hypertextuel.

Le dernier choix permet donc d'accéder directement à la liste complète des ouvrages du corpus, triée par ordre alphabétique des titres. Cette liste est présentée sous la forme d'un objet particulier appelé *liste de choix déroulante*, ce qui en rend la manipulation pas très aisée dès que la liste dépasse une dizaine d'entrées.

La figure G.2 présente quant à elle la nouvelle page d'accueil, qui ne comporte plus la liste des ouvrages disponibles et a vu les paratextes s'organiser autour de l'entrée vers le corpus. Un long texte de description et de présentation est lui aussi

FIG. G.1: La page d'accueil *CoLisSciences* de la version 1

venu accueillir le visiteur pour essayer de lui expliquer rapidement ce qu'il doit s'attendre à trouver sur ce site. Pour obtenir la liste des ouvrages et commencer une navigation, il faut maintenant cliquer sur le lien central *Accès corpus* qui mène à une page où il est possible de trier la liste de ouvrages suivant plusieurs critères : par auteur, par discipline ou par domaine. Il est aussi devenu possible de lancer une navigation notionnelle sans pour autant avoir choisi un ouvrage particulier au préalable. Les quatre liens proposés par la page d'accueil sont :

- Présentation du site
- Le projet *CoLisSciences*
- Acteurs et Soutiens
- Mode d'emploi

En bas de page sont aussi proposées deux icônes qui permettent d'aller directement soit à la liste des paratextes, téléchargeables sous la forme de fichiers au format PDF, soit la liste des figures « remarquables » du corpus.

Nous allons voir comment ajouter et modifier ces entrées de menu ainsi que les documents qui s'y rattachent. Cette procédure est restée identique depuis la première version du site. Pour toutes ces opérations, il faut tout d'abord se connecter sur *CoLisSciences* au travers de l'interface d'administration, ce qui se fait en ouvrant la page

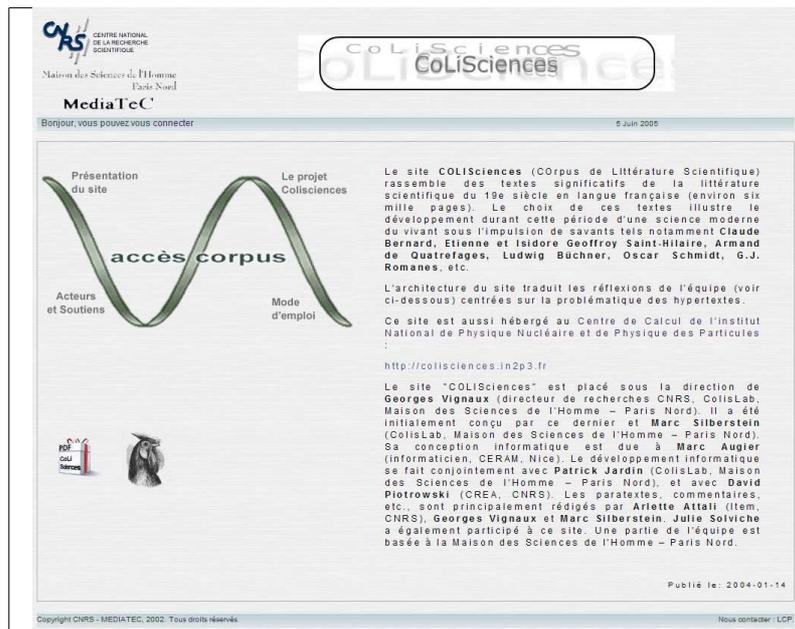


FIG. G.2: La page d'accueil actuelle

située à l'adresse : <http://www.colisciences.net/admin.php> et qui est visible sur la figure G.3.



FIG. G.3: La page de connexion de l'administrateur

Il faut donc avoir reçu d'un administrateur déjà existant un identifiant et le mot de passe qui lui est associé. Après avoir entré son identifiant (*nom d'administrateur*) et le mot de passe associé dans les champs prévus à cet effet sans erreur c'est la page que nous voyons sur la figure G.4 qui est affichée.

La figure G.5 présente quant à elle la même page mais cette fois sur la dernière version du site. L'interface a été nettoyée des options devenues inutiles parce qu'elles provenaient de PhpNuke que nous avons adapté pour nos besoins et correspondaient



FIG. G.4: Les icônes du menu d'administration

à des modules fonctionnels que nous n'avons pas utilisés. En revanche les fonctions spécifiques *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>* ont été revues, enrichies et améliorées.

Nous voyons sur cet écran regroupées toutes les options disponibles pour un administrateur et représentées sous la forme d'une icône et d'un court texte. Sur la droite dans la première version, en bas de l'écran dans la dernière, sont affichés plusieurs menus qui sont disponibles sur toutes les pages de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*, soit pour les utilisateurs enregistrés (catégorie dont les administrateurs sont une partie), soit pour les administrateurs sans pour autant être des fonctions de mise à jour. Nous trouverons là par exemple l'accès aux statistiques et le lien vers la partie administration qui permet de naviguer entre le corpus et la zone visible du site et les paramètres ajustables de la zone invisible du site.

Dans le cas qui nous intéresse ici, c'est-à-dire pour accéder à la mise à jour de la page d'accueil, il faut choisir et cliquer sur l'icône « Manager le contenu » qui est la quatrième en partant de la gauche, sur la première ligne dans la première version. Dans la dernière version, cette fonction est beaucoup mieux mise en évidence puisqu'il s'agit maintenant de la première icône du menu, et qu'elle indique « modifier les paratextes ».

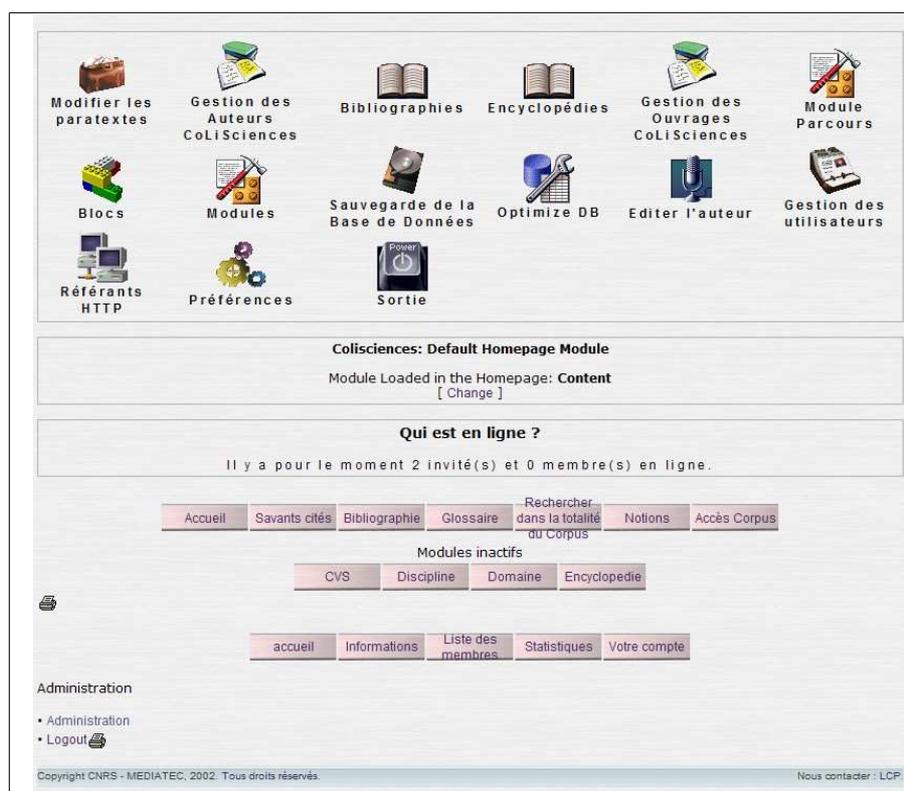


FIG. G.5: Les icônes du menu d'administration

Après avoir cliqué sur l'icône voulue et une fois la nouvelle page affichée, les mêmes icônes réapparaissent, parce qu'une fois connecté comme administrateur le menu regroupant toutes les fonctions d'administration est toujours disponible sur le haut de toutes les pages de cette partie de gestion du site. En revanche, cette fois elles sont suivies des options de gestion des contenus comme cela est visible sur la figure G.6.

Ce contenu est présenté dans un tableau, chaque ligne correspondant à une page de contenu, identifiée par son titre. Les autres informations attachées à ce contenu sont :

- Le statut qui permet de préparer une page et de la garder masquée (inactive) aux yeux des visiteurs tant qu'elle n'est pas prête (active).
- La catégorie, qui est en fait l'entrée de menu sous laquelle ce contenu apparaîtra dans la page d'accueil.

- Les fonctions disponibles, soit la modification et la suppression pour agir sur le contenu lui même et l'activation désactivation pour mettre en ligne/hors ligne.

Nous pouvons aussi vérifier que les catégories que nous voyons affichées dans ce menu correspondent très exactement aux entrées de menu que nous venons de voir sur la page d'accueil de *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*.

Dans la dernière version, la page de gestion des contenus est restée la même tout en s'adaptant encore une fois de plus près aux besoins des administrateurs. Elle est présentée complète sur la figure G.7 et nous pouvons retrouver encore une fois le menu icônique, la liste de page et les fonctions de mise à jour mais cette fois les intitulés ont été adaptés à *C<sup>o</sup>L<sup>i</sup>S<sup>ciences</sup>*. Ils sont donc beaucoup plus précis, par exemple c'est « Menu » qui est utilisé à la place de « Catégories » qui n'était pas très compréhensible à ce niveau. Une nouvelle entrée propose d'accéder directement à l'édition du « texte d'annonce de la page d'accueil ».

**Colisciences**  
Corpus de littérature scientifique

Bonjour ! 29 Mai 2002

**Menu Administration Administration des Bannières**

- Ajouter l'article
- Sauvegarde de la Base de Données
- Blocs
- Manager le contenu
- Editer l'auteur
- Gestion des utilisateurs
- Référents HTTP
- Messages
- Modules
- Newsletter
- Optimize DB
- Module Parcours
- Préférences
- Propositions
- Affectation des Sujets
- Liens Web
- Sortie

**Modules**

- Liste des membres
- Statistiques
- Votre compte

**Modules inactifs**  
(Pour des tests administratifs)

- News
- Sections
- Submit News
- Web Links
- wysiwyg editor

**Administration**

- Administration
- NEW Story
- Change Survey
- Content
- Logout

**Contenu en attente**

- Propositions: 1
- Comptes rendus en attente: 0
- Liens en attente: 0
- Téléchargements: 0
- \_WAITINGUSER: 0

**Manager le contenu**

Titre	Statut actuel	Catégorie	Fonctions
Du texte à l'hypertexte	Actif	Informations sur le programme Colisciences	[ Editer   Désactiver   Supprimer ]
Introduction à la médecine expérimentale de Claude Bernard	Actif	Informations sur le programme Colisciences	[ Editer   Désactiver   Supprimer ]
Pourquoi Claude Bernard	Actif	Informations sur le programme Colisciences	[ Editer   Désactiver   Supprimer ]
Hypertexte et cognition	Actif	Comment naviguer dans le contenu du site	[ Editer   Désactiver   Supprimer ]

**Ajouter une nouvelle catégorie**

Titre: Acteurs et Partenaires du programme

Description: gauche

**Editer la catégorie**

Catégorie: Acteurs et Partenaires du programme

**Ajouter une nouvelle page**

Titre: Les Acteurs

Catégorie: Acteurs et Partenaires du programme

Sous-Titre: LCP - Médiatéc

Haut du Texte: LCP Médiatéc - Unice|

Texte de la page:  
Si vous voulez plusieurs pages, vous pouvez écrire <!--pagebreak--> là où vous voulez un saut de page.

FIG. G.6: La gestion du contenu

**Modifier les paratextes**

**Gestion des Auteurs CoLiSciences**

**Bibliographies Encyclopédies**

**Gestion des Ouvrages CoLiSciences**

**Module Parcours**

**Blocs**

**Modules**

**Sauvegarde de la Base de Données**

**Optimize DB**

**Editer l'auteur**

**Gestion des utilisateurs**

**Référants HTTP**

**Préférences**

**Sortie**

**Modifier les paratextes**

La page d'accueil			
Titre	Statut actuel	Menu	Fonctions
Pourquoi Claude Bernard	Actif	Le projet CoLiSciences	[ Editer   Désactiver   Supprimer ]
Typologie des relations	Actif	Mode d'emploi	[ Editer   Désactiver   Supprimer ]
Soutiens Institutionnels	Actif	Acteurs et Soutiens	[ Editer   Désactiver   Supprimer ]
Présentation du site	Actif	Présentation du site	[ Editer   Désactiver   Supprimer ]
Glossaire et notions : leurs rôles	Actif	Mode d'emploi	[ Editer   Désactiver   Supprimer ]
Acteurs et Partenaires du programme	Actif	Acteurs et Soutiens	[ Editer   Désactiver   Supprimer ]
Le projet Colisciences	Inactif	Le projet CoLiSciences	[ Editer   Activer   Supprimer ]
Mode d'emploi	Actif	Mode d'emploi	[ Editer   Désactiver   Supprimer ]
Processus d'ajustement et opérations langagières	Actif	Cognition et langage	[ Editer   Désactiver   Supprimer ]
Les savants cités	Actif	Mode d'emploi	[ Editer   Désactiver   Supprimer ]
Hypertexte : de l'accès aux documents à leur structuration	Actif	Hypertexte	[ Editer   Désactiver   Supprimer ]
L'hypertexte	Actif	Hypertexte	[ Editer   Désactiver   Supprimer ]
Réflexions sur l'hypertexte	Actif	Le projet CoLiSciences	[ Editer   Désactiver   Supprimer ]
Cognition et langage	Actif	Le projet CoLiSciences	[ Editer   Désactiver   Supprimer ]
Réflexions sur la numérisation	Actif	Le projet CoLiSciences	[ Editer   Désactiver   Supprimer ]
Conversion de documents imprimés	Actif	Réflexions sur la numérisation	[ Editer   Désactiver   Supprimer ]
Construction d'un système hypertextuel sur le Web	Actif	Hypertexte	[ Editer   Désactiver   Supprimer ]
Opérations langagières, opérations cognitives un essai d'analyse	Actif	Cognition et langage	[ Editer   Désactiver   Supprimer ]
Reconnaissance automatique de l'écriture et du document	Actif	Cognition et langage	[ Editer   Désactiver   Supprimer ]
De l'hypertextualité en général	Actif	Réflexions sur la numérisation	[ Editer   Désactiver   Supprimer ]
La notion de << collection >>:	Actif	Hypertexte	[ Editer   Désactiver   Supprimer ]
Représentation de la connaissance dans un corpus de textes numérisés	Actif	Hypertexte	[ Editer   Désactiver   Supprimer ]

**Ajouter un nouveau menu sur la page d'accueil**

**Titre:**

**Le texte d'annonce de la page d'accueil**

Titre	Statut actuel	Menu	Fonctions
Annonce	Actif	Aucun	[ Editer   Désactiver   Supprimer ]

**Les autres textes**

Titre	Statut actuel	Menu	Fonctions
L'ensemble des paratextes à télécharger	Actif	Aucun	[ Editer   Désactiver   Supprimer ]

**Editer le menu**

**Menu:** Acteurs et Soutiens

**Ajouter une nouvelle page**

**Titre:**  **Menu:** Aucun

**Sous-Titre:**

**Haut du Texte:**   
**Texte de la page:**   
Si vous voulez plusieurs pages, vous pouvez écrire <!--pagebreak--> là où vous voulez un saut de page. **Bas du Texte:**   
**Signature:**

**Activer cette page?**  
 Oui  Non

Accueil
Savants cités
Bibliographie
Glossaire
Rechercher dans la totalité du Corpus
Notions
Accès Corpus

**Modules inactifs**  
CVS
Discipline
Domaine
Encyclopedie

accueil
Informations
Liste des membres
Statistiques
Votre compte

**Administration**  

- Administration
- Logout

Copyright CNRS - MEDIATEC, 2002. Tous droits réservés.
Nous contacter : LCP

FIG. G.7: La gestion du contenu  
581

A la suite sur la même page nous trouvons une zone « Ajouter une nouvelle catégorie » représenté dans la première version sur la figure G.8 et qui permet de créer une nouvelle entrée sur la page d'accueil.

The screenshot shows a web interface for managing categories. At the top, there is a table of existing categories with columns for title, status, description, and actions. Below this is a form titled 'Ajouter une nouvelle catégorie' with a 'Titre:' field containing 'Acteurs et Partenaires du programme' and a 'Description:' text area containing 'gauche'. An 'Ajouter' button is below the form. Below the form is another section titled 'Editer la catégorie' with a dropdown menu for 'Catégorie:' set to 'Acteurs et Partenaires du programme' and an 'Editer' button.

Titre	Statut	Description	Actions
Introduction à la médecine expérimentale de Claude Bernard	Actif	Informations sur le programme Colisciences	[ Supprimer ] [ Editer   Désactiver   Supprimer ]
Pourquoi Claude Bernard	Actif	Informations sur le programme Colisciences	[ Editer   Désactiver   Supprimer ] [ Editer   Désactiver   Supprimer ]
Hypertexte et cognition	Actif	Comment naviguer dans le contenu du site	[ Editer   Désactiver   Supprimer ]

**Ajouter une nouvelle catégorie**

**Titre:**

**Description:**

**Editer la catégorie**

**Catégorie:**

FIG. G.8: L'ajout d'une catégorie

Par exemple, pour ajouter « Acteurs et Partenaires du programme » il suffit d'entrer cette phrase dans la boîte « titre » de la zone « Ajouter une nouvelle catégorie », puis d'ajouter dans la boîte « description » l'endroit où on veut voir apparaître le nouveau bouton sur la page d'accueil : « gauche », « milieu » ou « droite ». Sur cet exemple, nous avons choisi à gauche.

Après avoir cliqué sur « Ajouter » la nouvelle catégorie est enregistrée et il ne reste plus qu'à lui ajouter un contenu pour la voir affichée à l'endroit adéquat sur la page d'accueil.

Il est aussi possible de modifier une catégorie en utilisant le menu déroulant de la zone « Editer la catégorie » pour voir notre nouvelle catégorie dans la liste comme représenté sur la figure G.10. Si nous cliquons sur « Editer » nous pourrions modifier l'intitulé de la catégorie et décider de la placer au milieu ou à droite.

Plus bas apparaît la zone de création de contenu « Ajouter une nouvelle page » que nous allons utiliser pour créer un contenu attaché à cette nouvelle catégorie.

Ajouter

Editer la catégorie

Catégorie: Acteurs et Partenaires du programme [Editer]

Acteurs et Partenaires du programme  
Autour de CoLiSciences: études et travaux  
Comment naviguer dans le contenu du site  
Informations sur le programme Colisciences

nouvelle page

Titre:

Catégorie: Aucun

Sous-Titre:

Haut du Texte:

FIG. G.9: La liste des catégories

Une fois entré le titre et sélectionné la catégorie, l'administrateur est libre de rentrer le texte comme il l'entend, les options de présentations sont assez sommaires, elles reprennent les fonctions les plus courantes d'un éditeur de texte.

Il faut descendre jusqu'au bout de la zone d'édition pour voir apparaître le bouton « Envoyer » qui permet de sauvegarder la page qui sera visible directement en ligne.

**Ajouter une nouvelle page**

**Titre:**

**Catégorie:**

**Sous-Titre:**

**Haut du Texte:**

**Texte de la page:**  
Si vous voulez plusieurs pages, vous pouvez écrire <!--pagebreak--> là où vous voulez un saut de page.

FIG. G.10: L'ajout d'une page

ANNEXE H

**RÉFÉRENCES SUR LES FONDATEURS DE  
L'HYPERTEXTE**

---

## Références sur les fondateurs<sup>192</sup>

### Bush et le Memex

- Une brève présentation de Vannevar Bush se trouve à l'adresse : <http://www.iath.virginia.edu/elab/hf10034.html>
- Son article fondateur *As we may think* est lui aussi en ligne : <http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm> ou <http://www.isg.sfu.ca/~duchier/misc/vbush/>
- Une image du Memex : [http://www.kerryr.net/pioneers/memex\\\_pic.htm](http://www.kerryr.net/pioneers/memex\_pic.htm)
- Le Memex et au-delà : <http://www.cs.brown.edu/memex/>

### Douglas Engelbart

- Une brève présentation : <http://www.iath.virginia.edu/elab/hf10035.html>
- La démonstration du prototype de 1968 : <http://sloan.stanford.edu/MouseSite/1968Demo.html>
- Une interview pour les enfants : <http://www.superkids.com/aweb/pages/features/mouse/mouse.html>
- Le travail coopératif : <http://www.bootstrap.org/>
- Augmenting Human Intellect : A Conceptual Framework (<http://www.histech.rwth-aachen.de/www/quellen/engelbart/ahi62index.html>)

### Ted Nelson et l'hypertexte

Brève présentation : <http://www.iath.virginia.edu/elab/hf10155.html>

---

192. Ces références proviennent de (Bruillard, 2002).

---

Projet Xanadu : <http://xanadu.com/> Sa home page : <http://www.sfc.keio.ac.jp/~ted/> Ted Nelson et le mot hypertexte en 1965 : [http://iberia.vassar.edu/~mijoyce/Ted\\\_sed.html](http://iberia.vassar.edu/~mijoyce/Ted\_sed.html)

## **Histoire des interfaces graphiques**

- GUI : *Graphical USer Interface*
- Bush et Engelbart. [http://www2.kenyon.edu/people/adamsal/gui/bush\\\_engelbart.htm](http://www2.kenyon.edu/people/adamsal/gui/bush\_engelbart.htm)



ANNEXE I

## LICENCES LIBRES

---

### I.1 La licence GPL

Version 2, June 1991

Copyright © 1989, 1991 Free Software Foundation, Inc.

59 Temple Place - Suite 330, Boston, MA 02111-1307, USA

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

## **1. Preamble**

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change free software—to make sure the software is free for all its users. This General Public License applies to most of the Free Software Foundation’s software and to any other program whose authors commit to using it. (Some other Free Software Foundation software is covered by the GNU Library General Public License instead.) You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs ; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid anyone to deny you these rights or to ask you to surrender the rights. These restrictions translate to certain responsibilities for you if you distribute copies of the software, or if you modify it.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must give the recipients all the rights that you have. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

We protect your rights with two steps : (1) copyright the software, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the software.

Also, for each author's protection and ours, we want to make certain that everyone understands that there is no warranty for this free software. If the software is modified by someone else and passed on, we want its recipients to know that what they have is not the original, so that any problems introduced by others will not reflect on the original authors' reputations.

Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that redistributors of a free program will individually obtain patent licenses, in effect making the program proprietary. To prevent this, we have made it clear that any patent must be licensed for everyone's free use or not licensed at all.

The precise terms and conditions for copying, distribution and modification follow.

## GNU GENERAL PUBLIC LICENSE

### TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License applies to any program or other work which contains a notice placed by the copyright holder saying it may be distributed under the terms of this General Public License. The "Program", below, refers to any such program or work, and a "work based on the Program" means either the Program or any derivative work under copyright law : that is to say, a work containing the Program or a portion of it, either verbatim or with modifications and/or translated into another language. (Hereinafter, translation is included without limitation in the term "modification".) Each licensee is addressed as "you".  
  
Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running the Program is

not restricted, and the output from the Program is covered only if its contents constitute a work based on the Program (independent of having been made by running the Program). Whether that is true depends on what the Program does.

1. You may copy and distribute verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty ; keep intact all the notices that refer to this License and to the absence of any warranty ; and give any other recipients of the Program a copy of this License along with the Program.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Program or any portion of it, thus forming a work based on the Program, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions :

- (a) You must cause the modified files to carry prominent notices stating that you changed the files and the date of any change.
- (b) You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License.
- (c) If the modified program normally reads commands interactively when run, you must cause it, when started running for such interactive use in the most ordinary way, to print or display an announcement including an appropriate copyright notice and a notice that there is no warranty (or else, saying that you provide a warranty) and that users may redistribute the program under these conditions, and telling the user how to view a copy of this License. (Exception : if the Program itself is interactive but

does not normally print such an announcement, your work based on the Program is not required to print an announcement.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Program, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Program, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you ; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Program.

In addition, mere aggregation of another work not based on the Program with the Program (or with a work based on the Program) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may copy and distribute the Program (or a work based on it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you also do one of the following :
  - (a) Accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange ; or,
  - (b) Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections

1 and 2 above on a medium customarily used for software interchange ;

or,

- (c) Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source code, even though third parties are not compelled to copy the source along with the object code.

- 4. You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.
- 5. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the

Program or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it.

6. Each time you redistribute the Program (or any work based on the Program), the recipient automatically receives a license from the original licensor to copy, distribute or modify the Program subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.
7. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Program at all. For example, if a patent license would not permit royalty-free redistribution of the Program by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Program.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims ; this section has the sole purpose of protecting the integrity of the free software distribution system, which is implemented by public license practices. Many people have

made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system ; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

8. If the distribution and/or use of the Program is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Program under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.
9. The Free Software Foundation may publish revised and/or new versions of the General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies a version number of this License which applies to it and “any later version”, you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of this License, you may choose any version ever published by the Free Software Foundation.

10. If you wish to incorporate parts of the Program into other free programs whose distribution conditions are different, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation ; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all

derivatives of our free software and of promoting the sharing and reuse of software generally.

## NO WARRANTY

11. BECAUSE THE PROGRAM IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.
12. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

## END OF TERMS AND CONDITIONS

## **Appendix : How to Apply These Terms to Your New Programs**

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty ; and each file should have at least the “copyright” line and a pointer to where the full notice is found.

<one line to give the program’s name and a brief idea of what it does.>

Copyright (C) <year> <name of author>

This program is free software ; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation ; either version 2 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY ; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program ; if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA.

Also add information on how to contact you by electronic and paper mail.

If the program is interactive, make it output a short notice like this when it starts in an interactive mode :

Gnomovision version 69, Copyright (C) <year> <name of author>

Gnomovision comes with ABSOLUTELY NO WARRANTY; for details  
type 'show w'.

This is free software, and you are welcome to redistribute it under certain  
conditions; type 'show c' for details.

The hypothetical commands `show w` and `show c` should show the appropriate parts of the General Public License. Of course, the commands you use may be called something other than `show w` and `show c`; they could even be mouse-clicks or menu items—whatever suits your program.

You should also get your employer (if you work as a programmer) or your school, if any, to sign a “copyright disclaimer” for the program, if necessary. Here is a sample; alter the names :

Yoyodyne, Inc., hereby disclaims all copyright interest in the program  
'Gnomovision' (which makes passes at compilers) written by James  
Hacker.

<signature of Ty Coon>, 1 April 1989

Ty Coon, President of Vice

This General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Library General Public License instead of this License.



ANNEXE J

## **BREVETABILITÉ DES LOGICIELS**

---

## **J.1 Les brevets logiciels**

### **A. Pourquoi faut-il être contre les brevets logiciels ?**

Un logiciel est un texte, il est donc protégé par les droits d'auteurs au même titre qu'un livre ou un morceau de musique. Le système des brevets protège, lui, des inventions industrielles. Le coût d'un tel brevet est presque négligeable au regard d'investissements industriels lourds mais reste inaccessible pour un simple développeur talentueux. Pourquoi changer un système qui a fait ses preuves ? Et dans l'intérêt de qui ? Certainement pas celui de la France ni de l'Europe. En connaissance de cause, le vote du Parlement Européen avec ses amendements est en contradiction totale avec le texte en faveur des brevets logiciels proposé par la Commission.

Pour mieux comprendre ce qu'ont à gagner les promoteurs des brevets logiciels, il est souhaitable de commencer par chercher du côté de la cible qu'ils cherchent à atteindre : déstabiliser les Logiciels Libres, mais en fait c'est toute l'industrie des logiciels, et en particulier les PME, qui seraient impactées par cette réglementation.

Les Logiciels Libres, plus qu'une technologie, sont une philosophie de développement et de distribution des logiciels. Le concept ne date pas d'hier puisque leur origine remonte à la création de la FSF par Richard Stallman en 1984. En revanche ils ne sont sortis du cercle fermé des « hackers » et la bulle technologique que depuis que GNU/Linux a commencé à faire son entrée dans le monde de l'entreprise. En effet depuis maintenant plus de quatre ans GNU/Linux est de moins en moins perçu comme un sympathique gadget technologique, mais une vraie alternative aux systèmes propriétaires Unix ou Windows pour les entreprises, voire les grandes administrations.

Le contexte Internet et celui des serveurs Web est bien entendu le domaine où GNU/Linux s'est le mieux imposé du fait des liens étroits qui lient Internet à Unix et finalement à GNU/Linux par simple évolution entre architectures équivalentes. Depuis 1995, Netcraft, une société anglaise, tiens à jour des statistiques sur le nombre de sites Web et les logiciels qu'ils utilisent. En juillet 2005, le graphique représenté

sur la figure J.1<sup>193</sup> montre très clairement l'hégémonie du serveur Apache, développé sous une licence de Logiciel Libre et installé principalement sur des architectures de type GNU/Linux par rapport aux serveurs PWS, IIS, ... de Microsoft.

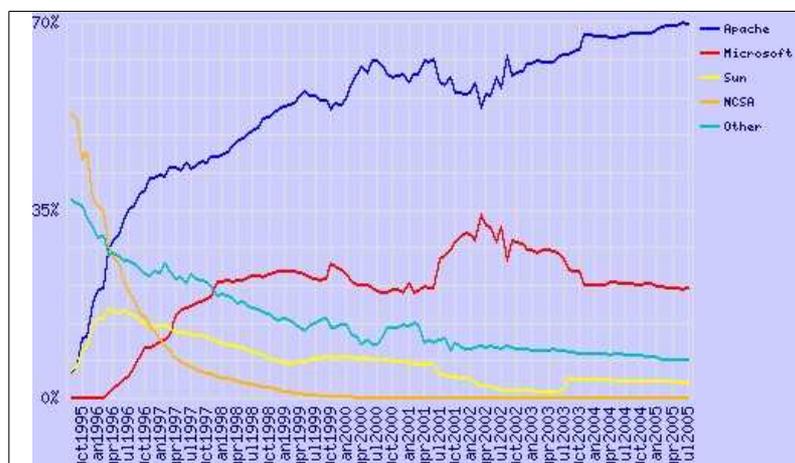


FIG. J.1: Les parts de marché Apache *vs* Microsoft

D'autre part, même sur son cœur de métier, Microsoft doit subir les assauts de ses challengers Logiciels Libres comme cela a été mis en évidence par plusieurs déclarations de gouvernements, d'administration ou de grand groupes sur leur volonté de migrer vers une plateforme GNU/Linux, ou plus modestement d'utiliser plus largement des Logiciels Libres. Plusieurs projets de lois sont plus ou moins avancés dans de nombreux pays : Argentine, Australie, Belgique, Brésil, Colombie, Costa-Rica, Danemark, France, Allemagne, Italie, Pérou, Portugal, Espagne, Ukraine, Uruguay et même Etats Unis d'Amérique (voir <http://proposicion.org.ar/doc/referencias/index.html> pour plus de détails). Plus précisément, nous pouvons noter dans l'actualité récente :

- Le gouvernement français veut diviser par deux le coût des logiciels informatiques utilisés par l'Etat en s'adressant à des éditeurs de « logiciels libres », qui seront ainsi mis en concurrence avec le géant américain Microsoft, a déclaré vendredi à Reuters Renaud Dutreil. (18 juin 2004, source REUTER).

193. Ce graphique est mis à jour tous les mois et disponible en ligne à l'adresse <http://www.netcraft.com>, nous l'avons déjà présenté dans la section 3., page 187 du chapitre 8 traitant des Logiciels Libres.

- Le projet LiMux de migration vers Linux des 14 000 ordinateurs de la ville de Munich (source Suse).

La première entreprise à souffrir de cette nouvelle donne est sans doute Microsoft : colosse aux pieds d'argile, sa suprématie sur le monde des ordinateurs personnels est quasi totale mais repose sur une ligne très étroite de produits : un système d'exploitation Windows et une suite bureautique Office. En comparaison, les autres produits de la marque sans être marginaux ne sont pas en mesure d'assurer la pérennité de l'entreprise. C'est sans doute la raison pour laquelle Microsoft doit se battre sur tous les fronts, commerciaux tout d'abord mais maintenant aussi légaux pour se maintenir en position de leader et contenir la progression de GNU/Linux.

Dans ce contexte chaque attaque contre GNU/Linux est interprétée comme les conséquences d'un complot derrière lequel l'ombre de Microsoft n'est jamais très loin : le procès intenté par SCO contre IBM, pour atteinte à leur propriété intellectuelle, puis contre les utilisateurs de Linux. L'ouvrage très controversé de la fondation Alexis de Tocqueville sur Linus Torvald et sa capacité à développer Linux, etc... Les brevets logiciels sont le dernier avatar que les sociétés commercialisant des logiciels propriétaires ont inventé pour couper l'herbe sous les pieds des sociétés et organisations développant des Logiciels Libres. L'impact des brevets logiciels est loin d'être neutre sur la société humaine, sur la notion de pérennisation du patrimoine de l'humanité par rapport à la monopolisation de ce savoir par quelques groupes industriels et finalement un impact sans doute très grave sur l'innovation.

La formulation de cette proposition va engendrer des dérives dans tous les domaines touchant à l'informatique. Les brevets logiciels constituent un danger économique, politique, juridique et technique pour des raisons très simples :

**Brevétisation de tout :** Le texte est trop permissif et tout est brevetable (logiciels, algorithmes, formules, idées... jusqu'au plus basique réflexe cérébral). La confiscation du savoir humain et de l'information par quelques-uns deviendra possible.

**Monopole absolu grâce à des pratiques tarifaires abusives :** Seules les plus grosses sociétés auront les moyens de déposer des brevets. Les SME déposeront elles le bilan, comme les développeurs indépendants. Le dépôt à lui seul étant excessivement cher (49000 euros en France, on peut comprendre la motivation des organismes de propriété industrielle!), sans concurrence, les multinationales pourront pratiquer les tarifs de leur choix.

**Frein à l'innovation :** Les promotions internes et les stratégies commerciales des multinationales seront évidemment fondées sur l'économie de brevet. Les produits grand-public que l'on connaît aujourd'hui en sont une preuve navrante. Or ce que nous sommes capable de faire aujourd'hui en Europe grâce à l'interopérabilité, les dépasse et ils le savent. Nous perdrons notre avance, et nous devrons regresser sous leur domination. A l'image de ce qu'il se passe aux USA, les grosses compagnies se contenteront d'acquérir des portefeuilles et d'attaquer en justice les contrevenants pour gagner de l'argent facilement au détriment de l'innovation (Les brevets triviaux rapporteront le plus).

**Fin des Logiciels Libres et du monde Open-Source :** Ils représentent pourtant un intérêt majeur pour l'Europe, avec l'adaptabilité que l'on connaît aux collectivités. La vie de la communauté Linux, très nombreuse et internationale, est suspendue à cette décision.

**Fin de l'indépendance de l'Europe face aux USA :** Les multinationales présentes sur le marché des brevets sont en majorité américaines et japonaises. Elles ont déjà commencé : plus de 30 000 brevets ont été déposés de façon illégale à l'Office Européen des Brevets en attendant d'être validés par la loi. C'est brevets Européens sont des monopoles accordés sur le marché Européen, mais sont détenus dans leur écrasante majorité par des sociétés Américaines (22778 brevets logiciels) et Japonaises (10580 brevets logiciels) plutôt que françaises (1634 brevets logiciels).

Cette liste n'est malheureusement pas exhaustive.

Vous ne pouvez pas ignorer la spécificité des logiciels qui, de par l'extraordinaire facilité de combinaisons de fonctionnalités qu'il permettent, fait que les brevets ne couvrent pas des produits mais des fragments de fonctionnalités (comme la barre de progression!). Développer un logiciel sans enfreindre aucun brevets relève alors de la mission impossible et l'incertitude juridique est alors inévitable car les petites entreprises innovantes n'ont pas de ressources à disperser en frais d'expertise juridique.

Il n'est pas acceptable de voir les multinationales développer ce marché sans l'ombre d'un débat au sein des gouvernements des pays de l'Union. Ce sujet étant difficile d'accès, il nécessite une réelle information. Etant donné les propos tenus par les défenseurs du projets qui ont manifesté une incompétence flagrante, on peut parler de manipulation. Le même problème se retrouve dans les grenades entreprises qui font l'erreur de confier leur communication sur le sujet à leur département de P.I. qui vit par et pour les brevets et en surestime l'intérêt global (gains moins coûts directs de licences mais aussi coûts indirects induits par l'entrave au jeu de la concurrence).

Il n'est pas normal que le processus co-décisionnaire ait été aussi rapide depuis l'arrivée du Commissaire Bolkenstein et de Madame Mac Carthy. Après le vote du 18 mai qui n'a remporté qu'une timide majorité dans une ambiance douloureuse, il demande à ce que la décision finale soit prise à la rentrée des parlementaires. Des parlementaires seuls devant une décision dont ils ignorent la gravité, sans information véritable ni échange avec leurs gouvernements respectifs.

Les brevets ont permis l'innovation quand les investissements matériels pour faire de la recherche et du développement étaient très lourds. Quand la recherche passe du matériel au logiciel, de la construction de machines à celle d'idées, les investissements ne sont plus les mêmes. Il devient très difficile d'identifier l'antériorité d'une idée, surtout quand il est fait référence aux outils de base d'une interface graphique comme l'a fait Microsoft en déposant un brevet sur l'action de "double-clic" qui permet de sélectionner et lancer une application informatique avec une telle interface. C'est donc le pari qu'a tenté de prendre Microsoft. La conséquence immédiate est que

toute société souhaitant développer et distribuer aux USA un logiciel qui utilise le “double-clic” doit acheter une licence d’utilisation de cette “technologie” à Microsoft, même les logiciels qui ne sont pas basés sur la plate-forme Windows. Par voie de conséquence, compte tenu de l’universalité du “double-clic”, ce brevet consiste à la création d’une sorte de “taxe” Microsoft qui s’applique sur tous les logiciels, GNU/Linux compris.

Les conséquences économiques, en particulier pour un jeune créateur d’entreprise qui souhaiterait se lancer dans une activité liée au logiciel sont désastreuses. Il serait intéressant d’analyser quel impact ce type de brevet va avoir sur le dynamisme et la réactivité de l’ensemble des acteurs de la “nouvelle économie”, mais ce n’est pas le premier objet de cette étude. D’ailleurs, pour conclure et bien mettre en évidence le problème suscité par les brevets logiciels, il faut ajouter que ce brevet est loin d’être unique, c’est sans doute le meilleur exemple des abus qu’il est possible de tirer de l’utilisation de brevets logiciels.

Mais là où l’impact de l’utilisation du « verrou » brevet logiciel pour assurer la dominance d’une seule société commerciale au détriment du reste de la Société est moins flagrant mais peut avoir des conséquences bien plus graves, c’est au niveau de l’équilibre mondial ; nous avons déjà vu que c’était un domaine dans lequel l’Europe serait très désavantagée, parce qu’elle « jouerait le jeu ». En effet, les brevets ne sont pour l’instant en vigueur qu’aux USA, l’Europe est en train de statuer et il n’est pas tout à fait clair comment la balance va pencher. Dans ces conditions, si nous imaginons les brevets logiciels mettant à mal l’économie des logiciels libres, voire leur disparition, cette disparition n’aura lieu que dans les pays sous le contrôle des brevets logiciels.

Nous savons déjà que la Chine a beaucoup d’ambitions dans le domaine des Logiciels Libres, parce que cela paraît être un tremplin pour brûler les étapes, gagner et développer leur indépendance dans le domaine des TIC qui reste pour l’instant complètement lié aux Etats-Unis. Si GNU/Linux disparaissait, combien de temps faudrait-il à la Chine pour développer un nouveau système d’exploitation ? Nous

pouvons aussi nous poser la question de ce que serait ce nouveau système : une continuation de l'architecture GNU/Linux ou pourquoi pas un système compatible cette fois avec l'environnement Windows ? Si la Chine développait un tel système, quel impact aurait-il sur notre propre économie ? Sans aucun doute commenté et documenté en Chinois, ce système nous serait en pratique complètement hermétique et de toute manière légalement hors de notre portée puisque sans doute sous le coup de nombreux brevets. Comment réagirait notre économie, comment nos outils pourront-ils continuer à interagir avec ce système ?

Nous ne voulons pas avoir à nous poser ces questions. En tant que professionnel de l'informatique et utilisateur de logiciels libres, je me devais de vous alerter sur un problème qui touche l'innovation, l'entrepreneuriat, les nouvelles technologies et la globalisation, mais qui malheureusement n'a pas la médiatisation qu'il mérite.

Nous sommes tous concernés de très près mais pourtant l'inquiétude qui grandissait au sujet des brevets logiciels Européens n'a que peu dépassé le cercle des défenseurs des Logiciels Libres, alors qu'il aurait du soulever toutes les entreprises Européennes du monde logiciel, en particulier les PME et les sociétés de services (SSII). Malgré la campagne pour les Européennes, le débat n'est jamais réellement sorti de la sphère des lobbyistes et n'a pas intéressé ni le public, ni les journalistes. Malgré tout, la directive a été complètement abandonnée le 6 juillet 2005 par le Parlement Européen, après que les amendements qu'il proposait d'apporter au projet initial ne satisfassent plus ni les pour, ni les contre.

## B. Conséquences pour *C<sub>o</sub>L<sub>i</sub>S<sub>ciences</sub>*

Un examen rapide d'une page « d'exemples assortis de brevets logiciels en Europe »<sup>194</sup> nous a permis d'identifier rapidement une courte liste de brevets que *C<sub>o</sub>L<sub>i</sub>S<sub>ciences</sub>* utilise sans pour autant en avoir acquis les droits au préalable. Il faut tout de même noter en guise de mise en garde préalable que lorsque le développement de *C<sub>o</sub>L<sub>i</sub>S<sub>ciences</sub>* a débuté ces brevets étaient illégaux en Europe, il faudrait alors

---

194. <http://swpat.ffii.org/patents/samples/index.en.html>

s'interroger sur la pertinence de l'enregistrement de tel brevets par l'Office des Brevets Européens (OBE), hors de tout contexte légal sur ce territoire. De plus, la trivialité des « technologies logicielles » protégées par ces brevets devrait suffire à prouver notre bonne foi et à justifier le fait qu'il n'a s'agit d'aucune façon d'une utilisation illégale de technologies propriétaires, mais plutôt de mise en œuvre de choses déjà vues des centaines de fois sur d'autres sites, programmes ou objets numériques.

Ces exemples devraient suffire à montrer la dangerosité de l'extension du système des brevets au monde logiciel. Nous avons cherché à mettre notre projet à l'abri de l'obsolescence des solutions informatiques propriétaires en utilisant des logiciels libres et voici que l'obsolescence et les logiciels propriétaires reviennent à la charge pour mettre en péril notre travail dans un simple but de gains commerciaux à très court terme.

## 1. Pages dynamiques

patent number :

EP0747840

Name of the Invention :

A method for fulfilling requests of a web browser

Ce brevet, déposé par la compagnie IBM en 1996 semble porter sur l'utilisation d'un serveur Web pour dépouiller des formulaires et renvoyer une page html construite à partir de plusieurs sources et en fonction des données envoyées par l'internaute. C'est plus ou moins ce qui est mis en œuvre dans *CoLiSciences* pour construire une page Web à partir d'informations provenant de bases de données, de fichiers XML et en fonction des requêtes des lecteurs.

## 2. Aide contextuelle

patent number :

EP517486

Ce brevet couvre l'utilisation d'un système d'aide spécifique au contexte dans lequel se trouve l'utilisateur, comme par exemple les informations qui s'affichent en titre de page pendant la lecture d'un ouvrage *C<sup>o</sup>L<sup>i</sup>Sciences*.

### 3. Systèmes d'onglets

Sur plusieurs écrans nous utilisons un système d'onglets tels que ceux d'un répertoire alphabétique papier. Cette « technologie » est brevetée par la société Adobe Systems Inc (US).

publication date :

1995-12-27

patent number :

EP0689133

Name of the Invention :

Method of displaying multiple sets of information in the same area of a computer screen

Claim 1 :

ML ep689133c

Procédé de combinaison sur un affichage d'ordinateur d'un ensemble additionnel d'information affiché dans une première zone de l'affichage et possédant en association un indicateur de sélection dans un groupe de plusieurs ensembles d'information requis sur une base récurrente affiché dans une seconde zone de l'écran, comprenant les étapes suivantes :

1. l'établissement de la seconde zone sur l'affichage d'ordinateur dans laquelle le groupe de plusieurs ensembles d'information est affiché, la seconde zone ayant une dimension qui est inférieure à la zone total de l'affichage d'ordinateur, la seconde zone affichant un premier parmi plusieurs ensembles d'information ;
2. la prévision dans la seconde zone d'une pluralité d'indicateurs de sélection, chacun d'eux étant associé à un ensemble correspondant de la pluralité d'ensembles d'information,

3. la sélection d'un second ensemble de la pluralité d'ensembles d'information pour un affichage dans la seconde zone par activation d'un indicateur de sélection associé à un second ensemble de la pluralité d'ensembles d'information, le second ensemble de la pluralité d'ensembles d'information étant ainsi substitué au premier de la pluralité d'ensembles d'information dans la zone de l'affichage ; et
4. la combinaison d'un ensemble additionnel d'information, affiché dans la première zone de l'affichage dans le groupe des multiples ensembles d'information de telle façon que l'ensemble additionnel d'information puisse être sélectionné à l'aide de son indicateur de sélection de la même façon que les autres ensembles du groupe.



ANNEXE K

## ACHAT D'OUVRAGES NUMÉRIQUES EN LIGNE

---

## K.1 Petit voyage à travers DRM, cas pratique

Nous avons abordé d'un plan tout à fait théorique les raisons du manque d'enthousiasme de la part du public en général pour les ouvrages numérisés. Nous n'allons pas revenir sur le prix d'achat ni la complexité technique des dispositifs permettant la lecture des ebooks. Nous ne reviendrons pas plus sur les problèmes de portabilité, ni d'autonomie, ni sur le plaisir de feuilleter un beau velin qu'il faudrait comparer au plaisir de manipuler un gadget coûteux. Nous allons précisément prendre comme postulat de départ que nous sommes un lecteur convaincu de livres électroniques, et que nous sommes muni du dispositif technique adéquat pour lire des ebooks, en l'occurrence un ordinateur personnel. Il nous faut maintenant nourrir notre bibliothèque numérique, pour cela quoi de plus naturel que de se connecter par Internet et le Web à une librairie virtuelle pour procéder à un achat en ligne. Nous voici donc sur le site de NumiLog<sup>195</sup>, mais nous pourrions faire la même chose sur Amazon, Fnac et d'autres. Choisissons quelques ouvrages en les ajoutant à notre *caddie*, les prix sont assez attractifs du fait de leur statut d'ouvrage numérique et certains sont affichés à 50% du prix éditeur. Ceci fait, nous procédons à la validation de la commande : jusqu'à présent tout allait bien, mais à présent il s'agit de valider nos droits à lire les ouvrages que nous allons acheter. Cette vérification est prise en charge par diverses technologies, rassemblées sous le vocable *DRM (Digital Rights Management)*, pour ce faire les messages de précautions et le voyage commencent.

Nous sommes tout d'abord alertés, comme il est possible de le vérifier sur la copie d'écran de la figure K.1, sur le fait que notre commande comprend des fichiers au format PDF protégé et d'autres au format LIT, bien entendu protégé lui aussi. Ces fichiers peuvent être lus uniquement avec certaines versions spécifiques des logiciels Adobe Reader ou Microsoft Reader. Bien entendu ces mêmes logiciels réclament en cascade des versions spécifiques du système d'exploitation de l'ordinateur qui les héberge, Windows, Pocket PC, Mac OS et Palm sont représentés, en revanche pour

---

195. <http://www.numilog.com/>

les utilisateurs de GNU/Linux, il faudra attendre, continuer à lire des ouvrages du domaine public ou papier, aucune information n'est disponible à ce sujet.

Au niveau des restrictions d'utilisation, il est de plus précisé que, contrairement à ce qu'il est possible de faire avec un livre papier, ces livres ne peuvent pas être prêtés à un ami, ni même à son conjoint, à un ascendant ou un descendant. Ce n'est même pas très sûr qu'il soit possible de lire sur son deuxième ordinateur...

Dernière contrainte, ces fichiers ne peuvent être ni imprimés ni lus à haute voix, l'ordinateur a permis pendant longtemps aux non-voyants d'accéder enfin à la lecture de façon autonome, les progrès des livres numériques vont-ils supprimer cette possibilité ? Heureusement, parallèlement les audio-books se développent aux aussi.

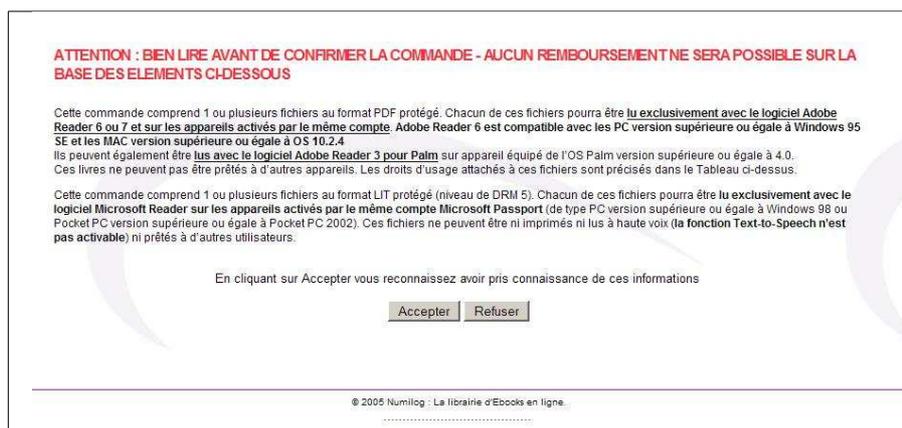


FIG. K.1: Confirmation commande NumiLog

Si ces mises en gardes n'ont pas découragé le futur lecteur, il lui suffit de cliquer accepter pour pouvoir continuer. En fait, le téléchargement ne pourra s'effectuer qu'après la réception d'un mail qui contient les informations et les liens sur lesquels les ebooks sont en attente pendant un temps limité. Un extrait de ce mail est visible ci-dessous.

```
Bonjour et merci pour votre commande chez NUMILOG (numéro de facture
: 0542484171748)
*****
C'est votre 1er achat ? Nous vous recommandons de vérifier que vous
```

*Annexe K. Achat d'ouvrages numériques en ligne*

---

avez installé le(s) logiciel(s) requis pour lire vos ebooks.

#### Vous avez acheté un livre au format PDF : vérifiez que vous avez bien installé le nouveau logiciel Adobe Reader 7.0 (disponible à l'adresse suivante :

<http://www.adobe.fr/products/acrobat/readstep2.html>)

ATTENTION : les utilisateurs de Mac OS9 doivent utiliser le logiciel Acrobat eBook Reader pour lire certains ebooks au format PDF. Pour télécharger le fichier d'installation :

<http://www.numilog.com/download/eBookReaderInstaller.bin>

#### Vous avez acheté un livre au format LIT : vérifiez que vous avez bien installé le logiciel Microsoft Reader (disponible à l'adresse suivante :<http://www.microsoft.com/reader/fr/>)

ATTENTION : Ces liens ne sont utilisables que pendant 2 jours calendaires.

RAPPEL : de nombreux livres au format LIT signalés en ligne par un pictogramme << Passeport MSR >> requièrent pour être lus sur un PDA de posséder le système d'exploitation Pocket PC 2002.

VOICI LE(S) LIENS SUR LESQUELS IL FAUT CLIQUER POUR ACCEDER A VOS LIVRES :

<http://www.numilog.com/download.asp?id=izydqj2prjbbibwhccopqzp2ze2sj>

...

\*\*\*\*\*

Pour d'autres précisions sur le téléchargement et sur le mode

d'emploi des logiciels de lecture, vous pouvez consulter nos rubriques FAQ et READERS sur <http://www.numilog.com/readers.asp>

Visualisez votre facture en cliquant sur le lien suivant :  
<http://www.numilog.com/facturefinale.asp?id=0542484171748>

En cas de problème, contactez-nous à l'adresse e-mail suivante :  
[commande@numilog.com](mailto:commande@numilog.com)

Nous vous rappelons que ces livres sont protégés par le code de la propriété intellectuelle et que leur copie est strictement réservée à l'usage privé.

Bonne Lecture,  
L'équipe de Numilog.

FIG. K.2: Le mail de confirmation

Un clic sur le lien cité au milieu des messages d'avertissement permet de voir apparaître l'écran de téléchargement, tel qu'il est reproduit sur la figure K.3. Bien entendu, tout cela reste très technique et d'une logique dont seule la science informatique a le secret. D'une part, il nous est dit qu'il suffit d'avoir la version 6.0 du fameux logiciel Adobe Reader, mais il est seulement proposé de télécharger les livres dans la version 7.0. D'autre part, la version 6.0 étant installée sur la machine qui a servi à faire cet essai, voir apparaître ce message paraît incohérent ou un problème de détection de la version de notre *Reader*<sup>196</sup>. La technologie étant sans doute encore en train d'être mise au point et considérant que la version que nous utilisons peut lire ce livre, nous cliquons le lien qui propose de télécharger nos livres numériques « *directement dans Adobe Reader 7.0* »(sic).

196. D'ailleurs, nous pouvons aussi nous inquiéter sur le fait qu'un site Web soit capable ou pas de détecter les versions des logiciels installés sur notre machine, ce qui est un autre propos.



FIG. K.3: Version d'Adobe Reader non reconnue

En fait de livre, c'est uniquement un tout petit fichier qui est transféré sur notre machine. Il va donc falloir continuer notre voyage en ouvrant ce fichier par un double-clic qui lance Adobe Reader 6.0. Un message nous avertit au démarrage de ce dernier que notre logiciel n'est pas « activé ». Ce qui est peut-être l'explication du fait que notre Adobe Reader n'était pas reconnu à l'étape précédente. Pour « activer » Adobe Reader, il faut se connecter au site Adobe, où l'information telle qu'elle est présentée figure K.4 est affichée : il faut donc nous identifier, pour attacher notre identité numérique à ce livre. C'est ce qui est appelé DRM (*Digital Rights Management*) et visiblement cela passe en priorité par l'ouverture d'un compte *Passport*, géré par la société Microsoft.

Pour les lecteurs qui n'ont pas de compte *Passport* et qui ne souhaitent pas en utiliser un, il reste la possibilité de suivre un autre lien qui mène à la page visible sur la figure K.5. Dans ce cas, c'est auprès de Adobe qu'il faudra nous identifier, ce qui va donc de toute façon passer par la création d'un compte pour s'identifier.

C'est donc un nouveau lien à suivre, qui nous amène à l'écran présenté sur la figure K.5, pour nous demande de préciser le pays où nous habitons. Cet écran est

## K.1. Petit voyage à travers DRM, cas pratique

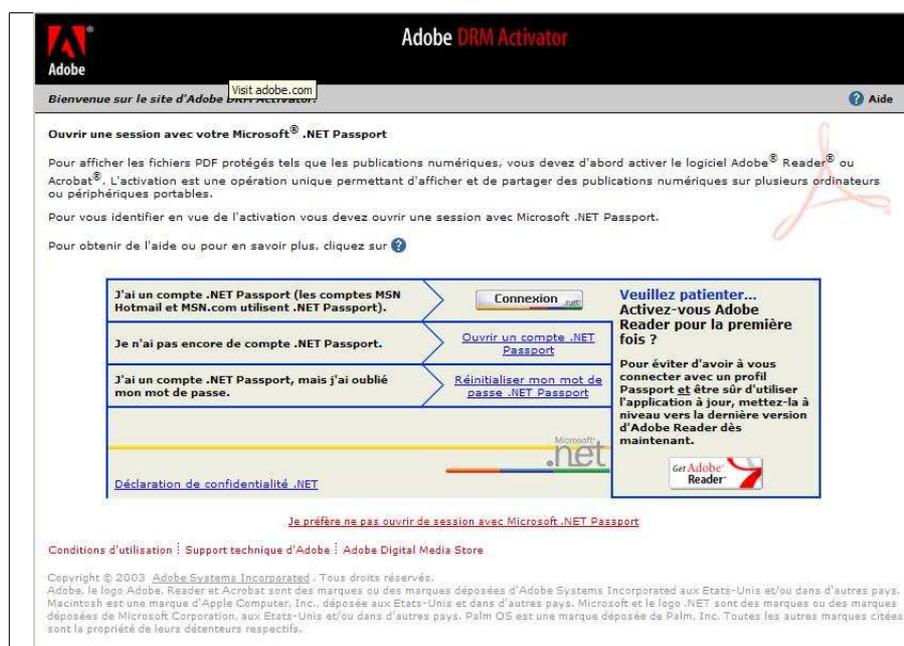


FIG. K.4: DRM et Adobe



FIG. K.5: DRM avec Adobe Reader sans ouvrir de compte Passport

uniquement en anglais, il semble qu'il y ait des restrictions pour certains pays, mais nous voulons simplement lire le ebook dont nous avons fait l'acquisition auparavant.

Pour cela, il nous faut cliquer sur le bouton *continuer* pour aboutir à une autre page, divisée en deux sur les figures K.7 et K.2 pour des raisons de taille.



FIG. K.6: id Adobe

La création d'un « id »<sup>197</sup> Adobe demande la collecte d'un assez grand nombre de données personnelles, avec un risque non-négligeable d'intrusion dans notre vie privée si elles venaient à tomber dans de mauvaises mains. Est-il concevable que toutes nos lectures, du moment qu'elles sont électroniques, puissent être enregistrées dans une base de données et qu'elles puissent servir à créer un « profil » de nos goûts, de nos choix politiques et philosophiques. *Dis moi ce que tu lis et je te dirais qui tu es.* Le risque éthique paraît immense et pourtant nous n'avons pas entendu de débat, de rapport de force entre des « pour » et des « contre ». Du moment que la technologie vient faire écran, nous sommes sensés accepter ce genre de dérive ? La question ne sera pas résolue ici, mais il y a lieu à un vrai débat.

Pour l'instant, nous ne pouvons que constater que même si nous avons dit habiter en France, les pages et les explications sont écrits uniquement en anglais. Effectivement une mention spéciale s'adresse aux Européens dont la vie privée est un peu mieux protégée et réglementée, mais le commentaire précise en fait que dans ce cas, ils abandonnent leurs droits et leur protection légale.

197. pour les Nord-Américains, c'est le nom consacré de l'identifiant personnel. C'est le même terme qu'ils utilisent pour ce que nous appelons le numéro de sécurité sociale, c'est-à-dire le seul numéro qui nous suit toute notre vie, en revanche pour eux il est lié à la carte d'identité.

Adobe Sign in | Search | Contact Us | Adobe Stores Worldwide

Products Solutions Support Purchase Company

## enter your account information

Please provide information for all required fields (marked with an asterisk).

**Create your Adobe ID and password**  
You will use your Adobe ID and password to access and modify your registration and account information.

\* E-mail address (this will be your Adobe ID)

\* Password (6-12 characters)

Password hint

\* Confirm e-mail address

\* Confirm password

Remember my login on this computer

The password hint is optional. If you forget your password we will display it to remind you. Your hint should be obvious to you, but not to others who might see it. It should not contain your password.

---

\* First name

\* Last name

Country or region: FRANCE

FIG. K.7: Création d'un compte Adobe (1<sup>ère</sup> partie)

**Uses of my information**

I would like information on Adobe software products and related services, including announcements of special promotions.

E-mail  Mail  Phone

I would like for parties other than Adobe to send me information on their products and services, including announcements of special promotions.

Mail

If you are a resident of a member state of the European Union, we cannot complete your purchase or registration transaction unless you check the boxes below.

\*  I agree that Adobe may process my information to complete my purchase or registration transaction.

\*  I agree that Adobe may transfer my personal information to Adobe subsidiaries and contractors outside of the European Union, including the United States, for the purpose of processing my registration or order and otherwise for the purposes set forth in the Adobe Online Privacy Policy. (Note: You will NOT be contacted about the products and services of Adobe or third parties unless you provided your consent in the first two boxes.)

See our [Online Privacy Policy](#) for details on how your information may be used by Adobe Systems.

Copyright © 2004 Adobe Systems Incorporated. All rights reserved.  
[Terms of Use](#) | [Online Privacy Policy](#) | [Accessibility](#) | [Preventing software piracy](#)  
[Permissions and trademarks](#) | [Product license agreements](#)

Approved by  
**TRUSTe**  
site privacy statement

FIG. K.8: Création d'un compte Adobe (2<sup>ème</sup> partie)



ANNEXE L

**OUVRAGES NUMÉRIQUES ET DOMAINE PUBLIC**

---

## **L.1 Numérique contre domaine public**

La numérisation d'un ouvrage, même du domaine public, peut justifier d'une nouvelle protection par un copyright. Cela n'est pas aberrant au premier abord : les ouvrages papiers, même d'œuvres provenant du domaine public, doivent être achetés. Mais ce sont aussi des objets, l'acheteur comprend que même s'il n'a plus besoin de payer pour accéder aux idées qui sont écrites sur le papier, il doit au moins payer pour celui-ci. Dans le cas du numérique, les choses sont plus complexes. Les idées sont libres de droit, le support est virtuel, pourquoi devoir payer ? Il reste le travail de mise en forme et les coûts de distribution. Dans ce cas, il paraît raisonnable de devoir payer la personne qui stocke et distribue l'ouvrage. Mais une question reste non résolue : pourquoi ne pas pouvoir redistribuer l'ouvrage ? Certains ouvrages sont même distribués gratuitement, sur Numilog<sup>198</sup> par exemple, pourtant la copie et le prêt sont toujours interdits, la pratique est véritablement paradoxale. En fait, du moment que la numérisation utilise une technologie protégée, ajoute ou mets en scène des contenus additionnels, comme il est très facile d'ajouter des restrictions sur l'utilisation de l'ouvrage : prêt interdit, impression ou copie impossible ou limitée du fait de l'industrialisation de ce type de processus, les restrictions sont ajoutées quelque soit le modèle économique de distribution de l'ouvrage.

---

198. <http://www.numilog.fr>

# **Hypertextualité, construction du sens et représentation des connaissances.**

## **Résumé**

Le but général de notre recherche est de présenter un dispositif de traitement d'un grand corpus de textes sous forme hypertextuelle, mis en ligne comme un site « Web » accessible à tous. L'apparente simplicité de cette description pourrait laisser penser qu'il ne s'agit là que d'« un autre site » offrant des textes électroniques en libre accès sur le Web.

Le premier objet d'étude est lié à l'analyse des nouveaux usages d'un corpus de littérature sous format électronique. En fait, la mise en place de ce site est conduite par une réflexion sur les dispositifs hypertextuels, dans le cadre de l'étude des activités cognitives impliquées dans la lecture et la compréhension, au travers de l'analyse de parcours de lecture et de la construction de sens par ces parcours.

Notre analyse porte également sur les modalités de construction des bibliothèques numériques, combinant les aspects techniques et patrimoniaux.

## **Discipline**

Sciences de l'information et de la communication

## **Mots-clefs**

hypertexte, lectochromie, parcours.

## **Intitulé et adresse du laboratoire :**

I3M - Information, Milieux, Médias, Médiations

Université Sophia Antipolis BP 209 06204 NICE CEDEX

## **Title**

**Hypertext, construction of the meaning and knowledge representation.**

## **Abstract**

The general scope of our research is to present an apparatus for the treatment of a large corpus of texts in hypertext form, implemented online as a Web site accessible to everybody. The apparent simplicity of this description could lead one to think that it is merely « another Web site » presenting freely accessible electronic texts online.

The first object of study is in fact the analysis of the new uses of a corpus of literature in electronic format. In fact, the development of this site results from thinking of hypertext systems, within the framework of the study of the cognitive activities involved in reading and comprehension, through analysing reading paths and the consequent construction of meaning.

Our analysis also involves the construction modes of digital libraries. This combines both technical aspects and those of cultural heritage.