Vers une synthèse d'information orientée tâche - Application à la conception et l'évaluation de Tissue MicroArrays

Julie Bourbeillon

THÉSE de DOCTORAT

de

L'UNIVERSITÉ GRENOBLE I - JOSEPH FOURIER

ÉCOLE DOCTORALE INGÉNIERIE POUR LA SANTÉ LA COGNITION ET L'ENVIRONNEMENT

Spécialité INFORMATIQUE - BIOLOGIE

présentée par

Julie BOURBEILLON

pour obtenir le titre de

DOCTEUR de L'UNIVERSITE JOSEPH FOURIER

Sujet de la thèse :

Vers une synthèse d'information orientée tâche Application à la conception et l'évaluation de Tissue MicroArrays

---

Soutenue le 23 Octobre 2007

devant le jury composé de :

H. MARTIN	Professeur à l'Université J. Fourier de Grenoble	*Président*
M-D. DEVIGNES	Chargée de Recherche CNRS au LORIA à Nancy	*Rapporteur*
F. SEDES	Professeur à l'Université P. Sabatier de Toulouse	*Rapporteur*
C. GARBAY	Directeur de Recherche CNRS	*Directrice de thèse*
F. GIROUD	Maître de Conférence à l'Université J. Fourier de Grenoble	*Co-Directrice de thèse*

Thèse préparée au Laboratoire Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble (TIMC-IMAG) et au Laboratoire d'Informatique de Grenoble (LIG)

Remerciements

Je tiens à exprimer tout d'abord mes remerciements aux membres du jury, qui ont accepté d'évaluer mon travail de thèse. Merci à M. Hervé Martin, de l'Université Grenoble I, d'avoir accepté de présider le jury de cette thèse, et à Mmes Marie-Dominique Devignes, du LORIA à Nancy, et Florence Sèdes, de l'Université Toulouse III, d'avoir accepté d'être les rapporteurs de ce manuscrit. Leurs remarques et suggestions lors de la lecture de mon manuscrit m'ont permis d'apporter des améliorations à la qualité de ce dernier.

Merci à Mme Catherine Garbay et Mme Françoise Giroud, qui ont encadré ce travail de Master 2 puis de thèse, avec beaucoup de compétence, d'enthousiasme et de disponibilité. Merci Françoise et Catherine pour vos conseils, vos encouragements, la confiance que vous m'avez accordée au cours de ces années et surtout votre amitié. J'espère avoir été à la hauteur.

Merci ensuite aux équipes qui m'ont accueillie en leur sein, qu'il s'agisse de l'équipe MRIM du LIG ou de l'équipe RFMQ du laboratoire TIMC-IMAG, ainsi qu'à tous ceux que j'ai côtoyés dans les locaux de l'IN3S.

Merci aussi aux amis, thèsards ou non, et à ma famille, officielle et officieuse, pour leur soutien et leurs encouragements.

Merci enfin à Manu, sans qui ces travaux ne seraient pas ce qu'ils sont.

Table des Matières

Table des Figures

Liste des Tableaux

Table des Listings

Listings

Introduction générale

Depuis une dizaine d'années, l'émergence et l'expansion d'Internet se traduisent par une multiplication des sites Web institutionnels, académiques, commerciaux ou personnels. L'explosion documentaire associée est le contexte général classiquement évoqué dans le cadre de nombreux travaux. En particulier menés en Recherche d'Information, leur but est de faciliter l'accès à l'information pertinente au sein de corpus documentaires de taille de plus en plus gigantesque.

Mais les documents multimédia traditionnels (textes, images et vidéos) ne sont pas les seuls touchés par cette explosion documentaire. En effet, de plus en plus d'institutions, dans des domaines de recherche très variés, de la médecine à la sismologie, de l'astronomie à la météorologie, de la démographie à la sociologie, mettent à disposition en ligne les données acquises lors de leurs expérimentations ou les mesures réalisées par leurs instruments, dans un effort de mutualisation des ressources scientifiques et de collaboration inter-laboratoires.

Conjointement, des techniques et matériels nouveaux, privilégiant entre autres la miniaturisation des spécimens et l'automatisation des processus intensifient le traitement en masse d'échantillons, conduisant à une augmentation phénoménale du volume de données disponibles au sein d'une équipe de recherche.

Un exemple typique pourrait être la technologie des Tissue MicroArrays, de plus en plus utilisée en recherche en oncologie, qui permet le traitement simultané de centaines de micro-échantillons de tissus au sein d'une même lame histologique.

Ce type de technologie pose un double problème :

- ⋆ la conception de l'expérience, et en particulier le choix des échantillons à considérer afin de répondre à des questions biologiques précises,
- ⋆ l'utilisation des données acquises lors d'expériences précédentes, pour analyser de nouveaux problèmes biologiques ou extraire des informations pertinentes.

En particulier, la seconde problématique d'exploitation des données est une question qui devient classique pour les sciences expérimentales, dans un contexte où la conduite d'une expérience voit son coût en temps et en matériel augmenter, et où la réutilisation des données d'autres équipes ou d'expériences précédentes dans un nouveau cadre devient la norme.

Cette pratique de réutilisation pose pourtant aux chercheurs un gros problème d'appréhension de jeux de données qu'ils maîtrisent souvent mal, parce qu'ils sont le fruit de travaux d'autres équipes, ou ont été acquis en masse, hors du contexte de validation d'une hypothèse scientifique précise par une expérience au cadre expérimental et à la couverture bien définis et surtout soigneusement délimités.

Or, cette appréhension du jeu de données considéré est une étape indispensable, préalable à une exploitation plus dirigée des données. En effet, le recours à des outils de fouille de données se doit d'être dirigé, et la définition d'un objectif de fouille nécessite une connaissance préalable minimale de l'espace des données. Dans le même esprit, le jeu de données peut servir de base à la poursuite d'études selon une démarche expérimentale plus classique, par validation d'hypothèse sur un extrait du jeu de données. Il faut alors déterminer si les informations disponibles sont suffisantes à la validation d'une hypothèse. Ceci passe là encore par une appréhension du jeu de données.

Cette appréhension des données, dans la perspective considérée dans ma thèse, implique la résolution d'un ensemble de problèmes complexes :

- ⋆ la recherche et l'extraction des données intéressantes dans le cadre d'une étude particulière, en utilisant des sources d'information potentiellement multiples et distantes,
- ⋆ l'agrégation des informations intéressantes au sein d'un même pool informationnel,
- ⋆ l'organisation des éléments pertinents au sein d'une structure facilitant l'appréhension des données,
- ⋆ la présentation des éléments pertinents et de leur organisation structurelle.

Étant donné la complexité de ces problèmes, il apparaît un besoin croissant d'assistance informatique pour aider les chercheurs à les résoudre. L'objectif de ma thèse est alors de présenter une solution informatisée à cette problématique d'appréhension des données.

La réponse proposée est une notion de synthèse, qui fédère les activités de

recherche et extraction d'informations, agrégation, organisation et présentation des données, qui sont sous-jacentes à la problématique d'appréhension des données. Inspirée des principes de Recherche d'information, cette synthèse se base sur un modèle intermédiaire entre Recherche d'Information classique et vision comportementaliste de l'accès à l'information. Ce modèle donne une place centrale à l'objectif de fouille de données ou à l'hypothèse à tester, définissant une Recherche d'Information orientée tâche.

La notion de synthèse apporte alors un point de vue original sur la Recherche d'Information, en la plongeant comme élément d'une démarche scientifique expérimentale, qui se trouve bouleversée par la présence de masses d'information rendant difficile voire impossible la formulation d'hypothèse et plus généralement l'ensemble du cycle expérimental. Le point de vue tâche sur la Recherche d'Information, encore peu abordé, soulève des problèmes difficiles. Parmi ceux-ci, on peut évoquer la formulation de requête, la notion de modèle de tâche, ou les questions de fusion d'informations.

Dans le cadre de mes travaux, cette notion de synthèse est introduite dans un domaine applicatif particulier, celui de l'exploitation des données acquises dans le cadre de la technologie des Tissue MicroArrays. Ce domaine applicatif, ses caractéristiques vis à vis de l'exploitation des données et les limites des outils informatiques associés selon la perspective d'appréhension des données sont présentés au sein du Chapitre 1.

La description de ce contexte applicatif et des manques constatés en matière d'appréhension de données suggèrent le recours à la notion de synthèse. Celle-ci est introduite au Chapitre 2 comme une notion difficile, aux multiples facettes, qui peut être abordée selon diverses perspectives : appréhension des données qui peut être rapprochée de la fouille de données ou de la Visualisation d'Information, exploration des données, qui relève de la Recherche d'Information, et représentation d'entités, selon des approches de systèmes adaptatifs, ou d'Intelligence Artificielle.

Ces divers points de vue sur la synthèse conduisent au choix d'un point de vue Recherche d'Information, argumenté au Chapitre 3. Celui-ci permet la définition de bases conceptuelles pour la synthèse, par définition d'un modèle de synthèse inspiré des modèles de Recherche d'Information. Les diverses composantes de ce modèle sont détaillées au sein de ce même chapitre.

Un prototype de système, basé sur le modèle de synthèse, permet l'opérationnalisation des paradigmes de Recherche d'Information orientée tâche. Le fonctionnement de ce prototype est l'objet du Chapitre 4.

Des études de cas et une étude utilisateur, présentées au sein du Chapitre 5, permettent une validation expérimentale conjointe du prototype développé et du modèle de synthèse proposé, en tant que solution à la problématique d'appréhension

de gros volumes de données scientifiques.

Les résultats de ces expérimentations permettent tout à la fois la suggestion d'améliorations possibles au prototype, l'ouverture vers d'autres domaines applicatifs et l'extension du modèle de synthèse vers d'autres problématiques telles que la conception d'expériences.

Chapitre

1 Contexte applicatif : Tissue MicroArrays et recherche en oncologie

Ce chapitre expose le contexte applicatif qui a vu l'émergence de mon projet de thèse. Après un bref rappel de l'objectif de la recherche en oncologie, il présente le cadre expérimental mis en place au sein de l'équipe, et en particulier la technologie des Tissue MicroArrays (TMA) et son utilisation dans des études anatomopathologiques. Cette technique vise à dépasser les limites posées par les protocoles de laboratoire classiques par le recours à un traitement de masse d'échantillons biologiques. Mais les méthodes telles que celle-ci posent aux biologistes de nouveaux problèmes liés au volume de données acquises. Ces gros volumes de données font en effet apparaître de nouveaux besoins de stockage et d'exploitation d'informations, passant par leur informatisation. Ils altèrent aussi le processus établi de la démarche expérimentale. Ces constats induisent une prise de conscience d'un manque en ce qui concerne l'appréhension de données TMA, problématique centrale de ma thèse.

1.1 Introduction

La recherche en oncologie, pour proposer des tests de dépistage et protocoles thérapeutiques innovants, passe entre autres par l'étude des mécanismes de la transformation de cellules normales en cellules tumorales, par le biais d'études tissulaires. L'acquisition de telles données anatomopathologiques repose classiquement sur la construction, à partir d'échantillons de tissu d'archive, de lames histologiques sur lesquelles est révélée l'expression de molécules d'intérêt.

Les procédures de ce type se prêtent bien au cadre classique de démarche expérimentale où des hypothèses, relevant d'une question étudiée, sont évaluées par le biais d'expériences menées sur un groupe d'individus restreint. Mais ces procédés sont longs à mettre en œuvre, coûteux en réactifs, et surtout conduisent à l'épuisement de ressources non renouvelables : les échantillons de tissu mis à disposition de la recherche.

Afin de dépasser les limites posées par ce processus, la technologie des TMA, qui permet le traitement en masse de plusieurs centaines de micro-fragments de tissu sur une seule lame, paraît d'un intérêt tout particulier. Mais celle-ci fait émerger de nouvelles limites.

Au niveau du protocole expérimental, des interrogations entourent encore la stratégie d'échantillonnage ou d'organisation des éléments au sein du Tissue MicroArray. Le volume d'informations à traiter, que ce soit pour concevoir les expériences, stocker les données acquises ou les exploiter, implique un support informatique dédié à la technologie.

Enfin, comme dans nombre de domaines de recherche appliquée, et en particulier dans les sciences du vivant, le progrès technologique consistant à proposer de nouvelles méthodes expérimentales à haut débit a conduit à l'émergence de nouveaux besoins informationnels. Ces techniques permettent en effet l'acquisition d'importantes quantités de données, dont en premier lieu le stockage et la gestion, et en second lieu l'exploitation, sont rendus difficiles par leur nature multidimensionnelle, leur hétérogénéité, leur incomplétude, leur incertitude liée aux biais de manipulation ou erreurs de saisie... L'utilisation de ces informations par des méthodes de fouille de données ou dans le cadre d'une démarche expérimentale classique est elle-aussi rendue plus ardue.

Dépasser les problèmes informationnels posés par les technologies permettant le traitement en masse d'échantillons biologiques, et en particulier faciliter l'extraction de connaissances utiles à partir des données acquises, implique de répondre à des besoins d'appréhension de données. En effet, une meilleure appropriation de la collection de données à disposition permet tout à la fois de mettre en place une stratégie de fouille de données adaptée et de définir comment se replacer dans un

cadre expérimental classique. L'objectif de mes travaux de thèse est d'apporter une solution à ce problème.

Dans la suite de ce chapitre, la Section 1.2 introduit quelques notions basiques d'oncologie, les objectifs et quelques pratiques courantes de la recherche contre le cancer, en particulier les techniques classiques d'exploration de tissus, et leurs limites qui ont conduit à l'émergence de la technologie des TMA. Ensuite, la Section 1.3 explore la technique des TMA, de la naissance du concept à ses apports et limites, en passant par une description de la méthodologie. Le volume de données généré par cette méthode a conduit les scientifiques qui y ont recours à mettre en place de nombreux outils informatiques pour traiter cette information, outils qui sont évalués Section 1.4. Les limites de la technologie en elle-même et des outils associés a fait émerger un nouveau besoin qui est esquissé Section 1.5 et constitue la problématique centrale de ma thèse.

1.2 Oncologie et pratiques de recherche courantes

1.2.1 Cancer en tant que contexte de recherche

En France, les cancers représentent la deuxième cause de mortalité, et la première cause de mortalité prématurée. Entre 45 et 64 ans, plus de 45% des décès sont dus aux cancers, les plus courants étant les cancers du sein féminin, de la prostate, du côlon. Les données épidémiologiques de ce type ainsi que l'engagement gouvernemental, par exemple au sein du «Plan Cancer», font des cancers un problème et une priorité majeurs de santé publique.

Le cancer est le résultat d'une dérégulation des systèmes de contrôle de la croissance des cellules, dérégulation qui entraîne la prolifération anarchique et incessante des cellules, la perte des processus de mort cellulaire programmée (ou apoptose), etc. Ce processus aboutit à la formation, au sein du tissu, d'une masse de cellules anormales appelée tumeur. Bénignes, ces tumeurs sont petites, localisées et peuvent être retirées par la chirurgie. Lorsqu'elles sont malignes, ce sont alors des cancers, qui peuvent infiltrer les tissus voisins, récidiver après ablation de la tumeur, ou essaimer par le système circulatoire dans d'autres organes où se forment des métastases.

Dans ce cadre, la recherche en oncologie poursuit un double objectif : l'exploration de stratégies thérapeutiques innovantes et la mise en place d'actions de prévention ciblant les sujets à risques. Ces deux visées impliquent une étude des mécanismes fondamentaux de la transformation tumorale, c'est-à-dire de la métamorphose de cellules normales en cellules cancéreuses.

L'étude du processus de transformation tumorale suppose l'étude des mécanismes

de régulation du cycle cellulaire, mécanismes qui sont perturbés par les cancers. L'objectif poursuivi est l'identification des molécules impliquées, qui sont altérées par la maladie et qui peuvent être la cible de tests de dépistage et traitements. L'observation des interactions entre ces molécules permet aussi de définir des voies de cancérisation, combinaisons de mutations conduisant à des jeux de molécules disfonctionnelles expliquant la transformation d'une cellule normale en une cellule cancéreuse.

Une des techniques d'étude des mécanismes de cancérisation est l'analyse de l'expression de molécules intervenant dans le cycle cellulaire au sein de tissus normaux, pré-cancéreux, cancéreux à divers stades, etc. Pour ce faire, il est courant de recourir à l'anatomopathologie, spécialité médicale dont l'objectif est l'exploration de la composition, de la structure, du renouvellement des tissus pathologiques, ainsi que des échanges cellulaires en leur sein.

C'est dans ce cadre que se placent une partie des travaux réalisés au sein de l'équipe RFMQ (Reconnaissances des Formes et Microscopie Quantitative) du Laboratoire TIMC-IMAG (Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble). Ceux-ci, menés en collaboration avec le CRLCC (Centre Régional de Lutte Contre le Cancer) Val d'Aurelle de Montpellier, portent sur les cancers du côlon et du sein, qui sont parmi les plus courants. Ces travaux visent à analyser, par le biais d'études anatomopathologiques mises en relation avec les dossiers cliniques des patients, l'expression de protéines impliquées dans les processus de cancérisation et leurs relations avec des informations démographiques, diagnostic et pronostic.

Ces recherches sont réalisées en recourant à une technologie d'étude histologique récente, la technique des Tissus MicroArrays. Ce choix a été motivé tout à la fois pour pallier les difficultés rencontrées avec les pratiques classiques d'analyse de tissus et par intérêt pour l'exploration d'une nouvelle méthode de laboratoire. Il convient donc de présenter rapidement les pratiques histologiques utilisées en routine au sein de l'équipe, avant d'exposer les problèmes techniques qui ont motivé le recours à la technique des TMA.

1.2.2 Pratiques classiques d'exploration de tissus

1.2.2.1 Introduction

L'anatomopathologie, en tant que spécialité médicale qui se consacre à l'étude macro- et microscopique des tissus pathologiques, est avant tout l'examen indispensable pour établir un diagnostic et, en complément, intervient parmi les disciplines utilisées dans le cadre de la recherche en oncologie. Ces études tissulaires sont souvent

réalisées à partir de biopsies (fragments de tissus ou d'organes) qui, en oncologie, sont prélevées lors de l'ablation d'une tumeur. Ces fragments sont échantillonnés sur la pièce opératoire, afin de vérifier si l'ensemble des tissus pathologiques ont bien été retirés.

Ces prélèvements sont conservés par inclusion dans des blocs de paraffine ou par congélation. En ce qui concerne les biopsies disponibles pour les travaux de l'équipe, celles-ci sont stockées dans la paraffine. Certaines de ces biopsies peuvent en effet être mises à la disposition des chercheurs, en particulier quand plusieurs blocs et les dossiers cliniques complets des patients sont disponibles.

L'utilisation de ces biopsies dans des études anatomopathologiques nécessite la réalisation d'échantillons observables au microscope optique à partir des blocs, sous la forme de lames histologiques, ainsi que la révélation de l'expression de molécules d'intérêt, par exemple par immunohistochimie. Mais cette technique expérimentale présente un certain nombre de limites que la technologie des TMA vise à dépasser.

1.2.2.2 Technique expérimentale

1.2.2.2.1 Procédure classique d'obtention de lames histologiques

La réalisation d'études anatomopathologiques sur les biopsies d'archive contenant des tumeurs, ou du tissu jugé sain servant de référence, requiert tout d'abord, de façon courante, l'amincissement des échantillons trop épais pour être utilisés tels quels en microscopie optique. Cet amincissement est réalisé par le découpage, à l'aide d'un microtome, de lamelles de tissus très fines (5μm) qui seront ensuite placées sur des lames de verre, ou lames histologiques.

Il faut ensuite procéder à la coloration de chaque lame pour révéler les structures tissulaires et cellulaires ou évaluer la répartition et la concentration de molécules d'intérêt.

1.2.2.2.2 Une méthode courante d'analyse moléculaire : l'immunohistochimie

Une méthode couramment utilisée pour la localisation de molécules particulières dans les échantillons placés sur lames histologiques est l'immunohistochimie. Cette technique est basée sur les principes de la réaction antigène/anticorps du système immunitaire. Elle permet de localiser des antigènes (très souvent des protéines ou fragments de protéines) dans des tissus, cellules, organites cellulaires, etc. Le réactif principal est un anticorps dirigé contre l'antigène à marquer, anticorps qui a

été obtenu à partir de sérum d'un petit mammifère (lapin, cobaye...) dont l'organisme a préalablement été sensibilisé à l'antigène. Des traceurs fixés directement ou indirectement sur cet anticorps permettent de voir la réaction. Par exemple, en immunofluorescence, les anticorps incluent leur système révélateur, car ils sont eux-mêmes fluorescents, alors qu'en enzymo-immunologie, c'est le produit d'une réaction entre une enzyme et le couple antigène/anticorps qui est coloré. Le schéma général d'une réaction immunohistochimique est exposé Fig. 1.1.

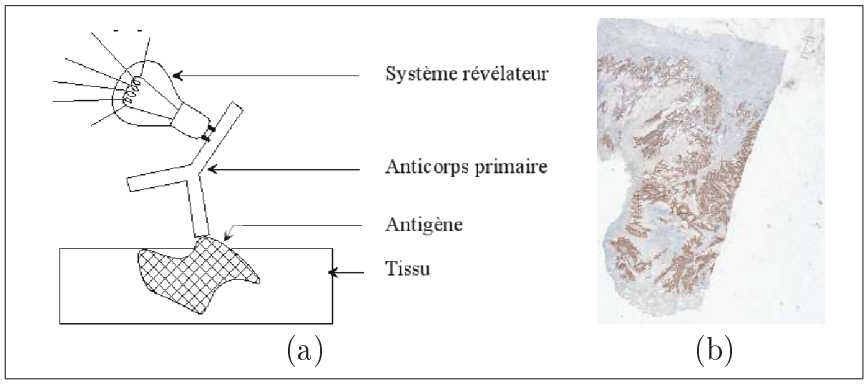


Fig. 1.1: Immnunohistochimie : (a) Les différentes composantes d'une immunoréaction. - (b) Image d'une lame histologique d'un cancer du côlon marquée pour révéler la β-caténine observée au microscope optique grossissement X4.

Une immunoréaction est donc composée de trois éléments principaux :

- ⋆ la préparation (tissu, cellule, organite subcellulaire...) contenant l'antigène à étudier,
- ⋆ un anticorps dirigé contre l'antigène recherché,
- ⋆ le système révélateur qui permet de visualiser l'immunoréaction.

Techniquement, l'échantillon à étudier est placé dans une série de bains. Un bain contient l'anticorps, d'autres permettent de laver l'échantillon pour retirer les molécules qui ne se sont pas fixées, d'autres enfin permettent de fixer le révélateur à l'anticorps, en particulier en enzymo-immunologie, où l'anticorps n'inclut pas son propre système de révélation.

Une observation microscopique des lames histologiques marquées permet d'évaluer la quantité et la localisation des anticorps fixés au sein de la cellule et des tissus, et par extrapolation d'évaluer la présence de l'antigène correspondant. Par abus de langage, anticorps et antigène sont souvent confondus du fait du lien existant entre les deux, et on parle généralement d'un marqueur.

Bien que bien maîtrisées car utilisées en routine depuis de nombreuses années, ces pratiques présentent des limites qui ont conduit à l'émergence de nouveaux protocoles expérimentaux tels que les TMA.

1.2.2.3 Limitations de ces technologies

Ces manipulations, bien qu'elles soient d'un intérêt majeur pour la recherche, présentent plusieurs inconvénients dans un contexte d'exploration des processus de cancérisation. En particulier elles sont coûteuses en tissus d'archive, en réactifs et en temps, tout en induisant une variabilité inter-lames.

Tout d'abord, afin de mener des études complètes en cancérologie, il faut disposer tout à la fois de matériel biologique et d'informations cliniques concernant le patient chez qui a été prélevé la tumeur ou le fragment de tumeur. Or, des échantillons biologiques associés à des dossiers cliniques complets sont des denrées rares. De plus, la coupe successive de lames conduit un jour ou l'autre à l'épuisement du bloc de biopsie dans lequel il n'est alors plus possible de réaliser de nouvelles coupes. Il faut donc utiliser les biopsies avec discernement lors des expérimentations, comme pour toute ressource non renouvelable.

Ensuite, les expériences permettant la mise en évidence de molécules, telles que l'immunohistochimie, ont un coût financier important, lié à l'achat de matériel pour réaliser les coupes (microtome, automate de coloration, etc.), les bains, auxquels s'ajoutent les lames, les réactifs, les anticorps...

De plus, les manipulations sont aussi coûteuses en temps, même si certains matériels permettent le traitement semi-automatique et en masse de lots de lames histologiques pour les étapes de révélation, une fois l'échantillon placé sur une lame vierge.

Enfin, les études incluant un nombre important de lames imposent leur traitement en lots successifs, de par les dimensions des matériels de marquage. Or il existe quasi systématiquement une certaine variabilité de marquage entre lots, même au sein du même laboratoire, avec le même matériel, et des bains réalisés par le même technicien. La comparaison de marquage entre lames appartenant à des lots différents est donc biaisée, certaines lames étant dès l'origine plus marquées que d'autres.

Il paraît donc intéressant de mettre en place des systèmes permettant de réaliser tout à la fois des économies de tissus et des économies de réactifs. La miniaturisation des échantillons et leur traitement en masse est l'une des solutions possibles à ce problème, induisant un compromis entre les économies et une potentielle perte d'information liée à la réduction en taille de l'échantillon. La technologie des Tissue MicroArrays paraît comme une technique qui a trouvé une bonne balance entre ces deux composantes du problème. Mais celle-ci fait émerger de nouveaux besoins d'assistance à sa réalisation et de gestion et exploitation des données, de telles techniques à haut débit bouleversant la pratique expérimentale. La complexité de ces problèmes méthodologiques et d'utilisation des données suggèrent le recours à l'outil informatique.

1.3 Une technologie innovante : les Tissue MicroArrays

1.3.1 Introduction

La technologie des Tissue MicroArray vise à dépasser les problèmes posés par les méthodes classiques d'exploration de tissus par la miniaturisation des échantillons, selon des principes qui ont fait leurs preuves pour le prélèvement d'échantillons cylindriques dans le bois, les sols ou la glace, à l'aide d'une tarière. Au lieu de réaliser des coupes complètes de chaque bloc de paraffine contenant une biopsie, des carottes de tissu sont prélevées au sein de chaque biopsie et sont organisées au sein d'une matrice lignes/colonnes dans un nouveau bloc, qui est alors traité comme un bloc classique.

Les premières publications présentant la technique des Tissue MicroArrays et ses avantages par rapport aux méthodes classiques d'exploration de tissus datent de la fin des années 90, avec des travaux tels que ceux exposés dans [Kononen et al., 1998, Kallioniemi et al., 2001, Bubendorf et al., 2001].

Depuis cette période, des centaines d'articles utilisant ce procédé ont été publiés, majoritairement pour la caractérisation et l'analyse de marqueurs tumoraux. Quelques exemples typiques pourraient être [Lugli et al., 2004] qui explore l'expression d'un marqueur selon l'organe, ou [Wu et al., 2003] qui étudie le rôle d'une molécule particulière dans la carcinogenèse colo-rectale et son lien avec le pronostic des patients. Plus rarement, la technologie est utilisée dans d'autres domaines que l'oncologie, par exemple dans [Wang et al., 2002] qui s'intéresse au domaine de la neuropathologie. Enfin, les Tissue MicroArrays sont aussi proposés comme outils de contrôle qualité au sein du laboratoire [Packeisen et al., 2002], ou entre laboratoires [Diaz et al., 2004], puisqu'ils permettent une évaluation peu coûteuse et rapide de la variabilité de coloration sur des dizaines de types tissulaires à la fois, au sein d'une seule lame.

La technique expérimentale des TMA, décrite plus précisément dans la suite de cette section, commence par la construction d'un plan de fabrication du bloc TMA, puis la réalisation de ce bloc. Le bloc est alors traité comme un bloc biopsie standard, avec coupes et coloration des lames obtenues. L'exploitation des lames passe par une étape extensive d'acquisition de données qui sont alors utilisées dans des études statistiques, d'analyse de données, etc. Mais ces techniques, si elles permettent de dépasser certains problèmes posés par les méthodes classiques, posent encore des problèmes techniques et conceptuels dont la résolution partielle peut éventuellement être apportée par l'outil informatique.

1.3.2 Technique Expérimentale

1.3.2.1 Construction du plan de fabrication du bloc

La première étape de la technique des TMA est la construction du plan de fabrication du bloc. Il s'agit de définir une carte TMA décrivant l'échantillon à placer dans chaque case de la matrice lignes/colonnes du bloc. Illustré par la Fig. 1.2 le processus relaté plus en détails ci-dessous n'est qu'une description technique qui sous-tend un problème de conception complexe.

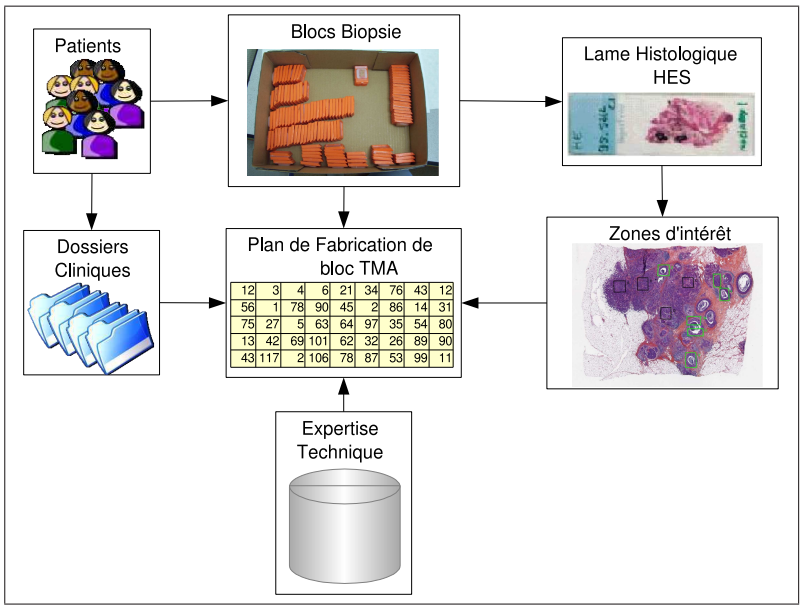


Fig. 1.2: Construction du plan de fabrication TMA - Pour les patients à inclure à l'étude, sont disponibles dossiers cliniques et blocs de biopsie. Sur une lame HES de chaque bloc sont repérées des zones d'intérêt où prélever des carottes. L'expertise technique de l'équipe permet de déterminer quelle carotte prélevée dans quelle zone d'intérêt de quel bloc biopsie donneur devra être insérée à chaque jeu de coordonnées dans le bloc TMA.

La taille de la matrice du bloc TMA (nombre de lignes et colonnes) est définie en fonction du modèle de bloc de paraffine qui recevra les carottes et du diamètre du prélèvement. Il s'agit alors de remplir la carte TMA vide.

Pour ce faire, une cohorte de patients à intégrer est définie. Pour chacun des patients, sont disponibles leurs dossiers cliniques et des blocs de biopsie d'archive. Si elle n'est pas déjà disponible, une lame histologique classique de l'ensemble du bloc biopsie, marquée avec une coloration standard (Hématoxyline-Eosine-Safran ou HES) est réalisée. Cette coloration permet la révélation de la structure du tissu : l'hématoxyline marque les noyaux en violet foncé, l'éosine colore les cytoplasmes en rose et le safran fait ressortir les fibres de collagène en orange. Un anatomopathologiste observe ces lames HES et définit des zones d'intérêt, portions de l'échantillon où il juge qu'il serait pertinent de réaliser des carottes.

L'ensemble des informations disponibles (zones d'intérêt définies par le pathologiste, informations cliniques, etc.) sont alors mises à contribution en conjonction avec l'expertise de l'équipe technique pour déterminer, pour chaque emplacement dans la carte du bloc TMA, quelle carotte devra être insérée à cet endroit. La définition de la carotte consiste entre autres en une description précise de ses coordonnées de prélèvement, au sein d'un bloc biopsie particulier.

La réalisation raisonnée d'un plan de fabrication d'un bloc TMA est un problème complexe qui implique des choix expérimentaux à plusieurs échelles : au niveau patient, il faut définir quels patients intégrer à l'étude ; au niveau bloc biopsie, la question est celle de la définition des zones d'intérêt de leur intégration au bloc TMA ; au niveau d'une zone d'intérêt, il s'agit de choisir une méthode d'échantillonnage et en particulier définir le diamètre des carottes et le nombre de répétitions à réaliser pour être représentatif.

En pratique, ces questions difficiles sont en général laissées de côté et l'objectif est de constituer des blocs TMA pour l'ensemble du matériel biologique disponible, de révéler l'expression du maximum de molécules, constituant ainsi le pool de données le plus complet possible. Le problème se trouve alors déporté au niveau de l'exploitation des données.

1.3.2.2 Construction du bloc TMA

Le plan de fabrication du bloc TMA étant défini, il s'agit alors de réaliser le bloc en tant que tel, selon le processus décrit Fig. 1.3.

Le bloc de paraffine vierge receveur est installé au sein du microarrayer, appareil permettant tout à la fois la perforation du bloc receveur, le prélèvement des carottes dans les blocs donneurs et leur insertion dans les trous du bloc receveur. Le plan de fabrication est parcouru une ligne après l'autre, comme le fait une machine à écrire.

Selon les instructions de ce plan, chaque échantillon est alors prélevé dans le bloc donneur correspondant, sous forme d'une carotte de tissu de 0.6 à 2mm de diamètre, à l'aide du microarrayer. Le trou correspondant est réalisé dans le bloc receveur, puis la carotte est insérée à son emplacement dans le bloc TMA.

Au final, on obtient un jeu de blocs donneurs carottés, mais encore utilisables pour réaliser de nouvelles carottes ou des lames histologiques classiques, ainsi qu'un bloc TMA. La difficulté majeure est ici purement temporelle, la réalisation des blocs étant une tâche longue et fastidieuse.

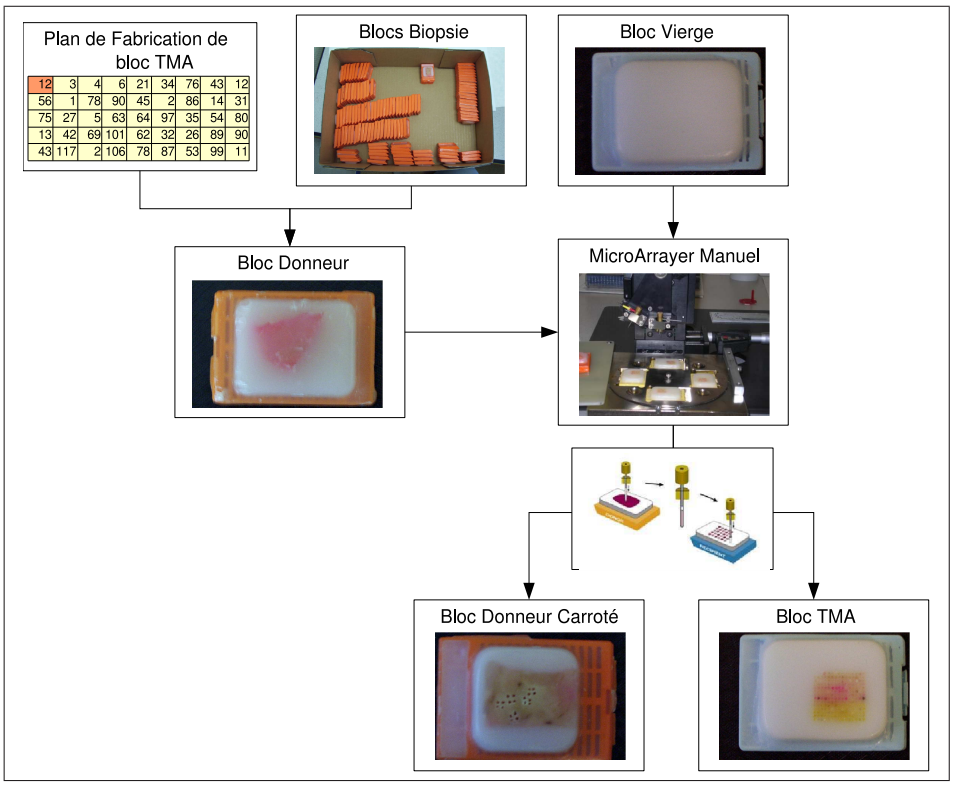


Fig. 1.3: Construction du bloc TMA - Pour chacun des blocs de biopsie et chaque emplacement de prélèvement définis dans le plan de fabrication du bloc, une carotte est réalisée puis insérée à la position prévue avec un microarrayer, ici un modèle manuel. Le résultat final consiste en blocs biopsie perforés et un bloc TMA correspondant au plan de fabrication.

1.3.2.3 Réalisation des lames TMA

1.3.2.3.1 Réalisation des coupes

L'exploitation anatomopathologique du bloc TMA passe tout d'abord par la construction de lames TMA dites blanches, c'est-à-dire sur lesquelles aucune coloration n'a encore été réalisée, selon le processus de la Fig. 1.4.

Pour réaliser ces lames blanches, le bloc TMA est coupé au microtome en lamelles de 5μm d'épaisseur, constituant un ruban de paraffine de coupes sériées. Les lamelles sont ensuite placées sur des lames de verre vierges. Le résultat est une lame TMA présentant une matrice de spots, chaque spot correspondant à la coupe d'une carotte.

La finesse des coupes rend leur transfert sur la lame vierge délicat : la perte ou l'altération de spots n'est pas rare. L'utilisation d'un scotch spécial collé sur la coupe lors des manipulations permet de limiter la disparition ou la casse des micro-échantillons. Mais le recours à cette pratique ne garantit pas un résultat parfait.

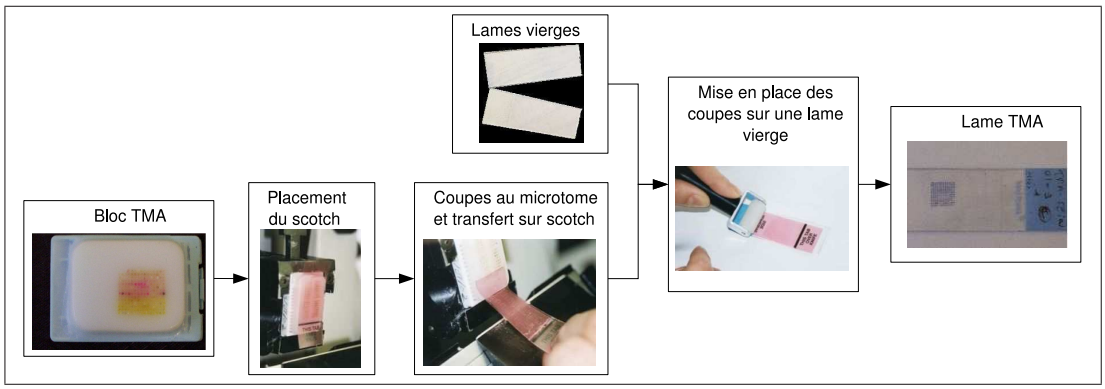


Fig. 1.4: Construction des lames TMA - Des coupes sériées sont réalisées dans le bloc à l'aide d'un microtome, puis positionnées sur des lames de verre et collées.

1.3.2.3.2 Réalisation de l'immunomarquage

L'expression de diverses molécules impliquées dans les processus de cancérisation sont alors révélées comme illustré Fig. 1.5.

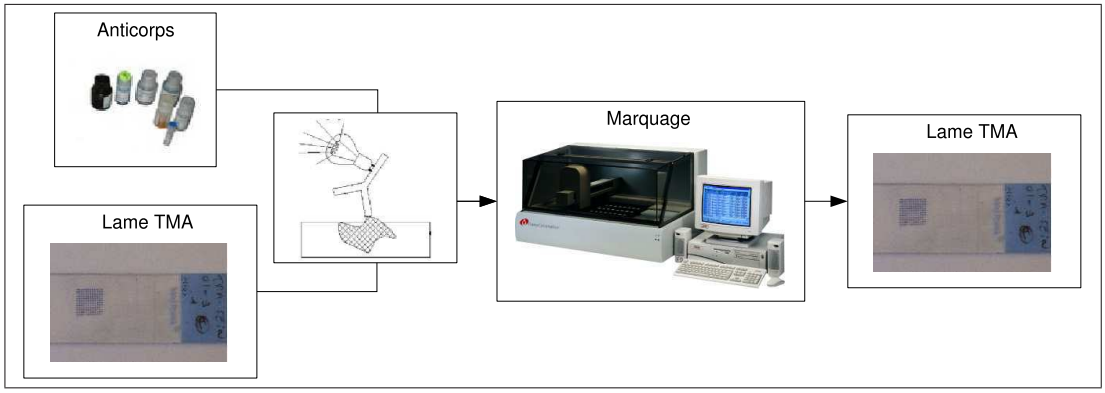


Fig. 1.5: Marquage des lames TMA - L'expression de molécules d'intérêt est révélée par immunomarquage en utilisant un automate de coloration.

À l'aide d'un automate de coloration, qui permet le traitement en masse des lames et un contrôle précis de la durée des bains, les lames sont exposées à des anticorps spécifiques des molécules à étudier, selon les mécanismes d'immunohistochimie décrits dans le Paragraphe 1.2.2.2.2. Pour les molécules étudiées au laboratoire, le système révélateur est une réaction enzymatique qui conduit à une coloration marron du couple antigène/anticorps d'intérêt. Le dépôt d'une résine synthétique permet la protection des lamelles lors des manipulations futures.

Mais, malgré le recours à un automate, les colorations ne sont pas pour autant nécessairement homogènes d'un lot à l'autre, d'une lame à l'autre et même d'une partie de lame à l'autre.

1.3.2.3.3 Molécules étudiées

Dans le cadre des travaux menés au sein de l'équipe, quatre molécules impliquées dans les processus de cancérisation ont été étudiées : la β-caténine, la Cycline D1, le Ki67 et le Bcl2. Le Tab. 1.1 résume les connaissances à propos de ces molécules. Leur rôle au sein de ces processus est en effet bien connu :

- ⋆ Certaines de ces molécules sont liées à l'augmentation de la prolifération cellulaire caractéristique des cancers :
  - ⋆ *β-caténine* : cette protéine est produite à proximité de la membrane des cellules où elle joue un rôle dans l'adhésion cellulaire. Dans les cellules normales, elle est rapidement détruite dans le cytoplasme à proximité de la membrane et elle atteint rarement le noyau. Dans les cellules tumorales, son processus de destruction est altéré et elle migre à travers le cytoplasme jusqu'au noyau. Dans le noyau, elle active les gènes de molécules de la famille des cyclines, causant la production de molécules de cyclines,
  - ⋆ *Cycline D1* : c'est une protéine de la famille des cyclines dont la production est activée par la β-caténine. Elle est produite dans le cytoplasme et migre vers le noyau où elle initie la division cellulaire,
  - ⋆ *Ki67* : cette protéine est présente dans le noyau des cellules actives, c'est-à-dire des cellules qui sont en cours de division ou se préparent à se diviser. Même si sa fonction exacte n'est pas précisément connue, elle est utilisée comme marqueur de la prolifération cellulaire : plus le nombre de cellules où le marquage en Ki67 est supérieur à 0 est grand, plus le nombre de cellules engagées dans le cycle cellulaire est grand,
- ⋆ La dernière molécule est liée à la perte des capacités d'apoptose, soit de mort cellulaire : *Bcl2* est une protéine dont la présence dans le cytoplasme de la cellule empêche sa mort.

Tab. 1.1: Résumé des faits biologiques connus à propos des molécules étudiées.

Molécule	*Compartiment cellulaire*	*Rôle*
β-caténine	Membrane, Cytoplasme, Noyau	Cellules normales : produite près de la membrane et détruite - Cellules tumorales : migre à travers le cytoplasme jusqu'au noyau où elle initie a production des cyclines
Cycline D1	Noyau, Cytoplasme	Produite dans le cytoplasme puis migre vers le noyau - Noyau : initie la division cellulaire et ainsi contribue à la prolifération cellulaire
Ki67	Noyau	Marque les cellules en division et est ainsi le témoin de la prolifération cellulaire
Bcl2	Cytoplasme	Empêche la mort cellulaire

1.3.3 Exploitation des lames TMA

1.3.3.1 Acquisition des données

L'acquisition de données à partir des lames qui ont été produites passe ensuite par une série d'étapes présentées Fig. 1.6. La première étape consiste en une acquisition d'images des lames à divers grossissements grâce à un système automatisé qui couple un microscope optique à une caméra et asservit leur contrôle à un programme informatique. Les images sont alors automatiquement partitionnées en images individuelles de spots, correspondant à la coupe de chaque carotte. Sur chaque image, diverses mesures de quantification de marquage sont évaluées :

- ⋆ *Hétérogénéité* : il s'agit d'une évaluation qualitative de la répartition du marquage au sein de l'échantillon. Cette hétérogénéité est subdivisée en quatre classes : homogène, dispersé, pavage et hétérogène,
- ⋆ *Intensité* : c'est une mesure du niveau de marquage pour les zones marquées, fournissant une idée de la concentration en molécule cible, de faible (couleur faible donc intensité faible) à forte (couleur intense donc intensité forte),
- ⋆ *Pourcentage de cellules marquées* : il s'agit d'une évaluation du rapport entre le nombre de cellules exprimant la molécule cible et le nombre total de cellules observées.

Ces mesures sont réalisées par une quantification manuelle, qui repose sur l'observation des images par un anatomopathologiste, et une quantification automatique, par le biais de programmes informatiques. La quantification de marquage peut être conduite spot par spot, ou au sein de lames TMA virtuelles qui regroupent plusieurs images de spots, issues potentiellement de lames différentes, au sein d'une même matrice. L'ensemble des informations (images et mesures) générées lors de cette phase sont stockées dans une base de données avec les informations cliniques des patients correspondants, pour une utilisation au sein d'études biologiques futures.

Les limites majeures à ce niveau concernent l'acquisition d'images et surtout l'évaluation manuelle du marquage, qui sont des tâches longues à réaliser. De plus, partitionnement des images et quantification automatique sont des problèmes difficiles à résoudre par des systèmes informatiques et sujets à erreurs.

1.3.3.2 Utilisation des données

Dans un contexte de recherche en oncologie, les données ainsi acquises doivent éventuellement permettre l'exploration du rôle des molécules étudiées dans la transformation tumorale ou l'identification de nouveaux marqueurs tumoraux, molécules impliquées dans le processus de cancérisation pouvant servir de cibles pour des traitements, des tests de dépistage ou pouvant être utilisées comme indicateurs prédictifs de pronostic pour les patients.

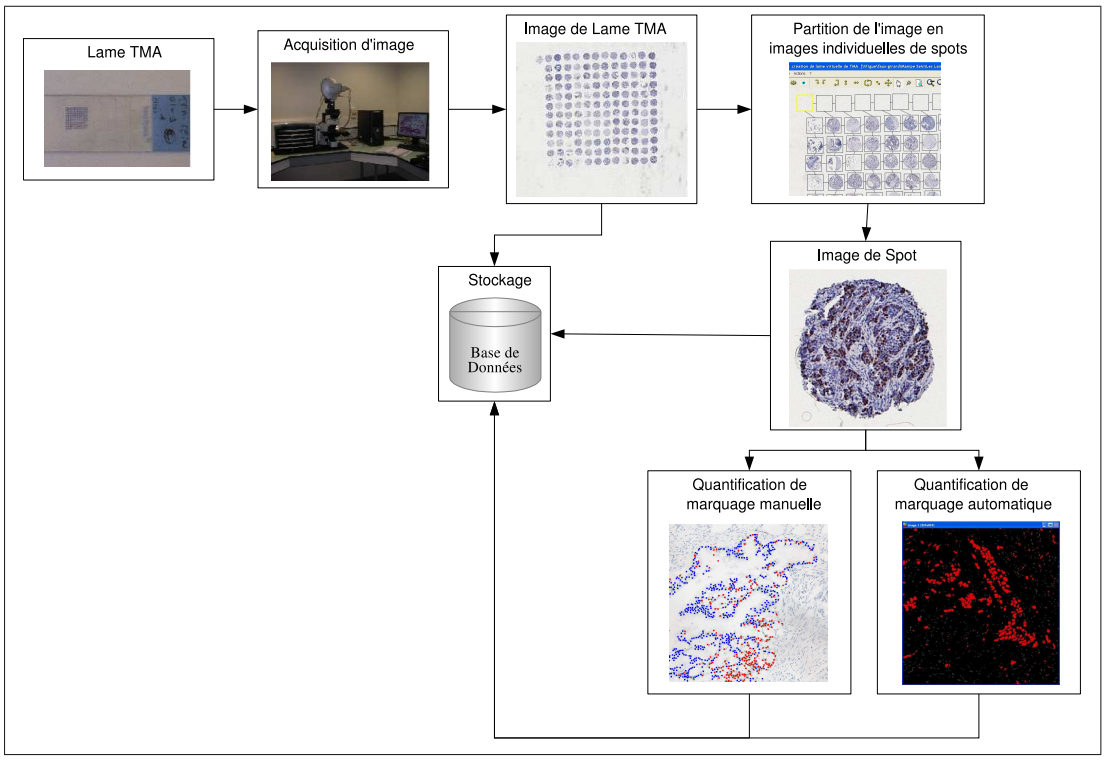


Fig. 1.6: Acquisition des données TMA - Après acquisition d'images de lames TMA et leur partition en images individuelles de spots, le marquage est quantifié automatiquement et par un anatomopathologiste. L'ensemble des données est stocké dans une base de données pour utilisation future.

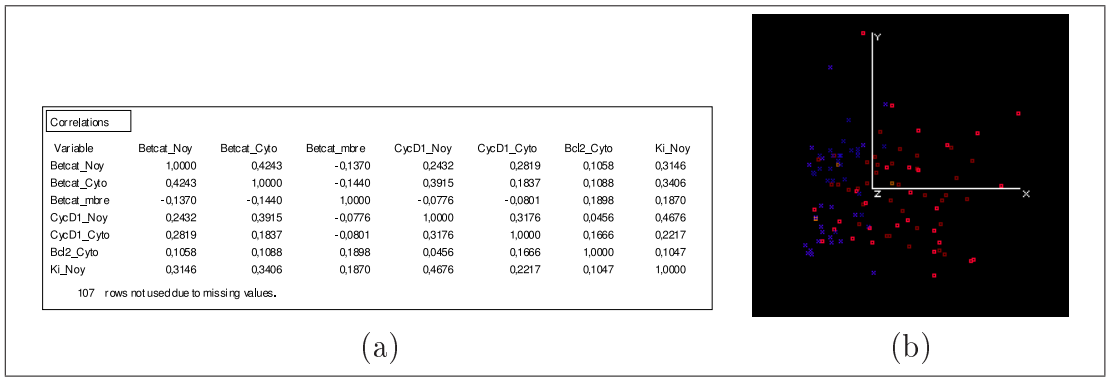
Or, le traitement en masse des échantillons place le biologiste face à de gros volumes de données hétérogènes, qui ont été construits hors de toute démarche expérimentale classique. Les méthodes standard de validation d'hypothèse par des tests statistiques sur des données spécifiquement acquises dans ce but ne sont alors plus applicables et il faut en général se tourner vers d'autres méthodes permettant la mise en lumière de structures, corrélations ou autres. L'extraction de telles connaissances relève du champ de la fouille de données, thématique présentée plus en détails Paragraphe 2.2.2.

En pratique, le recours à des outils de fouille de données pour le traitement des données TMA implique alors de constituer des fichiers de données «nettoyées». Cette notion de «nettoyage» sous-entend des données homogènes, des données valides, c'est-à-dire dans des plages de valeurs cohérentes, des données présentes pour tous les individus, etc. Ces fichiers doivent souvent être dans un format particulier, au format texte ou Microsoft Excel par exemple, afin de les soumettre aux logiciels de traitement. Ces logiciels quant à eux doivent être judicieusement choisis en fonction de la structure et de la nature des données à analyser.

Ainsi, par exemple, au sein de l'équipe, nous avons réalisé un fichier Microsoft Excel à partir de notre base de données TMA regroupant l'ensemble des patients

atteints d'un cancer du côlon et des évaluations manuelles de quantification de marquage pour quatre molécules d'intérêt, évaluations réalisées en fonction de la localisation par rapport à la tumeur et de la localisation intracellulaire du marquage. Ce fichier a servi de base à plusieurs traitements relevant de la fouille de données.

Par exemple a été appliquée une méthode qui relève de l'analyse de données : l'Analyse en Composantes Principales (ACP) est une méthode mathématique qui consiste à rechercher les directions de l'espace qui représentent le mieux les corrélations entre n variables aléatoires. Les résultats de cette ACP sont présentés Fig. 1.7.



Correlations

Variable	Betcat_Noy	Betcat_Cyto	Betcat_mbre	CycD1_Noy	CycD1_Cyto	Bcl2_Cyto	Ki_Noy
Betcat_Noy	1,0000	0,4243	-0,1370	0,2432	0,2819	0,1058	0,3146
Betcat_Cyto	0,4243	1,0000	-0,1440	0,3915	0,1837	0,1088	0,3406
Betcat_mbre	-0,1370	-0,1440	1,0000	-0,0776	-0,0801	0,1898	0,1870
CycD1_Noy	0,2432	0,3915	-0,0776	1,0000	0,3176	0,0456	0,4676
CycD1_Cyto	0,2819	0,1837	-0,0801	0,3176	1,0000	0,1666	0,2217
Bcl2_Cyto	0,1058	0,1088	0,1898	0,0456	0,1666	1,0000	0,1047
Ki_Noy	0,3146	0,3406	0,1870	0,4676	0,2217	0,1047	1,0000

107 rows not used due to missing values.

(a) (b)

Fig. 1.7: Utilisation des données TMA : ACP - Résultat d'une Analyse en Composantes Principales réalisée sur un groupe de patients atteints d'un cancer du côlon. Les variables étudiées sont les pourcentages de cellules marquées pour 4 marqueurs, selon la localisation intracellulaire du marquage : (a) Résultats numériques présentant la matrice des corrélations entre variables analysées - (b) Représentation graphique correspondante : les points rouges représentent du tissu tumoral et les points bleus du tissu adjacent à la tumeur. L'axe x correspondent à la première composante principale, l'axe y à la seconde et l'axe z à la troisième.

De façon très succincte, cette première analyse permet de montrer une assez bonne ségrégation entre tissus tumoraux et tissus adjacents à la tumeur sur la base des expressions moléculaires évaluées. En particulier, le pourcentage de cellules marquées en Cycline D1 nucléaire, en β-caténine nucléaire et membranaire, et en Ki67 nucléaire sont des mesures significativement discriminantes.

La difficulté majeure rencontrée à ce stade est le besoin d'une expertise conjointe :

- ⋆ des problématiques biologiques sous-jacentes aux données, afin d'être en mesure de diriger l'analyse et d'évaluer l'intérêt et la signification des résultats,
- ⋆ des outils de fouille, afin de sélectionner des méthodes adaptées aux données et à l'objectif poursuivi.

Ceci induit en général une collaboration entre biologiste et spécialiste de la fouille de données. De plus, la préparation des données est en général une activité longue à réaliser.

1.3.4 Apports et limites de la technologie des TMA

1.3.4.1 Résolution des problèmes posés par les lames histologiques standard

Le recours à la technologie des TMA permet de résoudre les problèmes posés par la procédure classique de construction de lames histologiques exposée précédemment Paragraphe 1.2.2.3.

En effet, en ce qui concerne l'utilisation des tissus d'archive, la réalisation de carottes de tissu permet d'échantillonner les biopsies. Cet échantillonnage fournit ainsi des représentants de chaque biopsie utilisables dans le cadre d'une étude tout en préservant le bloc d'origine. Ceci permet de réaliser des économies de tissus.

De plus, étant donné que chaque bloc TMA inclut plusieurs centaines de carottes, une seule lame est nécessaire là où il en aurait fallu plusieurs centaines, d'où une économie de réactifs et de temps de traitement des coupes.

Enfin, le problème d'hétérogénéité de marquage entre lames traitées dans des lots différents est lui aussi résolu : tous les échantillons à colorer se trouvant sur un nombre très faible de lames, celles-ci peuvent être systématiquement marquées au cours de la même manipulation.

Mais même si elle permet de résoudre certains des problèmes posés par les procédures classiques, la technologie des TMA pose de nouvelles difficultés qui lui sont propres, problèmes techniques ou liés à l'exploitation des données. En particulier, le volume de données pose des questions d'appréhension des données, qu'elle soit préalable à une fouille ou dans l'objectif de se replacer dans une démarche expérimentale classique. Ces diverses limites sont exposées dans le prochain paragraphe.

1.3.4.2 De nouvelles limitations

1.3.4.2.1 Problématiques techniques

Le premier jeu de problèmes posés par la technologie consiste en difficultés rencontrées lors de la réalisation de TMA au laboratoire. Cette technique étant relativement récente, l'expertise à ce sujet est encore peu partagée. Or, nombre d'interrogations entourent la méthode expérimentale, telles que sa validation pour divers types de tissus, la procédure d'échantillonnage, la conception des blocs TMA ou les méthodes permettant de réduire le temps de fabrication.

La première question qui peut être posée est la validité de la méthode présentée

dans d'autres cas que ceux exposés dans les premiers articles présentant la technologie. En effet, ceux-ci évoquent son utilisation avec des échantillons de tumeur du sein ou de la prostate, par exemple, mais qu'en est-il des autres types de tissus ? De plus, dans ces premiers travaux, les tissus échantillonnés sont stockés dans des blocs de paraffine, mais tous les échantillons disponibles pour la recherche ne sont pas sous cette forme : certains sont congelés, etc. Quelques travaux explorent ces sujets, en réalisant des études de faisabilités dans des cas particuliers. Par exemple, [Jourdan et al., 2003] valide l'utilisation de la technologie pour les cancers colo-rectaux, ou [Fejzo and Slamon, 2001] détermine si la méthode peut convenir pour des tissus qui ne sont pas conservés dans des blocs de paraffine, dans ce cas des tissus congelés. Mais ce type de validation reste limité.

Ensuite, une des problématiques les plus critiques est une problématique d'échantillonnage. La technologie des TMA repose sur la réalisation de carottes de très petit diamètre au sein de biopsies, qui sont elles-mêmes des échantillons de tissus hétérogènes dans les trois dimensions : la surface de coupe est hétérogène et cette hétérogénéité est aussi présente dans la profondeur. Dans le cas de coupes de blocs complets, l'hypothèse de similarité entre coupes successives est à peu près vérifiée, mais qu'en est-il pour des carottes d'environ 1mm de diamètre ? Se posent alors les questions du diamètre idéal des carottes et du nombre qu'il faut en prélever pour satisfaire une contrainte de représentativité. Une méthode consiste à évaluer le nombre nécessaire et suffisant de carottes pour obtenir le même résultat qu'avec des lames histologiques classiques [Camp et al., 2000]. Une autre approche consiste à estimer le nombre de carottes nécessaires à la mise en évidence d'une association connue entre une protéine et une variable histologique ou clinique [Rubin et al., 2002, Torhorst et al., 2001]. Les conclusions très variables de ces études suggèrent qu'il n'y a pas de solution unique au problème d'échantillonage.

De plus, se pose aussi un problème d'organisation des carottes au sein du bloc receveur. Tous les échantillons n'ont pas la même densité et sont plus ou moins «durs», les carottes les plus dures pouvant poser problème lors de la coupe, conduisant à des pertes de spots. Faut-il les regrouper ou au contraire les séparer au maximum ? Vaut-il mieux regrouper les carottes issues d'un même bloc donneur, ce qui limite le nombre de manipulations de blocs par le technicien, ou au contraire les séparer au maximum pour limiter les risques de pertes de spots ? Vaut-il mieux organiser les carottes de façon à faciliter le travail du technicien (par exemple en les ordonnant en fonction du rangement des blocs donneurs dans la boîte d'archive) ou de façon à faciliter le travail de lecture de l'anatomopathologiste, en prenant en compte des critères d'organisation liés à l'étude à réaliser (par exemple, l'étude compare des tumeurs issues de différents organes, donc les carottes sont regroupées par organe).

Enfin, l'ensemble du processus reste très coûteux en temps. Le Tab. 1.2 présente un ordre de grandeur du temps passé par les techniciens de l'équipe pour réaliser un bloc TMA.

Tab. 1.2: Construction de bloc TMA : Temps passé - Les estimations de temps passé à la mise en œuvre de la technique fournies ici concernent la construction de quatre répétions de trois blocs TMA différents regroupant 119 échantillons de tumeurs du sein, et la réalisation et l'analyse des lames TMA correspondantes, sur lesquelles a été révélée l'expression de 6 molécules d'intérêt.

Étape du processus	*Temps passé*
Construction du plan de fabrication du bloc :	*26 jours*
- Réalisation des coupes des blocs biopsie	6 jours
- Coloration des lames HES	3 jours
- Définition des zones d'intérêt	15 jours
- Réalisation du plan sous Microsoft Excel	2 jours
Construction du bloc TMA :	*20 jours*
Réalisation des lames TMA :	*15 jours*
- Réalisation des coupes	9 jours
- Réalisation de l'immunomarquage	6 jours (1 jour par marqueur)
Exploitation des lames TMA :	*68 jours*
- Acquisition d'images grossissement X4 et X20	12 jours (2 jours par marqueur)
- Partition des images	1 jour (1h par marqueur)
- Quantification de marquage automatique	4 jours (6h par marqueur)
- Quantification de marquage manuelle	48 jours (1 jour par lame)
- Analyse statistique des données	3 jours
Total :	*129 jours soit 6 mois*

La réalisation d'études utilisant la technologie des TMA apparaît donc une pratique difficile. Un certain nombre d'interrogations pourraient être levées par l'acquisition d'expertise technique supplémentaire, permettant ainsi tout à la fois une plus grande rapidité d'exécution des tâches de laboratoire et la mise en place de bases de connaissances pouvant guider le technicien, dans la conception des blocs par exemple.

L'outil informatique, en fournissant des outils d'assistance à la réalisation des blocs TMA ou à l'acquisition des données, qu'il s'agisse de simples systèmes permettant la documentation des tâches réalisées, pilotant des tâches répétitives, ou fournissant une aide à la décision, apparaît comme une solution possible à certaines des difficultés rencontrées.

1.3.4.2.2 Une altération de la démarche expérimentale

En plus des problèmes techniques qui viennent d'être évoqués, se pose aussi un problème plus fondamental au niveau de la démarche expérimentale en elle-même.

En effet, de manière schématique, la recherche scientifique dans des domaines tels que la biologie, où la démarche est très empirique, se base souvent sur une méthode hypothético-déductive incluant trois temps forts : l'identification d'une question à étudier, la formulation d'hypothèses dans le cadre de cette question et

l'argumentation permettant d'accepter ou rejeter les hypothèses. Ce dernier point est entre autres basé sur l'expérimentation. Ce processus simplifié est présenté Fig. 1.8.

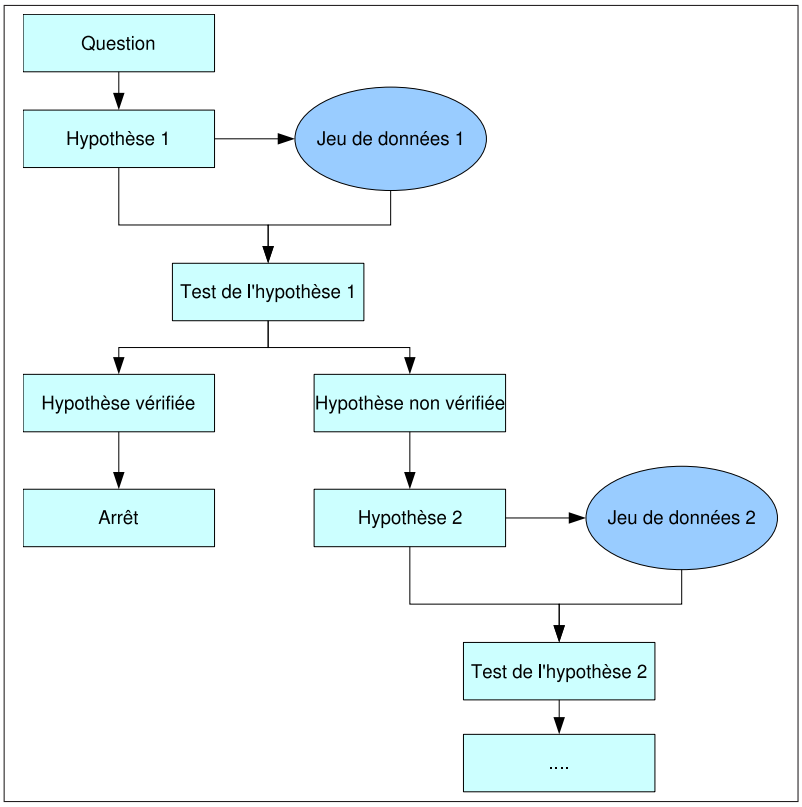


Fig. 1.8: Représentation schématique de la démarche expérimentale - Dans le cadre d'une question à étudier, une première hypothèse est formulée. Un jeu de données permettant d'évaluer l'hypothèse est constitué, par exemple des mesures d'expression de molécules réalisées dans le cadre d'une étude anatomopathologique. Celle-ci est testée, par exemple par des analyses statistiques. Si l'hypothèse est invalidée, de nouvelles hypothèses sont formulées pour expliquer le phénomène observé, sinon, la question est itérativement étudiée selon d'autres points de vue, correspondant à d'autres phénomènes, par des hypothèses différentes.

Les études anatomopathologiques réalisées avec des méthodes classiques, c'est-à-dire des lames histologiques complètes, rentrent plutôt bien dans ce cadre de démarche expérimentale. Par exemple, si la question concerne les facteurs influant les cancers, je peux formuler l'hypothèse que l'âge du patient joue un rôle et réaliser un jeu de lames histologiques de patients de diverses classes d'âge révélant l'expression de diverses molécules impliquées dans la cancérisation. Des études statistiques courantes permettront alors de valider ou invalider l'hypothèse du rôle de la classe d'âge.

Dans le cadre d'études utilisant la technologie des TMA, une telle démarche impliquerait la conception raisonnée de blocs TMA, conception guidée par l'étude à réaliser. Ce problème complexe commence tout juste à être évoqué, par exemple dans des travaux comme [Kajdacsy-Balla et al., 2007], qui définit diverses formes de TMA, selon l'objectif à atteindre : étude de pronostic, de la progression tumorale,

des stades de cancérisation, de l'hétérogénéité tissulaire.

Mais, dans le cadre d'études utilisant une technologie permettant un traitement en masse des échantillons telle que celle des TMA, le recours à une telle démarche est rarement possible. En effet, un bloc TMA peut inclure jusqu'à 600 carottes de tissu et il serait ridicule de réaliser un bloc de faible taille. Or certaines populations de patients ont une taille très réduite : certains types de cancers sont très rares. De même, il est peu courant de disposer d'échantillons de tumeurs aux stades précoces, celles-ci étant rarement diagnostiquées si tôt ou peu traitées par chirurgie. Construire des blocs représentant de tels ensembles ferait perdre tout l'avantage du traitement en masse de la technologie. Il est donc de pratique courante de réaliser, avant même la formulation d'hypothèses, des blocs TMA incluant tout le matériel biologique disponible.

Le biologiste se trouve alors confronté à une masse d'informations hétérogènes se rapportant à une question biologique générale. Certaines sont des données quantitatives (par exemple un pourcentage de cellules marquées) alors que d'autres sont qualitatives (par exemple une classe d'hétérogénéité de marquage). Ces données sont aussi de types très variables (texte, nombres, images...). Enfin, ces données sont acquises à des échelles différentes (au niveau patient, comme son âge ; au niveau organe, comme la partie de l'organe où se trouve la tumeur ; au niveau tissu, comme le degré d'envahissement tissulaire de la tumeur ; au niveau cellule, comme la localisation intracellulaire du marquage...). Ce nouveau contexte d'exploitation des données, où l'extraction de connaissances biologiques à partir des données acquises est problématique, est illustré Fig. 1.9.

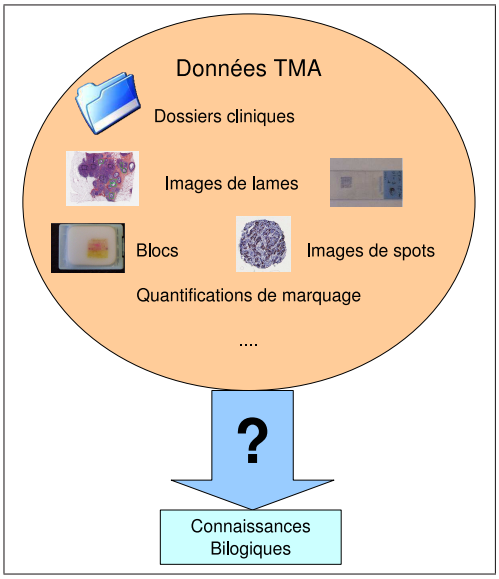


Fig. 1.9: Représentation schématique de l'altération du contexte expérimental - Les données, hétérogènes, à diverses échelles, et en volume important, sont présentes de façon préalable à toute formulation d'hypothèse. Se pose alors la question de leur mode d'exploitation pour en extraire des connaissances biologiques.

Dans ce contexte, le biologiste peut adopter diverses stratégies non exclusives, présentées Fig. 1.10.

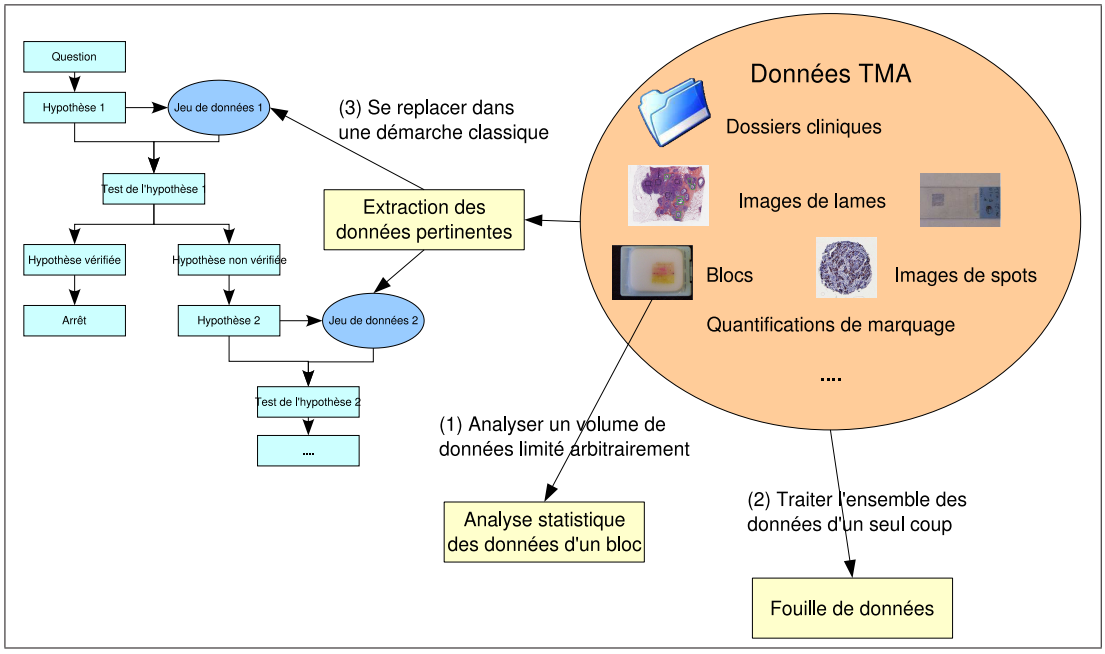


Fig. 1.10: Altération de la démarche expérimentale par la technologie des TMA - Confronté à un gros volume de données complexes, le biologiste peut essayer de réduire la taille de la collection à étudier en se limitant aux données issues d'un seul bloc TMA (1). Il peut aussi recourir à des méthodes de fouille de données adaptées aux gros volumes de données sur le jeu de données complet (2). Il peut enfin essayer de se replacer dans le contexte classique de démarche expérimentale, ce qui ouvre de nouvelles questions (3), entre autres le choix des données à utiliser pour valider une hypothèse donnée, etc.

La première option serait de limiter arbitrairement le volume de données à étudier, en réduisant l'analyse à l'étude d'un seul bloc TMA. Dans ce contexte, les individus à considérer correspondent aux carottes et les variables aux mesures de marquage réalisées sur les diverses lames TMA correspondantes. Ainsi, le jeu de données retrouve une taille raisonnable dans laquelle les méthodes statistiques et d'analyse de données généralement utilisées dans le cadre d'une démarche expérimentale standard sont facilement mises en pratique. Mais ce procédé conduit souvent à la construction d'une population hétérogène et inadaptée pour tester certaines hypothèses : certains individus sont clairement hors du cadre de l'étude, tandis que d'autres potentiellement intéressants ne sont pas inclus car les échantillons ont été traités dans un autre bloc TMA. Cette technique paraît donc peu pertinente et les deux autres solutions semblent les plus intéressantes mais aussi les plus compliquées au niveau traitement de l'information.

La seconde option serait de directement recourir sur l'ensemble des données disponibles à des méthodes de fouille de données de type exploratoire, algorithmes d'apprentissage automatique ou de reconnaissance des formes par exemple, ce qui implique l'usage d'outils informatiques spécialisés. Une fouille de données raisonnée

passe généralement par une étape d'appropriation de la collection de données par l'utilisateur, stade d'analyse descriptive permettant la définition des pré-traitements nécessaires pour l'utilisation des méthodes pertinentes étant données la structure, la nature, etc. de la collection. Ceci induit la mise en place d'outils assistant le biologiste dans son exploration manuelle de la base de données.

La dernière solution consisterait à dépasser les limites posées par l'option précédente, en considérant l'ensemble du jeu de données et en extrayant uniquement celles qui sont pertinentes par rapport à l'hypothèse à tester. Ceci revient à tenter de replacer le jeu de données dans un contexte de démarche expérimentale classique. Une telle démarche implique de pouvoir répondre à des types de questions totalement nouvelles :

- ⋆ quelles sont les hypothèses qui pourraient éventuellement être testées étant donné le jeu de données disponible ?
- ⋆ étant donnée une hypothèse qui n'est pas explorable avec le jeu de données disponible, quelles sont les données à acquérir pour tester l'hypothèse ?
- ⋆ quelles sont les données à utiliser pour tester une hypothèse particulière ?

Ce type de question ne peut pas trouver de réponse par un accès direct au jeu complet de données, du fait de leur quantité et de leur complexité. Des outils d'analyse descriptive de la collection de données doivent être mis à disposition du chercheur, afin de l'aider à s'approprier le jeu de données et lui fournir des indications pour répondre à ces interrogations.

1.3.4.2.3 Une ébauche de solution

Les paragraphes précédents, en mettant en exergue les difficultés liées à l'utilisation de la technologie des TMA, ont permis de suggérer l'intérêt potentiel d'une informatisation de la technologie. Comme il a été montré deux grandes classes de limites à la technique, il apparaît deux grandes classes de motivations pour la mise au point d'outils informatiques autour de la technologie des TMA.

Les premières sont des motivations techniques, face à un besoin de support à la mise en œuvre d'une technologie posant des problèmes de réalisation complexes et impliquant la gestion de gros volumes d'informations.

Les secondes recouvrent un besoin de systèmes permettant à l'utilisateur d'appréhender la collection de données, tout à la fois en tant qu'étape préalable à d'autres méthodes de fouille de données et en tant qu'outil permettant la constitution de populations sur lesquelles des hypothèses peuvent être testées par des méthodes statistiques classiques.

La prochaine section va explorer le champ des outils existants afin d'évaluer les besoins identifiés ici qui sont couverts par ces systèmes informatiques et ceux qui

restent à supporter.

1.4 Outils informatiques associés à la technologie des TMA

1.4.1 Introduction

Étant donné le volume de données générées et les difficultés techniques et d'exploitation des données rencontrées, un besoin d'une assistance informatique autour de la technologie a rapidement vu le jour parmi les scientifiques utilisant les TMA et les outils informatiques dédiés sont de plus en plus nombreux.

Ces outils visent soit à apporter un support guidé par ordinateur aux diverses étapes de réalisation des TMA, afin de réduire le temps passé à leur réalisation manuelle, soit à optimiser le stockage des données et leur accès, soit à faciliter l'exploitation des données. Les logiciels existants, qu'il s'agisse d'outils académiques ou commerciaux, interviennent donc aux diverses étapes de la technique : conception et réalisation des blocs TMA, acquisition et traitement d'images, gestion des données, exploitation des données, qu'il s'agisse de leur visualisation ou de leur analyse. Certains se focalisent sur certaines étapes de la méthode, d'autres visent à fournir une plateforme intégrée couvrant l'ensemble du processus.

Ces divers outils, s'ils apportent une solution à des problèmes soulevés par l'utilisation de la technologie, laissent un certain nombre de questions en suspens, en particulier celles de l'appréhension des données. La résolution de ce type de problématiques est l'objectif de ma thèse.

1.4.2 Accompagnement informatisé de la méthode de laboratoire

1.4.2.1 Conception et construction des blocs

Parmi les différentes étapes de la technologie des TMA, la conception de plans de fabrication des blocs et l'application de tels plans pour la construction de blocs physiques sont parmi les plus complexes et coûteuses en temps. La conception des blocs implique en effet la mise en relation de nombreuses informations sur les blocs donneurs disponibles et les dossiers cliniques correspondants, et le recours à une importante expertise technique. La construction des blocs selon les plans induit de

nombreuses manipulations fastidieuses avec un microarrayer manuel.

Les fabricants de matériel proposent quelques systèmes automatisant partiellement le processus. Ainsi, Becheer Instruments[1] commercialise un microarrayer automatique qui permet, à partir de la définition manuelle du plan de fabrication, une construction du bloc TMA entièrement guidée par ordinateur. Dans le même esprit, Alphelys[2] a conçu un système motorisé qui s'adapte aux microarrayers manuels offrant les mêmes fonctionalités que les microarrayers automatiques.

Ce type de systèmes, s'ils permettent un gain de temps dans la fabrication des blocs, présentent toutefois quelques limites. Tout d'abord, s'ils facilitent la gestion du plan de fabrication bloc, ils ne proposent aucune assistance à sa conception. De plus, à ma connaissance, ils ne permettent pas la mise en relation du plan avec les images des lames fabriquées selon ce plan et les mesures et informations cliniques associées, ce qui serait pourtant d'un grand intérêt.

1.4.2.2 Acquisition et traitement d'images

La seconde étape de la technologie, temporellement coûteuse et bénéficiant de plus en plus d'une informatisation, est celle de l'acquisition et du traitement des images de lames TMA. Un prérequis à cette informatisation est la validation de l'utilisation d'images de lames et spots TMA pour des évaluations anatomopathologiques. Une telle validation a été réalisée par exemple pour des images de spots de cancer de la prostate dans le contexte d'une application Web spécifique par [Bova et al., 2001]. De nombreuses compagnies, telles Alphelys, Bacus Laboratories[3] ou d-metrix[4], proposent des systèmes d'acquisition d'images de lames histologiques, ou plus spécifiquement de lames TMA.

Ces images de lames complètes font alors souvent l'objet de traitement par divers algorithmes de segmentation ou de quantification de marquage. Parmi les travaux académiques, [Soenksen, 2003] s'intéresse à la mise au point d'un système de partitionnement automatique des images de lames en images individuels de spots. [Chen et al., 2004] proposent un prototype pour l'acquisition d'images, l'analyse et l'archivage de données TMA, basé sur des technologies Web. Ce prototype est focalisé sur le contrôle du système d'acquisition d'images de lames, l'évaluation automatique de biomarqueurs et la détection de spots sur les images de lames complètes, permettant ainsi le partitionnement automatique en images individuelles de spots. Dans la même veine, [Rabinovich et al., 2006] présentent un système incluant le prétraitement des images, en particulier pour reconstruire les spots abîmés, l'évaluation

[1] http ://www.beecherinstruments.com
[2] http ://www.alphelys.com
[3] http ://www.bacuslabs.com/
[4] http ://www.dmetrix.net/

automatique du marquage, le stockage et l'affichage des données et images. Des outils commerciaux, tels TMAx de Beecher Instruments, TMAScore de Bacus Laboratories ou TMALab d'Aperio[5], permettent tout à la fois la segmentation automatique des lames et l'évaluation du marquage.

Ces outils, s'ils fournissent une plus grande automatisation du processus et donc un gain de temps certain, permettent surtout un enrichissement des données associées aux lames TMA, apportant ainsi une part de complexité supplémentaire aux problèmes de stockage et d'exploration de données.

1.4.3 Gestion et exploitation des données

1.4.3.1 Gestion des données

Le volume de données générées par la technologie a fait rapidement apparaître un besoin majeur de stockage de données, thématique qui a été parmi les premières à émerger autour de l'informatisation de la technologie. Le contexte actuel de globalisation de l'information et de mutualisation des ressources, en particulier dans le domaine scientifique, a fait aussi émerger un besoin d'échange et de partage des données.

Cette informatisation du stockage et du partage des informations liées à la technologie a en particulier bénéficié d'un gros effort de normalisation, par la définition d'un format de fichiers d'échange de données TMA, le TMA Data Exchange Format [Berman et al., 2003], une DTD XML dont la structure sert de base à de nombreux outils de gestion de données TMA.

De nombreux systèmes se limitent d'ailleurs au stockage de données. Les plus basiques utilisent des fichiers Microsoft Excel [Shaknovich et al., 2003], associés à des outils de classification et de visualisation d'images pour l'outil de l'université de Stanford [Liu et al., 2002a], ou utilisent un système de gestion de base de données très simple tel que Microsoft Access [Manley et al., 2001]. Mais de telles représentations de l'information ne sont pas suffisantes dans le contexte TMA où l'espace des données est extrêmement complexe et hétérogène. La plupart des outils de stockage et visualisation de données TMA se basent sur de vrais systèmes de gestion de bases de données : Sybase pour TMAJ de l'université John Hopkins de Baltimore [Faith et al., 2004], MySQL pour TmaDB de l'université de Leeds [Sharma-Oates et al., 2005], etc. S'ils permettent la persistance de l'information, ces outils ne sont pas adaptés à l'exploration des données. Les plus avancés, comme TmaDB, permettent juste un filtrage simple, par le biais de requêtes SQL de type SELECT. Ceci implique tout à la fois une connaissance du langage SQL et du schéma

[5] http ://www.aperio.com/

de base de données, et une exploitation de listes de résultats en mode texte qui sont peu pertinentes dans un contexte où l'image est primordiale. De tels outils sont donc peu adaptés à leur public de biologistes et médecins.

D'autres systèmes se focalisent sur la notion d'échange et d'accès distribué aux ressources, essayant ainsi de pallier les problèmes de volume de données à conserver. Ce sujet est en effet particulièrement critique pour les images utilisées en anatomopathologie, dont la haute résolution est aussi gourmande en espace disque. Une solution couramment proposée est le recours à des architectures distribuées, où chaque site ne conserve localement qu'une partie des données. Par exemple, [Schmidt et al., 2004] présente un système basé sur des technologies pair-à-pair, permettant l'acquisition, le stockage, l'analyse collaboratifs de données TMA. Dans le même esprit, [Viti et al., 2007] s'appuie sur une infrastructure de type grille informatique ou «grid», l'EGEE (Enabling Grid for E-sciencE, projet qui fédère 90 institutions de 32 pays). Mais ce type d'initiative, tout en apportant une solution au problème d'espace de stockage, exacerbe encore le problème de l'exploitation des données, en mettant un volume d'informations encore plus grand à disposition des chercheurs.

Les simples stockages, accès et partage des données, s'ils fournissent un pool de données à exploiter, ne sont donc pas suffisants, et les systèmes les plus évolués visent à proposer en plus une visualisation plus poussée et des outils de fouille de données.

1.4.3.2 Exploitation des données

1.4.3.2.1 Visualisation des données

Alors que les systèmes simples de gestion de données TMA ne proposent qu'un accès par image de lame TMA ou image individuelle de spot, les plateformes les plus avancées dans l'accompagnement de la technologie proposent quelques concepts intéressants autour de la problématique de visualisation des données.

Ainsi, le système Profiler [Kim et al., 2005] offre un support complet à l'expérimentation, de la conception des plans de construction des blocs à l'évaluation du marquage. Il permet la mise en relation des informations TMA avec des données cliniques ou issues de bases de données moléculaires et le filtrage de jeux de spots basé sur des critères de type diagnostic ou immunohistochimiques, offrant ainsi une possibilité de navigation dans une partie précise de la collection d'images. TMABoost [Demichelis, 2005] propose un support similaire, incluant en plus un lien avec le système d'acquisition d'images. Enfin, Virtual Tissue Matrix [Conway et al., 2006], tout en permettant le stockage et l'évaluation de données issues d'études conduites avec la technologie, introduit la notion de lame TMA virtuelle, affichage au sein d'une

grille d'images de spots sélectionnées au sein de la base de données de l'application, permettant leur évaluation et analyse conjointe même si les spots sont issus de lames différentes. Ce concept de lame virtuelle se retrouve au sein de TMALab d'Aperio sous le nom d'«array composite», ainsi que dans la base de données d'Alphelys, qui permet la comparaison de spots lame à lame. Ces deux derniers systèmes commerciaux semblent être ceux qui proposent les capacités de sélection et filtrage des données les plus avancées.

Toutes ces plateformes, si elles proposent des ébauches dans la direction d'une navigation dirigée au sein de la collection de données restent encore plutôt rudimentaires sur ce point, qui n'est pas leur objet central.

1.4.3.2.2 Fouille de données

L'extraction d'informations pertinentes à partir des données générées par la technologie repose, pour l'ensemble des systèmes observés, sur une fouille de données. Deux tendances sont à noter en ce qui concerne la relation entre les outils associés à la technologie des TMA et la fouille de données : l'exportation sous un format qui peut servir d'entrée à un logiciel de fouille tierce partie ou l'intégration de tels outils à la plateforme de support à la technologie TMA.

La plupart des systèmes incluant une fonctionnalité de gestion de données se reposent sur la première option. Certains, comme celui de l'université de Stanford [Liu et al., 2002a], conservent leurs données directement sous un format Microsoft Excel qui est facilement importé par de nombreux systèmes de fouille de données ou logiciels de statistiques. D'autres proposent une exportation à divers formats (comme TMALab d'Aperio). Ce mode de fonctionnement permet une grande souplesse dans le choix du système de fouille de données et dans le type de traitements que le biologiste peut appliquer sur son jeu de données. Mais il implique une étape d'exportation/importation qui peut être source d'erreurs, peut nécessiter un nettoyage intermédiaire du fichier de données, et se révèle au final coûteux en temps.

Plusieurs logiciels incluent donc des outils de fouille de données. Certains se limitent à de simples analyses descriptives. Ainsi, la plateforme d'Alphelys propose des calculs de moyennes, écarts-types et variances. Le système le plus abouti dans l'analyse statistique des données TMA est sans doute TMA Foresight, de Premier Biosoft[6]. Cet outil permet tout à la fois la préparation des données, avec une assistance à la transformation de données qualitatives en données quantitatives et au remplacement de données manquantes, le calcul de statistiques descriptives (moyenne, écart-type, etc.), des régressions selon les modèles de Cox ou Kaplan-Meier.

D'autres plateformes incluent des outils d'analyse exploratoire des données. TMA-

[6] http ://www.premierbiosoft.com/

Boost [Demichelis, 2005], de même que TMA Foresight, proposent des outils de classification, permettant l'exploration de sous-ensembles de données spécifiques. L'outil de l'université de Stanford met lui aussi des outils de classification à disposition, en adaptant des systèmes d'analyse de puces à ADN aux TMA, et permet de dépasser le problème posé par les réplications de spots par l'application de règles de décision [Liu et al., 2005].

1.4.4 Apports et limites des outils associés aux TMA

1.4.4.1 Résolution des problèmes posés par la technologie TMA

La réalisation d'études utilisant la technologie des TMA pose un certain nombre de problèmes liés aux techniques de laboratoire ou aux données générées. Les divers outils qui viennent d'être présentés visent à répondre à une partie de ces interrogations.

En ce qui concerne les aspects techniques, les divers outils disponibles sur le marché permettent surtout de réduire le temps passé par les équipes techniques à la réalisation des blocs ou à l'acquisition des données, en offrant un support informatique automatisant un certain nombre d'étapes telles que la construction des blocs, l'acquisition et la segmentation d'images, la quantification de marquage.

Au niveau données, la plupart des systèmes offrent des fonctionnalités de gestion et partage des données plus ou moins évoluées et un accès à l'information selon différents points de vue, des possibilités de navigation dans les images de lames et spots... L'exploitation des données au sein de plateformes intégrées est plus rarement rencontrée et la plupart des outils reportent cette problématique sur des logiciels dédiés.

Le gros éventail de l'offre de systèmes informatiques associés à la technologie apporte donc des réponses à de nombreuses problématiques mais certains thèmes restent des questions ouvertes.

1.4.4.2 Des questions en suspens

Malgré les nombreux efforts qui sont faits dans l'informatisation de la technologie des TMA, quelques difficultés subsistent avec les outils existants. Il n'existe pas de plateforme intégrée regroupant l'ensemble des outils intéressants, les tâches demandant le plus d'expertise technique sont peu supportées, la fouille de données reste souvent au soin de logiciels externes et l'exploration de la collection de données, permettant son appréhension par l'utilisateur, est difficile.

Le premier constat général qui peut être tiré d'une revue des outils accompagnant la technologie est que, si le champ des problèmes résolus par l'ensemble de ces systèmes est important, aucune plateforme à ma connaissance ne propose une solution intégrée, qui couvre l'ensemble du processus. La mise en place d'une telle plateforme à partir d'outils existants implique le recours à plusieurs outils mis en série, avec les éventuels problèmes de compatibilité, perte de temps lors des transferts de données entre outils et autres, inhérents à ce type de situation.

Ensuite les étapes du processus qui reposent sur une expertise technique spécifique font peu l'objet d'une assistance informatisée. Ainsi, les algorithmes standards de traitement d'images tels que des algorithmes de segmentation ou de mesures colorimétriques peuvent être facilement adaptés au contexte TMA et sont présents dans plusieurs systèmes. Par contre, la conception des blocs, liée à des problématiques d'échantillonnage et des contingences techniques propres aux TMA, reste un processus manuel, même si des outils proposent des interfaces de définition des plans de fabrication des blocs. Or cette activité représente une charge cognitive et temporelle importante, et une assistance informatique serait la bienvenue.

De plus l'exploitation des données est majoritairement laissée de côté au sein des plateformes au bénéfice de logiciels externes au système.

Enfin, les outils informatiques dédiés sont assez limités en ce qui concerne l'exploration de la collection de données et images par l'utilisateur. Les analyses descriptives disponibles restent très basiques (moyenne, écart-type, variance, etc.). Les systèmes de navigation dans les images sont en général basés sur des observations lame par lame ou spot par spot. Les lames virtuelles, pour les systèmes qui incluent une notion de ce type, restent des artefacts construits manuellement. Des outils permettant la construction de vues organisées sur la collection permettraient une meilleure compréhension du jeu de données par l'utilisateur.

Il apparaît donc un certain nombre de manques au sein des systèmes associés à la technologie. En particulier, la navigation dans les données TMA dans un objectif exploratoire pour le biologiste nécessite une aide plus important que la conception manuelle de lames virtuelles ou le filtrage simple des spots. C'est l'objectif de ma thèse qui vise à proposer un système informatique d'assistance à l'évaluation des données TMA, apportant un support à une démarche expérimentale modifiée par une technologie à haut débit.

1.5 Constat et problème posé

Dans le contexte de la recherche en oncologie en général, et des études anatomopathologiques en particulier, la technique des Tissue MicroArrays apparaît comme

une solution intéressante aux problèmes posés par les lames histologiques classiques, aussi bien au niveau économie de tissus et réactifs qu'au niveau d'une réduction du temps de traitement des lames.

Cette technique expérimentale laisse pourtant ouvertes un certain nombre de questions. En particulier, le traitement en masse des échantillons, s'il apporte une dimension statistique au travail de l'anatomopathologiste et se prête bien à des méthodes de fouille de données de type exploratoire, pose des problèmes de gestion de l'information et d'appréhension du jeu de données par le biologiste. L'acquisition d'expertise technique complémentaire, l'informatisation du processus de construction de blocs et lames TMA et la mise en place d'outils informatiques de gestion et d'exploitation de données TMA semblent nécessaires pour dépasser les limites rencontrées dans le recours à la méthode.

L'offre de systèmes informatiques dédiés à la technologie des TMA est de plus en plus complète, mais il apparaît tout de même des manques au sein des outils associés à cette technique en ce qui concerne :

- ⋆ l'assistance à la conception raisonnée de blocs TMA,
- ⋆ l'assistance à la construction de lames TMA virtuelles, permettant au biologiste une exploration guidée de la collection de données et ainsi son appropriation de l'espace informationnel,
- ⋆ la mise en place d'outils de fouille de données spécifiques aux TMA, qu'ils s'agisse de méthodes descriptives, exploratoires, prédictives etc.
- ⋆ l'intégration de l'ensemble des outils nécessaires, de la conception des blocs à l'exploitation des données au sein d'une plateforme unique dédiée à la technologie.

La mise en place d'une plateforme intégrée dédiée à la technique des TMA est l'objectif de l'équipe RFMQ, avec le projet TMA-Explorer. Dans le cadre de ce projet, je m'intéresse aux problématiques d'exploration du jeu de données qui ont été collectées, qu'il s'agisse de dossiers cliniques ou de données anatomopathologiques (images et évaluation de lames et spots).

En effet, une vision sur ces informations, globale ou orientée selon une étude biologique particulière, est nécessaire au biologiste, tout à la fois pour évaluer les besoins en nouveaux blocs et pour avoir un aperçu des analyses qui pourront être réalisées sur les données, c'est-à-dire pour préparer la fouille de données ou replacer ses travaux dans le cadre d'une démarche expérimentale classique. Cet objectif peut être atteint par la construction automatique de lames TMA virtuelles, qu'elles présentent un plan de construction pour un bloc TMA ou une combinaison de spots existants, construction qui doit être guidée par les besoins analytiques de l'utilisateur. Or, les systèmes existants ne permettent qu'une construction manuelle de plans de fabrication ou de lames virtuelles, ce qui limite leur utilisation en matière d'exploration des données par le biologiste.

Dépasser les limites des outils actuels associés à la technologie en ce qui concerne l'exploration du jeu de données implique au niveau conceptuel, selon le point de vue adopté dans ma thèse, de :

- ⋆ fournir une assistance à la formulation d'hypothèses, exprimant ainsi un besoin décrivant une étude à réaliser,
- ⋆ constituer un jeu de données pertinent sur lequel une hypothèse donnée pourra être testée, sous une forme permettant une première évaluation rapide de la validité des hypothèses,
- ⋆ prendre en compte les connaissances existantes dans le domaine TMA, que ce soit au niveau des éléments manipulés ou au niveau méthodologie expérimentale,
- ⋆ prendre en compte des contraintes qualité.

En fait ce problème, vues ses caractéristiques, peut être considéré comme un problème de synthèse. La synthèse est en effet décrite comme une opération intellectuelle qui consiste à réunir méthodiquement les divers éléments d'un ensemble, ce qui correspond au processus décrit précédemment. Ce processus est de plus courant dans nombre de disciplines où intervient une phase de confrontation de grandes quantités de données afin d'en tirer une vue structurée. On peut penser à l'archéologie (exploration virtuelle d'un site de fouilles), bibliométrie (relations entre auteurs), pédologie (pédocomparateur), botanique (navigation dirigée dans un herbier informatisé)...

Considérer le problème d'appropriation par un scientifique d'une importante collection de données hétérogènes, telles les données TMA, comme un problème de synthèse, implique dans un premier temps de replacer cette notion dans un contexte plus large. En effet, ce concept peut être rattaché à de nombreux domaines de recherche : fouille de données, Visualisation d'Information, Recherche d'Information, systèmes adaptatifs, raisonnement artificiel, représentation de connaissances etc. Cet état de l'art, objet du chapitre 2, m'a conduit à adopter un point de vue particulier sur la synthèse, qui considère ce problème comme un problème de Recherche d'Information augmenté de paradigmes issus d'autres domaines. Cette proposition, introduite dans le chapitre 3 a conduit à l'implémentation d'un prototype présenté chapitre 4, dont une première validation expérimentale dans le domaine des TMA est décrite dans le chapitre 5.

CHAPITRE

2 La synthèse d'information : un problème difficile aux multiples facettes

Le chapitre précédent a permis l'identification d'un problème d'appréhension des données TMA par le biologiste utilisant cette technologie. Or, cette appréhension des données est nécessaire tout à la fois pour planifier une fouille de données et surtout pour replacer les travaux dans le contexte d'une démarche expérimentale classique. Ces difficultés ont conduit à l'émergence d'un besoin d'exploration des informations collectées. Une solution possible à ce problème d'analyse descriptive des données est la construction assistée par ordinateur de documents présentant des vues de synthèse sur les données. Dans ce chapitre est abordée cette notion de synthèse, concept général mais difficile à appréhender. La vision de la synthèse choisie peut se rattacher à de nombreux domaines de recherche tels que la fouille de données, la Visualisation d'Information, la Recherche d'Information. Des systèmes d'assistance à la synthèse pourraient aussi tirer parti de concepts issus des domaines du Web sémantique et hypermédia adaptatif, de l'Intelligence Artificielle. La synthèse est donc une notion aux multiples facettes et ce chapitre propose une exploration rapide de l'état de l'art de ces diverses disciplines et envisage leur rôle potentiel dans la synthèse.

2.1 Introduction

La technique des TMA, en permettant le traitement en masse d'échantillons tissulaires, place le biologiste face à un gros volume d'informations que les outils dédiés à la technologie ne lui permettent pas d'exploiter facilement. En particulier, l'appréhension de la collection de données acquises, données qui sont de plus hétérogènes, est difficile. Or, une vision et compréhension globales des informations disponibles sont importantes, en tant qu'étape préliminaire à une fouille de données et surtout pour envisager l'utilisation des données dans une démarche expérimentale standard. Proposer au biologiste un système informatique construisant des synthèses adaptées à ses besoins est la solution envisagée dans le cadre de ma thèse. Le développement de ce point de vue implique de définir plus en détails cette notion de synthèse telle qu'elle est considérée dans le cadre de mes travaux et de mettre en relation ce concept avec les domaines de recherche qui lui sont apparentés ou dont des paradigmes peuvent être utilisés dans la réalisation de documents de synthèse.

Le concept de synthèse est une notion ancienne qui trouve ses racines dans la didactique et la rhétorique, en tant que point de conjonction entre thèse et antithèse et méthode de raisonnement qui, par opposition à l'analyse, va des principes aux conséquences, des causes aux effets, du simple au composé, de l'élément au tout. Il s'agit donc d'une notion difficile à appréhender.

Dans le milieu scientifique, la synthèse apparaît surtout comme une opération intellectuelle qui consiste à réunir méthodiquement les divers éléments d'un ensemble et elle émerge du besoin croissant de compacter une surabondance d'informations, par exemple au sein d'articles de type «review » ou de synthèses bibliographiques. Pour Goldschmidt, qui propose un guide pratique de rédaction pour une synthèse scientifique [Goldschmidt, 1986], elle résulte de la collecte systématique de résultats de recherche sur un sujet déterminé, pour un public particulier, dans un but précis ; de l'évaluation de la qualité des résultats utilisés ; de la présentation des résultats valides sous une forme utile pour le public visé ; de l'inclusion d'une discussion des manques informationnels majeurs qui devraient faire l'étude de recherches futures. Les chercheurs engagés dans une activité de synthèse ont donc pour objectif de construire une vision personnalisée du monde, compacte et argumentée, basée sur l'agrégation d'un ensemble potentiellement hétérogène, dirigée par un besoin informationnel précis. Il s'agit donc d'un concept aux multiples facettes, qui pose tout à la fois des problèmes d'appréhension des données, d'exploration des données et de représentation du problème de synthèse et des entités impliquées.

Tout d'abord, cette notion de synthèse pose un problème d'appréhension des données, au sens d'une vision globale sur l'information. Ainsi, en tant que construction d'une vue compacte sur un jeu important de données, la synthèse peut se rattacher au domaine de la fouille de données, qui vise à valoriser des gros volumes de données en en extrayant des connaissances par des méthodes mathématiques ou informa-

tiques. Elle peut aussi se rattacher à la Visualisation d'Information, qui permet la construction de vues générales sur les données, une vision englobante sur l'information qui tire parti des possibilités graphiques de plus en plus avancées des systèmes informatiques.

Ensuite, la synthèse pose un problème d'exploration des données, au sens d'une orientation de la vue proposée sur l'information en fonction d'un but ou d'un besoin. Ainsi, en tant de problème de manipulation d'information, la synthèse peut être considérée comme un objet d'étude pour la communauté des sciences de l'information. Celle-ci pose alors les trois problèmes classiques qui ont fait des sciences de l'information une discipline en tant que telle : la problématique de Recherche d'Information, la question de la pertinence des informations trouvées et le problème de l'interaction entre le système et l'usager.

Enfin, la synthèse pose des difficultés de représentation du problème de synthèse en lui-même et des entités impliquées. Ainsi, en tant que construction destinée à un public particulier, la synthèse peut être considérée comme relevant du domaine des systèmes adaptatifs, tirant parti d'une représentation de l'utilisateur pour proposer des services personnalisés. Conjointement, en tant que construction argumentée d'un objet composite complexe, la synthèse est une tâche difficile dont la résolution peut tirer parti de paradigmes d'Intelligence Artificielle, discipline qui s'oriente entre autres vers l'assistance à l'usager dans ses tâches quotidiennes répétitives ou à la charge cognitive importante, et qui tire de plus en plus parti de représentations des problèmes à résoudre et de connaissances.

Afin d'aider le scientifique engagé dans des activités de synthèse, il s'agit de proposer un outil tirant parti de concepts issus de ces divers communautés. Ceci implique une exploration rapide de l'état de l'art de chacune d'elles pour en extraire les éléments pertinents pour la notion de synthèse et le positionnement du point de vue adopté sur la synthèse par rapport à ces divers domaines de recherche.

2.2 Appréhender les données

2.2.1 Introduction

Construction d'une vue compacte et globale sur les informations disponibles, la synthèse pose tout d'abord un problème d'appréhension des données, qui peut être abordé tout à la fois d'un point de vue fouille de données et d'un point de vue Visualisation d'Information.

Ce problème relève de l'un des thèmes de la fouille de données. Cette discipline vise en effet à proposer des outils permettant tout à la fois l'appréhension globale

d'un jeu de données (analyse descriptive), l'exploration de son organisation (analyse structurelle), et la construction de modèles prédictifs (analyse explicative). La synthèse peut donc être considérée comme une méthode d'analyse descriptive, champ où les supports visuels ont souvent une grande importance.

Les limites des méthodes d'analyse visuelle des données et l'évolution des moyens informatiques et en particulier graphiques ont permis l'émergence de la Visualisation d'Information, qui complète les méthodes de fouille visuelle. Cette discipline semble alors d'un intérêt certain pour la synthèse.

Dans la suite de cette section, ces deux thèmes de fouille de données et de Visualisation d'Information ainsi que leur relation à la notion de synthèse vont être abordés.

2.2.2 Fouille de données

2.2.2.1 Introduction

La problématique d'appréhension des données, qui est sous-jacente au concept de synthèse, suggère de s'intéresser au domaine de la fouille de données. En effet, le data mining, ou fouille de données en français, est apparu au début des années 90, pour aider à l'exploitation des données de plus en plus nombreuses stockées par des systèmes informatiques aux capacités de stockage et de traitement de plus en plus importantes. La fouille de données s'est alors structurée en tant que maillon essentiel de la chaîne de traitement de l'extraction de connaissances à partir de données, tel qu'exposé dans [Frawley et al., 1992].

Ce dernier processus d'extraction de connaissances inclut tout d'abord des prétraitements : construction de corpus de données spécifiques ou datamarts, mise en forme des données, nettoyage des données (traitement des données manquantes...), etc. La fouille de données opère alors sur ces datamarts, par le biais de méthodes issues de domaines variés : statistiques, analyse de données, reconnaissance des formes, apprentissage automatique... Ces diverses méthodes, dont un aperçu est présenté dans [Fayyad et al., 1996], apportent alors des outils d'analyse descriptive, structurelle ou explicative, dans un objectif final de découverte de sens.

De nos jours, les logiciels dédiés à la fouille de données, aux interfaces attractives, proposant une automatisation poussée des traitements, sont très nombreux : [Goebel and Gruenwald, 1999] le notent dès 1999, en comparant une soixantaine d'outils, et des sites Web sur le thème du data mining comme kdnuggets [1] en répertorient tout autant. Mais le mythe d'une fouille de données presse-bouton perme-

[1] http ://www.kdnuggets.com

ttant l'extraction d'informations pertinentes sans connaissance, ni des algorithmes sous-jacents aux outils, ni des données à analyser, s'écroule de plus en plus. Dans cet esprit, [Friedman, 1997] suggère que l'engouement pour la fouille de données pourrait s'avérer au final plus profitable pour les vendeurs d'outils que pour leurs utilisateurs. Aussi, la fouille de données est de plus en plus perçue comme un processus interactif et itératif, dirigé vers un but particulier, résultant de la coopération d'un expert du domaine d'étude, capable de faire la distinction entre information utile et information sans intérêt, et d'un expert en fouille de données qui est à même d'évaluer les outils applicables aux cas analysés.

La question qui doit être résolue pour mener une fouille de données avec succès apparaît donc en définitive celle de l'adéquation entre le datamart construit pour l'étude et les algorithmes appliqués, la sélection et nettoyage des données et le choix des outils étant deux activités intimement liées. Dans ce contexte, des outils permettant l'appréhension des données, prennent tout leur sens. C'est l'un des objectifs de la synthèse, telle qu'elle est envisagée dans le cadre de mes travaux, ainsi que d'un champ particulier de la fouille de données : l'analyse descriptive de données. Cette dernière fait souvent appel à des représentations graphiques, par l'analyse visuelle des données. La notion de synthèse va donc être replacée dans le cadre de ces deux sous-domaines de la fouille de données dans les prochains paragraphes.

2.2.2.2 Analyse descriptive des données

Les méthodes de fouille de données descriptives visent à fournir à leur utilisateur un aperçu sur les données, tout comme dans le point de vue adopté ici sur la synthèse. En effet, une étude sophistiquée d'un ensemble de données est très souvent précédée d'une étude exploratoire, souvent à l'aide d'outils simples mais robustes, dont l'objectif, ainsi qu'il est exposé dans [Besse et al., 2001], est d'éviter de tomber dans des pièges grossiers liés à une mauvaise appréciation du contexte d'utilisation des outils de fouille.

Cette observation de la collection de données permet de se familiariser avec les données et de détecter des problèmes éventuels tels que valeurs manquantes, erronées ou atypiques, modalités trop rares, distributions «anormales », incohérences, liaisons non linéaires... Elle permet aussi de guider le choix de pré-traitements des données qui les rendront conformes aux méthodes de modélisation ou d'apprentissage qu'il faudra mettre en œuvre pour atteindre les objectifs fixés. Dans le cadre de fouille de données scientifiques, une telle étude permet enfin d'aider le chercheur à replacer ses données dans une démarche expérimentale classique, en lui fournissant une base sur laquelle émettre des hypothèses.

Parmi les méthodes relevant de l'analyse descriptive des données, on peut citer des méthodes statistiques simples, telles que des calculs de moyennes ou écarts-types, etc.

ainsi que des techniques d'analyse de données, telles que l'Analyse en Composantes Principales, qui permettent d'évaluer des tendances ou des dispersions. Parmi ces techniques, beaucoup se basent sur une présentation graphique, thème exploré plus précisément dans le prochain paragraphe.

2.2.2.3 Analyse visuelle des données

L'analyse descriptive des données recourt beaucoup à des représentations graphiques, qui sont souvent plus expressives que des ensembles numériques ou des tableaux de chiffres. Dans cette approche, l'analyse descriptive des données suit des résultats de sciences cognitives qui tendent à montrer que l'humain se représente mieux le monde sous forme schématique que sous forme discursive [Larkin and Simon, 1987]. Ainsi [Friendly and Kwan, 2003] insistent sur l'importance du choix de présentation des informations en statistiques et fouille de données, une représentation organisée d'une collection pouvant être porteuse d'informations complémentaires par rapport à un ensemble de documents ou de données, alors qu'une présentation inadaptée peut conduire à de fausses interprétations ou à une non détection de faits importants.

Dans le contexte des statistiques, ces représentations prennent la forme de camemberts, histogrammes, nuages de points, diagrammes en étoile, boîtes à moustaches... Pour la plupart conçues à une époque où l'outil informatique était inexistant ou à ses premiers balbutiements, elles présentent l'indéniable avantage d'une facilité de réalisation, même de façon manuelle, tout en permettant une certaine appropriation du jeu de données par le chercheur, en fournissant divers points de vues sur l'information.

Mais, les méthodes graphiques simples sont parfois inadaptées quand les informations à représenter sont complexes, ce qui est souvent le cas pour les données expérimentales. Cette problématique a attiré l'attention tout à la fois des statisticiens et des informaticiens, conduisant à une évolution du domaine. Des représentations permettant de gérer ces informations complexes ont donc vu le jour. Par exemple, [Noirhomme-Fraiture and Rouard, 1997] introduisent le concept de zoom stars, et en particulier les zoom stars 3D, qui combinent diagrammes en étoiles et histogrammes pour représenter des objets symboliques (classes d'individus représentés par un jeu de variables quantitatives, qualitatives, intervalles ou à valeurs multiples...) et permettre des comparaisons entre objets et des évaluations de corrélations entre variables.

De plus, l'augmentation de puissance des ordinateurs, et en particulier les capacités graphiques des postes de bureau, ont conduit à l'émergence d'outils tirant parti des performance des systèmes informatiques courants ou expérimentaux, à la marge entre les domaines de la fouille de données et de l'interaction homme/machine : la Visualisation d'Information. La synthèse, qui manipule des données complexes,

pourrait tirer parti de paradigmes issus de cette discipline récente, qui fait l'objet du paragraphe suivant.

2.2.3 Visualisation d'information

2.2.3.1 Introduction

Historiquement, la Visualisation d'Information est née dans les années 90 d'une convergence entre des thèmes de recherche issus de domaines plus généraux : visualisation scientifique, Recherche d'Information, interaction homme/machine ou hypermédia... Elle vise à faciliter l'analyse, l'interprétation et la supervision de phénomènes complexes, en facilitant l'exploration d'informations disponibles en très grand nombre, par le biais d'images digitales interactives ou animées. La visualisation, dans le sens de présentation, n'est pas un concept récent et est utilisée depuis des milliers d'années, en particulier dans les cartes géographiques. Le domaine de la fouille de données, et en particulier les statistiques recourent aussi souvent à des présentations visuelles sous formes de graphiques et autres.

L'ajout de l'aspect informatique à la notion de visualisation ne se réduit pourtant pas à la réalisation de ces graphiques simples avec l'aide d'un ordinateur. Les capacités graphiques des postes de travail actuels permettent en effet la construction de métaphores graphiques élaborées, qui dépassent les représentations classiques sur papier et même les interfaces classiques dites de type WIMP (Window, Icon, Menu, Pointer), avec des concepts de type «focus+context» [Cohen and Brodlie, 2004], les interfaces déformables, les interfaces zoomables à l'infini, etc.

La problématique de présentation structurelle reflétant une organisation conceptuelle, sous-jacente à la notion de synthèse, est alors celle de la Visualisation d'Information, telle qu'exposée plus en détails dans le prochain paragraphe.

2.2.3.2 Un domaine en pleine expansion

Comme le souligne [Eick, 2005], les débuts de la Visualisation d'Information ont vu l'émergence d'une pléthore de méthodes de présentation. Ainsi, [Kroeker, 2004] introduit quelques systèmes de visualisation basés sur des métaphores visuelles différentes : graphes, chronologies, arbres, tables, cartes, etc. Quelles que soient les méthodes utilisées, quelques questions basiques valent la peine d'être explorées : y a-t-il vraiment un apport notable pour l'utilisateur entre les représentations graphiques et le mode textuel ? quelles représentations utiliser ? comment tirer parti au mieux de l'espace disponible ?

La première question de l'apport effectif des visualisations 2D ou 3D par rapport à une présentation textuelle de l'information a été évaluée par [Sebrechts et al., 1999], [Cugini et al., 2000] ou [Heidorn and Cui, 2000] dans le contexte de la visualisation des résultats d'un système de Recherche d'Information et fait apparaître un coût cognitif croissant avec la dimensionnalité de la représentation, coût qui diminue avec l'expérience et semble dépendre de la tâche à réaliser (entre autres localisation, comparaison, ou groupement de documents). Les autres travaux concernent surtout des comparaisons entre 2D et 3D, en relation avec la mémoire spatiale des sujets, celle-ci étant en effet particulièrement impliquée dans les activités de visualisation. Ainsi, [Tavanti and Lind, 2001] montrent que les interfaces 3D améliorent les performances dans la tâche impliquant la mémoire visuelle qu'ils ont testée (retrouver la position d'un objet, ici une lettre, dans une hiérarchie). Mais, les contrôles des interfaces 2D et 3D différant, ceux-ci peuvent constituer un biais, dont [Cockburn and McKenzie, 2002] cherchent à s'affranchir en testant tout à la fois des interfaces et leur équivalent dans le monde réel. Les résultats sont alors plus mitigés et les utilisateurs semblent trouver que plus la dimensionnalité de l'interface augmente, plus l'affichage devient chargé et moins il est efficace. Il apparaît alors nécessaire de réaliser un compromis entre l'intérêt pour l'utilisateur et les coûts de calcul de la présentation, qui augmentent avec le nombre de dimensions. Ce compromis semble, d'après les travaux analysés ici, en faveur du 2D.

Se pose alors la question du type de représentation à utiliser. Le volume d'éléments à présenter au sein d'une structure géométrique simple reste une problématique majeure, même si elle a fait l'objet de nombreuses études. [Card and Mackinlay, 1997] et [Keim, 2002] proposent d'ailleurs des taxonomies des diverses méthodes proposées. Or une communauté de chercheurs s'intéresse depuis toujours à ces problèmes : ce sont les géographes et en particulier les cartographes qui, au sein des systèmes d'information géographique, cherchent à présenter de gros volumes de données pertinentes dans un espace à deux dimensions. De nombreux concepts de cartographie peuvent être appliqués à d'autres domaines [Skupin and Fabrikant, 2003], et la symbolique spatiale est souvent utilisée au sein des systèmes de Visualisation d'Information ou de fouille visuelle de données, afin de permettre l'appréhension de concepts abstraits multidimensionnels. Cependant, en Visualisation d'Information, le système de coordonnées est souvent arbitraire en pratique, et ne consiste finalement qu'en une métaphore graphique pour un algorithme de classification sous-jacent.

Enfin, dans un contexte où l'espace d'affichage est limité, l'objectif devient plutôt une représentation compacte de l'information, pour laquelle plusieurs approches sont possibles. Par exemple les cartes de Kohonen [Kohonen et al., 1996] utilisent un modèle de topologie auto-adaptative basé sur un réseau neuronal. Dans l'approche Treemaps [Shneiderman, 1992], une hiérarchie est représentée non pas sous forme d'arbre, mais comme un imbriquement de zones rectangulaires. [Wattenberg, 2005] propose de dépasser le modèle Treemaps, pour lequel aucun algorithme idéal n'a été trouvé, et explore des représentations non rectangulaires.

2.2.4 Des limites à l'appréhension de données

La synthèse a été abordée ici comme un outil d'analyse visuelle des données, mais ces outils présentent des limites que la Visualisation d'Information, qui tire parti des capacités des ordinateurs pour proposer des constructions graphiques interactives élaborées, vise à dépasser. Ces diverses méthodes permettent une appréhension globale des données ou informations disponibles, mais présentent quelques lacunes dans le contexte de la synthèse.

Tout d'abord, l'organisation des éléments au sein de la structure d'affichage reste centrée sur des attributs explicitement présents dans les données ou les documents, ce qui n'est pas forcément adapté dans le cas d'informations complexes et potentiellement implicites.

De plus, la représentation est en général figée. Les outils permettant un focus sur une partie des données ou une réorganisation consistent en général en traitements a posteriori, tels que des filtres. Or, la synthèse implique la construction de vues orientées a priori dans un but précis. La synthèse induit alors une problématique de sélection tout à la fois d'informations pertinentes, mais aussi d'outils permettant la construction d'une vue adéquate en fonction de l'objectif recherché.

Enfin, les phases de pré-traitement (sélection, nettoyage, etc.) sont souvent laissées à la discrétion du chercheur. En conséquence, ces méthodes apportent une assistance limitée en ce qui concerne la problématique de replacement des données dans une démarche expérimentale standard : par exemple, la réalisation de la sélection des informations pertinentes pour tester une hypothèse est généralement une étape manuelle.

Ces problématiques relèvent de l'exploration des données et passent par des manipulations d'informations. En particulier, la notion de sélection d'informations pertinentes pour tester une hypothèse évoque le domaine de la Recherche d'Information. Ceci suggère un intérêt pour les concepts de la discipline des sciences de l'information et en particulier de la Recherche d'Information, pour la synthèse. L'analyse de cette relation fait l'objet de la prochaine section.

2.3 Explorer les informations

2.3.1 Introduction

La section précédente a permis de cerner les problématiques d'appréhension des données sous-jacentes à la notion de synthèse, problématiques qui peuvent trouver

une solution dans les méthodes de fouille de données et de Visualisation d'Information. Mais cette analyse a conduit à quelques réserves qui suggèrent de considérer le domaine des sciences de l'information.

Les sciences de l'information ont vu le jour à la fin de la seconde guerre mondiale, en réponse au problème d'explosion de l'information pointé par [Bush, 1945], qui propose aussi une solution technologique à cette question de stockage et d'accès à l'information scientifique, avec «memex», l'intuition de ce qui deviendra les systèmes d'information informatisés d'aujourd'hui. Il s'agit d'un domaine pluridisciplinaire qui se focalise principalement sur la collecte, la classification, la manipulation, le stockage, la recherche et la dissémination d'informations. Les sciences de l'information ne se limitent donc pas à des problématiques informatiques et incorporent des aspects de bibliothéconomie, sciences cognitives ou sciences sociales. La synthèse, qui manipule de gros volumes d'informations hétérogènes, semble donc bien cadrer dans ce domaine.

D'après [Saracevic, 1999], l'histoire des sciences de l'information est très fortement liée à trois grandes idées qui ont été développées au cours du temps. La première notion originale qui a émergé dans les années 50 est la notion de Recherche d'Information, qui permet un traitement de l'information basé sur la logique formelle. La seconde notion, qui est apparue peu après, est la notion de pertinence, orientant et associant directement le mécanisme avec les besoins et évaluations humains. Enfin, la dernière idée forte est la notion d'interaction qui permet des échanges et retours directs entre systèmes et personnes engagées dans un processus de recherche. Ces trois notions de Recherche d'Information, pertinence et interaction, ainsi que leurs relations à la synthèse, vont être explorées dans la suite de cette section.

2.3.2 Recherche d'Information

2.3.2.1 Introduction

Les recherches autour de la notion de Recherche d'Information, en tant que processus facilitant l'accès aux informations pertinentes dans le cadre d'une tâche que l'utilisateur doit accomplir, se sont rapidement scindées en deux groupes distincts [Saracevic, 1999] :

- ⋆ une vision comportementaliste : ces travaux visent à modéliser les manifestations de l'information telles que les documents et leur usage. Ils s'intéressent au comportement des utilisateurs en cours de recherche d'information, au processus de Recherche d'Information en tant que communication dans un contexte social,
- ⋆ une vision opérationnelle : ces études se consacrent à la théorie et aux algorithmes de Recherche d'Information, à la construction de systèmes informa-

tisés efficaces.

Ces deux visions vont faire l'objet des prochains paragraphes.

2.3.2.2 Une vision comportementaliste

Selon la vision comportementaliste, il s'agit de mieux comprendre le processus de Recherche d'Information en analysant les entités impliquées et les relations entre elles à un haut niveau d'abstraction. Selon ce point de vue, l'utilisateur est considéré comme entité pensante dans un contexte qui peut être socio-économique, ou affectif, décrit dans [Kuhlthau, 1991] ; l'usager est confronté au problème cognitif posé par une représentation incomplète du monde qu'il cherche à dépasser, comme évoqué par exemple dans les «Anomalous States of Knowledge» de [Belkin et al., 1982]. Le processus de Recherche d'Information vise dès lors à étendre les connaissances de cet utilisateur en facilitant son accès à l'information.

Ces travaux peuvent rester très théoriques mais la plupart se basent sur des études utilisateurs. Ces études peuvent concerner des groupes d'utilisateurs particuliers, comme par exemple les chercheurs en sciences sociales dans [Meho and Tibbo, 2003] ou dans la série d'articles introduite dans [Spink et al., 2002], les astronomes, chimistes, mathématiciens et physiciens dans [Brown, 1999]. Elles peuvent aussi se focaliser sur un contexte de recherche spécifique, comme par exemple l'Internet pour [Johnson et al., 2003], qui tentent de proposer un environnement d'évaluation adapté.

Ces multiples études ont conduit à un ensemble de modèles. Par exemple, Dervin [Dervin, 1983] conçoit l'information comme un outil humain conçu pour faire sens d'une réalité perçue tout à la fois comme chaotique et ordonnée ; Ellis [Ellis, 1989] dote les comportements de recherche observés empiriquement d'un ensemble de caractéristiques dont les interactions dépendent de l'activité de recherche en cours à un instant donné ; Kuhlthau [Kuhlthau, 1991] complémente ce dernier modèle en attachant aux stades du processus de recherche des sentiments, pensées, actions. Dans [Wilson, 1999], Wilson compare ces diverses conceptions avec son propre modèle, qui considère l'émergence du comportement de recherche comme conséquence de la prise de conscience d'un besoin, besoin qui sera satisfait ou non à la fin du processus. Ces modèles visent à refléter un mécanisme générique de Recherche d'Information, ou se focalisent sur un contexte particulier, comme la recherche par l'intermédiaire d'un bibliothécaire pour [Kuhlthau, 1991], ou sur l'Internet dans [Choo et al., 1999].

Ces approches permettent une meilleure appréhension du fonctionnement de l'interaction entre l'utilisateur et les sources d'informations. Cette problématique est particulièrement intéressante dans l'objectif de la construction d'un système informatisé d'assistance à la synthèse d'informations, puisqu'elle permet d'asseoir un cadre théorique sur lequel un modèle de synthèse pourrait être défini. Mais les mod-

èles construits dans ce contexte comportementaliste ne peuvent conduire directement à des implantations de systèmes de Recherche d'Information, étant donnée la grande part de subjectivité impliquée. Cet aspect système est par contre au centre des préoccupations de la vision opérationnelle de la Recherche d'Information.

2.3.2.3 Une vision opérationnelle

L'opérationnalisation du processus de recherche par la construction de systèmes informatiques dédiés peut servir de base pratique à la mise en place de systèmes de synthèse. Cette opérationnalisation est tout d'abord passée par des implantations basées sur un modèle très simple et mécanique visant à réaliser une correspondance entre une requête constituée de mots clés et des représentations de documents sous forme de vecteurs de mots, ces représentations étant construites lors d'une phase préalable d'indexation.

Mais ce modèle simpliste n'est plus suffisant lorsqu'il s'agit de répondre aux besoins d'utilisateurs réels dans un environnement non contrôlé. La représentation d'un corpus documentaire sous forme de vecteurs de mots ne permet en effet pas de refléter la complexité inhérente à l'information, en particulier de nos jours où la place du multimédia, images, sons et vidéo, ne permet plus de se limiter aux seules informations textuelles. De plus, le fossé cognitif et sémantique entre le besoin réel de l'utilisateur et la représentation de ce besoin sous forme de liste de mots clés est important.

L'amélioration des résultats des systèmes de Recherche d'Information passe tout d'abord par l'évaluation de ces résultats, permettant ainsi une comparaison. Elle conduit ensuite à l'évolution du système de Recherche d'Information, modification dont il est espéré qu'elle induira une amélioration du système. L'évaluation des résultats est la problématique qui a donné naissance à la notion de pertinence. De façon pratique, la modification des systèmes passe souvent par l'ajout de nouvelles dimensions au modèle sous-jacent de correspondance entre requête et documents. Ces thèmes de pertinence et extensions du modèle simple de Recherche d'Information sont exposés entre autres dans le classique ouvrage «Modern Information Retrieval» de [Baeza-Yates and Ribeiro-Neto, 1999] et sont abordés dans le paragraphe suivant.

2.3.3 Pertinence : mesure et stratégies d'amélioration

2.3.3.1 Introduction

La notion de pertinence a été introduite comme réponse au problème d'évaluation des résultats des systèmes de Recherche d'Information. C'est une problématique

centrale pour nombre de travaux en Recherche d'Information, ainsi que le montre Mizarro dans son historique de la thématique [Mizzaro, 1997]. Elle reste aussi d'actualité dans le cadre du problème de synthèse, une évaluation objective ou subjective de la qualité du résultat produit par le système étant un élément standard de la qualité logicielle.

Cette notion de pertinence pose tout d'abord le problème conjoint de sa définition et de sa mesure, préalable à la mise en place de stratégies d'amélioration des performances des systèmes de Recherche d'Information.

2.3.3.2 Une pertinence multidimensionnelle difficilement mesurable

Dans le contexte d'évaluation des systèmes de Recherche d'Information, les mesures qui ont rapidement émergé et restent parmi les plus courantes sont la précision (nombre de documents jugés pertinents parmi ceux retournés par le système) et le retour («recall» en anglais, nombre de documents retournés parmi le nombre total de documents disponibles). Il s'agit de mesures simples et très mathématiques, qui présentent l'avantage de permettre des évaluations et comparaisons entre systèmes au sein de campagnes telles que TREC[2], mais qui posent pourtant un double problème, de calcul et d'adéquation au problème d'évaluation.

Tout d'abord, même ces métriques simples, qui sont bien acceptées et utilisées au quotidien dans la communauté, sont difficiles à mesurer dans les corpus documentaires actuels. Il est en effet, même dans des collections de test, difficile d'envisager une revue de tous les documents, à la lumière de toutes les requêtes possibles, par des juges humains. Ainsi, pour TREC, les jugements de pertinence de référence sont en partie ceux des divers systèmes en compétition.

Indépendamment du problème d'évaluation de la mesure, se pose la question de la nature de la mesure, et plus particulièrement de ce qu'il faudrait mesurer. Les mesures de type précision et retour sont très proches du système, pertinence algorithmique qui n'est pas forcément suffisante alors que l'objectif de la Recherche d'Information est de répondre à un besoin informationnel de l'utilisateur. Ce constat a conduit à l'émergence de diverses notions de pertinence, tendant vers la prise en compte de valeurs subjectives, toutes aussi valides les unes que les autres selon le point de vue adopté [Saracevic, 1975]. Ainsi [Mizzaro, 1998] considère les diverses notions de pertinence comme des relations entre deux dimensions à plusieurs niveaux d'abstraction : le besoin (besoin réel, besoin perçu, requête intentionnelle et requête formulée) et l'information (information, documents et représentation de documents). [Borlund, 2003] présente une analyse similaire des différents degrés et niveaux de pertinence et insiste sur l'aspect dynamique de la notion, celle-ci évoluant en cours de processus de recherche en conjonction avec une évolution des besoins utilisateur.

[2]http ://trec.nist.gov/

La notion de pertinence apparaît alors comme une notion floue et multifacettes qui est mal reflétée par les mesures simples communément admises. Cette multidimensionnalité présente aussi l'avantage de permettre d'envisager des améliorations des performances des systèmes de Recherche d'Information selon divers axes, correspondant à des dimensions différentes de la pertinence considérée, thème évoqué dans le prochain paragraphe.

2.3.3.3 Multiples stratégies d'amélioration des performances

Les diverses notions de pertinence se sont complexifiées en prenant en compte des dimensions de haut niveau d'abstraction telles que le besoin utilisateur réel ou perçu ou la notion d'information par opposition aux documents qui sont censés la contenir. De la même façon, l'amélioration des performances des systèmes de Recherche d'Information passe généralement par une complexification des représentations des éléments intervenant dans le système.

Classiquement, les systèmes de Recherche d'Information font intervenir une fonction de correspondance entre une représentation du corpus documentaire et une représentation du besoin sous forme de requête. Les diverses stratégies d'évolution des logiciels de recherche en général cherchent à aller vers une représentation de plus haut niveau d'abstraction selon une ou plusieurs de ces trois dimensions.

Ainsi, pour les algorithmes de correspondance, le modèle booléen où chaque document contient ou non des termes de la requête, et est ainsi pertinent ou non, est dépassé par des algorithmes aux résultats plus nuancés. Ces algorithmes permettent une évaluation plus fine d'un degré de pertinence, autorisant alors un ordonnancement des documents par une valeur de pertinence. Dans ce cadre, on peut citer entre autres les modèles vectoriels ou probabilistes.

En ce qui concerne les dimensions corpus documentaire et utilisateur, a été introduite la notion de Recherche d'Information en Contexte qui regroupe l'ensemble des paradigmes visant à une meilleure prise en compte tout à la fois des documents, de leur contenu et en particulier leur sémantique, des besoins de l'utilisateur et leur expression, de l'usager, sa connaissance du monde et ses préférences, autant de dimensions qui sont introduites dans [Cool and Spink, 2001], par exemple, et explorées dans les divers articles de [Ingwersen et al., 2005].

Plus précisément, est introduite une notion d'utilisateur en tant que tel, qui n'est plus réduit à l'expression de ses besoins et ne se limite plus à une liste de mots clés. L'utilisateur devient une entité individuelle qui interagit dans un processus cyclique avec le système. Cette notion d'interaction, devenue centrale dans les systèmes de Recherche d'Information, est l'objet du prochain paragraphe.

On peut aussi noter la prise en compte de niveaux d'abstraction plus élevés que les simples vecteurs de mots, par l'enrichissement conjoint de la représentation du corpus documentaire et de la requête. Cet enrichissement sémantique du corpus documentaire, qui passe souvent par la prise en compte de représentation de connaissances telles qu'évoquées Paragraphe 2.4.4.2, permet tout d'abord de résoudre partiellement le problème du fossé sémantique entre le multimédia, de plus en plus présent dans les bases documentaires, et le textuel, en fournissant une base solide à l'annotation d'images, vidéos, sons, etc. Un exemple pourrait être [Bontas et al., 2004] dans le domaine de la pathologie. L'enrichissement sémantique de la collection de documents apporte aussi de nouvelles dimensions sur lesquelles bâtir des requêtes, comme présenté dans le paragraphe suivant.

2.3.4 Interaction entre système et utilisateur

2.3.4.1 Introduction

Introduire la notion d'interaction au sein des systèmes de Recherche d'Information, c'est considérer qu'il s'agit d'un processus et prendre en compte sa nature intriquée et dynamique. C'est aussi rendre une place centrale à l'utilisateur, en tant qu'individu interagissant avec un système informatique, en tant qu'origine du besoin à satisfaire et juge des résultats.

En effet, les systèmes les plus simples de Recherche d'Information considèrent que les diverses étapes de recherche sont déconnectées. Chaque requête est traitée de manière indépendante des autres. Or, les études utilisateurs tendent à montrer une dépendance entre les diverses requêtes posées au cours d'une session de recherche. Le comportement de recherche ne consiste plus en une somme d'événements de recherche isolés, mais en un processus évolutif plus ou moins itératif.

De plus, le fossé cognitif et sémantique entre la requête et le besoin utilisateur, qui est au final, plus que la requête, l'élément à satisfaire, est important. Réduire ce fossé implique de réduire la distance entre l'utilisateur et ses besoins réels d'une part, et la requête d'autre part. Il s'agit alors de tenter de mieux comprendre l'utilisateur et de mieux le représenter au sein du système et de lui fournir des possibilités de formulation de requêtes plus expressives que des listes de mots clés.

Enfin, la présentation d'une liste ordonnée de documents jugés pertinents par le système n'est pas forcément la plus adaptée à la formation d'un jugement de pertinence par l'utilisateur et d'autres paradigmes de visualisation des résultats peuvent être envisagés, permettant une navigation plus souple et interactive dans l'espace des résultats.

Ces constats conduisent à la problématique d'interaction entre utilisateur et système d'information. Cette interaction peut se décliner selon plusieurs axes. Une première perspective est la nature dynamique du processus, qui a conduit à de nouveaux modèles de Recherche d'Information. Ensuite, il peut s'avérer pertinent d'introduire une meilleure prise en compte de l'utilisateur en tant qu'individu interagissant avec le système informatique. Cet axe n'est pas spécifique aux systèmes de Recherche d'Information et peut tirer parti des travaux réalisés dans le domaine des systèmes adaptatifs et adaptables introduits Section 2.4.2. Puis, l'interaction peut servir de base à la réduction de la distance entre le besoin non formalisé de l'utilisateur et la requête. Enfin, l'interaction avec l'utilisateur peut, tout comme pour les résultats d'un système de fouille de données, s'exprimer dans la dimension visualisation des résultats.

2.3.4.2 Un processus dynamique

Comme présenté précédemment, la Recherche d'Information est de plus en plus perçue comme un processus dynamique et non une juxtaposition d'épisodes de recherche indépendants. Ce constat a conduit à la mise en place de plusieurs modèles de Recherche d'Information incluant cette dynamique. En particulier, on peut citer le retour de pertinence (ou «relevance feedback» en anglais) ou des modèles comportementalistes qui ont conduit aux méthodes d'expansion de requête en collaboration avec l'utilisateur.

Le retour de pertinence est une notion qui est apparue relativement tôt dans les systèmes de Recherche d'Information, pour répondre à deux types de problèmes. À l'origine, il s'agit de pallier les difficultés des utilisateurs à exprimer une requête précise, sur une collection de documents qu'ils maîtrisent peu, pour représenter un besoin qu'ils ont du mal à cerner. Une conséquence de ces premières difficultés est le nombre souvent trop important de documents jugés pertinents par le système, parmi lesquels l'utilisateur doit trouver ceux qui sont vraiment intéressants. La solution du retour de pertinence est une contribution de l'usager en complément de la requête : parmi les nombreux documents que le système juge pertinents, l'utilisateur peut identifier des exemples de documents qui l'intéressent dans le cadre de sa recherche, afin d'aider le logiciel à affiner le processus de façon quantitative (retourner plus de documents semblables) et qualitative (chercher d'autres documents similaires). [Ruthven and Lalmas, 2003] présente une revue de cette notion ainsi que les divers cadres d'utilisation.

Historiquement un processus automatique qui modifie la requête de l'utilisateur de façon transparente, ainsi que le décrivent [Salton and Buckley, 1990], les processus de retour de pertinence ont évolué vers des systèmes collaboratifs. Dans ce cadre, les nouveaux termes jugés intéressants par le système ne sont que suggérés à l'utilisateur par le biais d'interfaces adaptées, type de systèmes qui d'après

[Koenemann and Belkin, 1996] sont plus appréciés des usagers, comme tous logiciels qui dépassent l'approche «boîte noire». Ce dernier constat a conduit à l'intégration aux campagnes d'évaluation TREC d'une piste interactive, pour laquelle les propositions de systèmes ont été explorées par [Belkin et al., 2001].

De plus, cette notion d'interaction a conduit à l'émergence de certains modèles comportementalistes qui prennent en compte cette dynamique. Ainsi, [Bates, 1989] définit le modèle de «berrypicking» (cueillette, butinage), qui considère que les utilisateurs ne cherchent pas uniquement de l'information de manière itérative autour d'un problème fixe, comme dans le retour de pertinence, mais que l'idée qu'ils se font de leur problème évolue, lorsque confrontée aux informations déjà accédées. Il ne s'agit pas d'un processus circulaire ou en spirale, mais d'un chemin aux multiples convolutions.

Surtout, cette dynamique de l'interaction fait apparaître la nature éphémère d'une requête représentant un besoin, étant donné la dynamique temporelle de ce besoin, ainsi que l'importance de la formulation de la requête.

2.3.4.3 Distance entre besoin et requête

2.3.4.3.1 Formulation de requête

En Recherche d'Information, la formulation de la requête joue un rôle déterminant pour la sélection des éléments d'intérêt, mais il s'agit pour l'utilisateur d'un problème complexe. En effet, la collection documentaire est souvent inconnue de l'utilisateur. De plus, il a souvent une idée très floue de ce qu'il veut rechercher, et encore plus de difficultés à exprimer cette idée floue en termes utilisables par un système de Recherche d'Information, étant donné le fossé cognitif et sémantique entre les deux. Ainsi, [Jansen et al., 2000], tout comme [Aula, 2003] montrent les difficultés des utilisateurs à créer des requêtes facilement exploitables par les systèmes de Recherche d'Information, quelle que soit leur expertise sur le domaine exploré : ils notent peu de reformulations, des requêtes courtes, peu d'utilisations correctes des opérateurs booléens, etc.

Afin de faciliter cette formulation, de nombreux auteurs proposent de s'appuyer sur une expression à un niveau d'abstraction supérieur, par le biais de requêtes dites «conceptuelles» [Bloesch and Halpin, 1997], qui permettent l'expression de notions familières aux utilisateurs, ou «sémantiques» [Stuckenschmidt et al., 2004, Dongilli et al., 2004], qui permettent l'exploration de volumineuses collections de documents en se basant sur des ontologies ou thésaurus, ou même en langue naturelle, se basant sur des techniques de traitement automatique des langues. Certains systèmes proposent même des visualisations graphiques de l'espace d'interrogation [Catarci et al., 2004], pour faciliter l'expression de la requête qui devient

non plus textuelle mais graphique. Dans le même esprit, [van Zwol and Apers, 2001] proposent une exploration par thème de la collection documentaire, en support de la requête. [Mäkelä et al., 2005] introduisent la notion de vues, permettant une saisie de requêtes selon diverses facettes de la vue à construire. La plupart de ces techniques impliquent le recours à une forme de représentation de connaissances, sujet du Paragraphe 2.4.4.2.

Mais malgré ces efforts d'extension de la requête, le mode d'interrogation demeure une requête de sélection par le contenu : l'expression de la requête est facilitée et mieux ancrée dans les données, mais le problème de la distance entre le besoin non formulé de l'utilisateur et sa formulation sous forme de requête reste inchangé. Il faudrait alors permettre l'expression fonctionnelle de l'objectif de la recherche et des besoins d'information de l'usager, c'est-à-dire introduire un niveau d'abstraction supérieur dans la requête. Une piste pourrait être l'introduction de la notion de tâche.

2.3.4.3.2 Introduction de la notion de tâche

Le sujet de la distance cognitive entre le besoin de l'utilisateur et son expression sous forme de requête reste central, en particulier dans le domaine de la recherche scientifique, où le problème est non seulement la recherche de l'information pertinente, mais aussi son intégration avec des connaissances préalables.

Ceci soulève la question de l'usage qui est fait de l'information trouvée. Ce problème est évoqué dans le domaine de la bioinformatique par [Bartlett and Toms, 2005], où la prédiction de la fonction d'une protéine d'après la séquence du gène correspondant implique plusieurs étapes successives de raffinement d'hypothèses. Dans le même esprit, dans [Hunter et al., 2004], la Recherche d'Information est intégrée comme faisant partie d'un protocole expérimental impliquant des stades progressifs de raffinement d'hypothèse. Mais introduire la notion de résolution de problème ou de tâche dans le processus de Recherche d'Information apparaît comme un problème majeur, comme le notent [Järvelin and Ingwersen, 2004]. Cette dimension reste encore sous-exploitée ou implicite à cause de ses difficultés de compréhension, analyse et expression.

Parmi les approches dans cette direction, [Johnson et al., 2003] visent à concevoir des taxonomies de tâches en observant le comportement de groupes d'utilisateurs pendant les sessions de recherche et en conduisant des entretiens avant et après la recherche pour avoir un aperçu des motifs, stratégies et état d'esprit des individus impliqués. Ces représentations de tâches peuvent être prises en compte dans la formulation de la requête en étendant les requêtes avec un vocabulaire orienté tâche, sélectionné par l'utilisateur [Liu et al., 2002b]. Mais ces taxonomies restent difficiles à exploiter, parce qu'elles sont soit trop générales [Choo et al., 1999], soit limitées à une tâche particulière [Bartlett and Toms, 2005].

Une autre méthode pour aborder l'intégration de connaissances est la construction d'une vue synthétique sur l'information ; plutôt que réduire l'affichage à une liste d'items, l'information retournée est organisée en documents synthétiques qui fournissent une vue générale sur l'information, cette vue étant conçue en fonction des besoins de l'utilisateur [Blake and Pratt, 2002]. Cette notion de document de synthèse implique de se focaliser sur des problèmes d'organisation et de présentation : ceci suggère de profiter de concepts du domaine de la Visualisation d'Information.

2.3.4.4 Visualisation de résultats d'une recherche

La Section 2.2 a permis de montrer l'intérêt des aspects visualisation dans l'appréhension d'un corpus informationnel. Mais elle a aussi mis en lumière un problème de sélection des éléments à intégrer à la présentation, ce qui suggérait la combinaison d'un système de Recherche d'Information avec les systèmes de visualisation. Mais la collaboration peut être aussi bénéfique de l'autre point de vue. En effet, en Recherche d'Information, se pose classiquement le problème du compromis entre retour et précision : un système précis retourne peu de documents pertinents, avec un risque de laisser de côté des informations très intéressantes ; un système visant l'exhaustivité retourne beaucoup de documents mais le taux de documents retournés qui sont peu ou pas pertinents est élevé. Beaucoup de systèmes courants, tels les moteurs de recherche sur Internet font le choix de maximiser le retour aux dépens de la précision, ce qui pose à l'utilisateur un problème majeur d'exploration de la liste des documents retournés. Des techniques de Visualisation d'Information pourraient aider à résoudre ce problème. Diverses approches dans cette perspective sont présentes dans la littérature, permettant en général d'apprécier la structure de l'espace des documents pertinents.

La première approche consiste à introduire d'autres dimensions de classification que la pertinence, pour résoudre en outre le problème du hors sujet, classique en Recherche d'Information. En effet, un grand nombre de documents non pertinents pour l'utilisateur sont souvent retournés pour cause de polysémie ou d'homonymie, etc. Une solution à ce problème est de présenter une vue organisée par groupes de documents, permettant ainsi à l'utilisateur de laisser de côté d'un seul coup des groupes de documents qui sont hors de son contexte d'intérêt. Ainsi que le présente [Hearst, 2006], la construction des groupes peut être soit prédéfinie, comme par exemple les catégories de Yahoo[3], soit réalisée à la volée par classification. Groupement basé sur des termes ou des phrases, à un seul niveau ou hiérarchique, les approches sont nombreuses et très différentes. On peut citer par exemple [Campos et al., 2006] ou [Ferragina and Gulli, 2005], qui proposent des meta-moteurs de recherche pour l'Internet utilisant une classification hiérarchique construite à partir d'extraction de phrases. [Xu et al., 2006] introduisent le même type d'approche pour la recherche de produits sur des sites marchands. [Glover et al., 2002] utilisent les liens entre pages

[3]http ://www.yahoo.com

Web et intra-page pour construire la hiérarchie et classifier les pages. La croissante popularité de ces approches a d'ailleurs conduit [Cigarrán et al., 2005] à proposer des métriques d'évaluation spécifiques, les mesures classiques de Recherche d'Information n'étant plus pertinentes.

Un second type de représentations vise à rendre tangibles les relations entre documents. Dans cette optique, on peut citer les représentations sous forme de graphes, comme par exemple dans le moteur de recherche Kartoo[4]. Dans le même esprit, [Fox et al., 2006] proposent un support pour les requêtes composées basé sur une métaphore de traversée d'une rivière sur des pierres : la représentation est réalisée sous forme de graphe des chemins entre les pierres que sont les sous-requêtes. Un autre type de visualisation considère une approche sous forme de tableaux à deux dimensions. Ainsi, [Kunz, 2003] propose une visualisation matricielle des conjonctions entre deux termes. [Arentz and Øhrn, 2004] associent des méthodes de classification avec des cartes de niveaux sous forme de tableaux à deux dimensions où la couleur d'une case correspond au niveau d'occurrence de l'intersection entre deux thèmes. [Chen, 2006] présente le même genre d'approche pour des données spatio-temporelles permettant l'association avec des cartes géographiques. [Carey et al., 2003] combinent plusieurs visualisations permettant d'apprécier les relations entre documents et termes : carte de Sammon, carte dendritique, visualisation radiale...

Ensuite, d'autres approches induisent une représentation de la dispersion des informations. Une première technique utilise des nuages de points, les axes correspondant en général à des attributs de la requête. Par exemple, [Atkinson et al., 2001] présentent un système de recherche d'attaques informatiques présentant les résultats sous forme de nuages de points en 3D. La métaphore géographique se prête bien à ce type de représentation, comme dans les cartes conceptuelles de [Driessen et al., 2006]. Les treemaps sont aussi utilisées dans ce contexte. Par exemple, [Großjohann et al., 2002] utilise des treemaps pour visualiser la distribution des termes recherchés dans des documents XML, tandis que des résultats de recherches sur les agences gouvernementales américaines sont visualisés ainsi par [Kules and Shneiderman, 2004].

Enfin, certains systèmes proposent de multiples représentations permettant d'apprécier diverses dimensions de la collection, comme dans [Mußler and Reiterer, 2001] avec des nuages de points, tilebars, barcharts...

2.3.5 Des limites à l'exploration d'informations

La synthèse a ici été évoquée dans une optique Recherche d'Information. Mais comme pour les problèmes de Recherche d'Information classiques, la résolution du problème de synthèse ne peut pas se limiter au modèle simple de correspondance

[4] http ://www.kartoo.com

entre une requête composée de mots clés et une représentation de la collection documentaire sous forme de vecteurs de mots.

Pour la Recherche d'Information, les tendances actuelles visent de façon générale à une meilleure prise en compte de l'utilisateur et son interaction avec le système, ainsi que du contexte de Recherche. Cette évolution implique l'intégration de paradigmes issus d'autres disciplines, telles que la Visualisation d'Information, la représentation de connaissances ou l'adaptation à l'utilisateur, autant de thématiques qui ont été identifiées comme d'intérêt pour la résolution de problèmes de synthèse.

Ce qui paraît surtout pertinent pour la synthèse dans le cadre des travaux sur les sciences de l'information est la notion de tâche, qui commence à apparaître en Recherche d'Information. L'objectif affiché est une prise en compte du contexte fonctionnel de la requête, de l'usage qui va être fait des informations pertinentes. Or, dans le cadre de la synthèse, la collecte d'informations n'est qu'une étape parmi d'autres et cette question de l'usage de l'information devient complètement centrale pour tout système d'assistance à la synthèse.

Cette notion de tâche utilisateur, conduisant au traitement personnalisé du problème complexe de synthèse, pose la question de la représentation des entités impliquées dans le processus de synthèse, étape préalable à toute résolution. Cette problématique de représentation tout à la fois de l'utilisateur, du problème à résoudre, et des connaissances nécessaires à sa résolution, fait l'objet de la prochaine section.

2.4 Représenter les entités impliquées

2.4.1 Introduction

Les sections précédentes ont permis d'identifier une tendance à la complexification des méthodes mises en œuvre pour la résolution des problèmes d'appréhension et exploration de l'information. Cette tendance est particulièrement notable pour la problématique d'exploration et la Recherche d'Information, avec la prise en compte de niveaux d'abstraction de plus en plus élevés. Elle pose la question de la résolution d'un problème complexe, surtout quand l'objectif est d'augmenter la pertinence en prenant en compte de façon plus précise le besoin utilisateur.

Comme présenté précédemment, cette ligne de recherche sous-entend une individualisation de l'interaction, l'introduction d'une notion de tâche et la prise en compte de connaissances du domaine. La résolution d'un problème de Recherche d'Information prenant en compte ces éléments de contexte pose le problème de la représentation de ces entités sous un format compréhensible par un système infor-

matique et par l'humain. Il devient alors nécessaire de définir tout à la fois une forme et des méthodes d'interprétation et exploitation de ces représentations.

Ainsi, en ce qui concerne l'individualisation de l'interaction, il s'agit de représenter l'utilisateur et adapter l'interface et le contenu en fonction de l'utilisateur interagissant avec le système. La notion de tâche implique de représenter le problème à résoudre et mettre en place des modes de raisonnement sur ce problème. Utiliser des connaissances requiert une représentation de ces connaissances et leur utilisation comme support au raisonnement et la Recherche d'Information. Ces trois thématiques de représentation et exploitation des notions d'utilisateur, de problème et de connaissances sont développées dans les prochains paragraphes.

2.4.2 Représenter l'utilisateur

2.4.2.1 Introduction

L'individualisation de l'interaction avec la machine est une problématique relativement ancienne, née des difficultés que nombre d'utilisateurs potentiels rencontrent face à l'outil informatique : apprentissage de l'utilisation d'un logiciel coûteux en temps, organisation des éléments de menu pas forcément intuitive pour tous, etc. La synthèse, qui induit la construction d'une vue destinée à un public particulier, sous-entend cette problématique d'individualisation, qui devient de plus en plus fréquente tout à la fois dans les logiciels commerciaux et en tant que thème de recherche.

La personnalisation de l'interaction avec la machine est en particulier l'objectif de l'hypermédia adaptatif, tel que présenté par [Brusilovsky, 2003]. L'ajout d'éléments de personnalisation passe par la réalisation d'un modèle de l'utilisateur, représentant ses connaissances, buts et intérêts, ou tous autres éléments permettant de distinguer un utilisateur donné. Il s'agit alors d'exploiter ces profils au sein des systèmes pour proposer des services adaptés.

2.4.2.2 Construction d'un profil utilisateur

La modélisation de l'utilisateur passe souvent par la définition d'un stéréotype d'utilisateur, représentant par des jeux d'attributs les caractéristiques d'un utilisateur générique ou d'un groupe d'utilisateurs donné, ainsi que [Kobsa, 2001] le montre dans sa revue du domaine. Il s'agit alors de permettre une personnalisation de ce modèle générique par chaque utilisateur particulier.

La construction d'un profil utilisateur ou la spécialisation d'un modèle utilisateur

est réalisable selon un panel de méthodes très étendu, ainsi que [Koutri et al., 2002] le montrent dans leur exploration de cette thématique. Schématiquement, deux grandes tendances sont à noter. La personnalisation peut consister en un processus interactif impliquant des saisies dans une interface particulière, dans le cas des systèmes adaptables. La génération du profil peut aussi être automatique, résultant par exemple de l'analyse de traces d'usage de l'application, dans le cas des systèmes adaptatifs. Une illustration de cette dernière approche pourrait être [Trousse, 2000], qui présente un système qui vise à prédire le comportement de l'utilisateur en se basant sur son comportement passé, afin de lui faire des suggestions d'actions. [Kim et al., 2002] comparent deux approches (saisie directe et construction par le biais d'agents logiciels représentant l'utilisateur) et suggèrent une combinaison des deux.

Le problème qui se pose alors est de déterminer quelles connaissances sur l'utilisateur sont pertinentes afin de proposer le contenu le mieux adapté. Dans cette optique, [Nguyen et al., 2006], au cours de leur processus de construction de communautés d'utilisateurs similaires, évaluent les variables pertinentes pour les profils d'usagers d'un système de recommandation de films.

2.4.2.3 Modes d'adaptation

L'adaptation passe par l'exploitation du profil ou des préférences de l'utilisateur. Cette adaptation peut être réalisée à deux niveaux différents : l'interface du système et le contenu proposé à l'usager.

En ce qui concerne l'interface, les systèmes les plus simples utilisent des valeurs d'attributs comme paramètres d'éléments d'affichage, par exemple un thème de gestionnaire de fenêtres décrivant des couleurs de bordure et de fond, etc. Des systèmes plus évolués ont recours à des outils de construction d'interface adaptables. Par exemple, [Furtado et al., 2003] proposent une plateforme qui se base sur des ontologies décrivant le domaine d'application du système et des stéréotypes d'utilisateurs et qui génère des interfaces différentes selon la catégorie d'utilisateurs.

La seconde problématique est l'adaptation du contenu. Celle-ci induit souvent l'utilisation d'éditeurs spécialisés, qui permettent à l'auteur de définir plusieurs versions d'un même document, la version proposée à l'usager dépendant du contexte d'utilisation. Ainsi, [Garlatti and Iksal, 2003] se basent sur des ontologies pour permettre à l'auteur la construction de documents spécifiques des niveaux connaissances du domaine des lecteurs, [Tran-Thuong and Roisin, 2003] propose un éditeur de documents multimédia permettant la création de contenu variable selon le matériel de présentation, et [Chevallet et al., 2005] permettent l'affichage d'informations touristiques sur un téléphone mobile en fonction de la localisation géographique de l'utilisateur, localisation déterminée par GPS et une photo des lieux envoyée au système

par l'usager. La difficulté réside alors principalement au niveau de l'auteur, qui se trouve contraint de construire et maintenir plusieurs versions concurrentes d'un même document.

2.4.3 Représenter le problème

2.4.3.1 Introduction

Depuis l'apparition des premiers calculateurs, des problèmes de plus en plus complexes ont été soumis aux systèmes informatiques, faisant émerger des besoins d'apprentissage, de raisonnement de la part des ordinateurs, donnant naissance à la notion d'Intelligence Artificielle. Mais des machines capables d'un comportement intelligent, d'une compréhension de leurs propres raisonnements et même d'une conscience de soi, qui soient en bref à même de réussir le test de Turing [Turing, 1950] sont encore bien loin de la portée humaine.

Une approche plus pragmatique de la notion d'Intelligence Artificielle vise à augmenter l'autonomie des systèmes et à simuler l'intelligence humaine sans la reproduire, apportant des solutions pratiques à des problèmes difficiles, parmi lesquels le problème de synthèse trouve sa place. Évoluant depuis la Recherche Opérationnelle des années 60 et les systèmes experts ou d'aide à la décision des années 80, l'Intelligence Artificielle, de nos jours, s'inscrit dans une perspective de programmation d'un apprentissage et de collaboration avec l'humain ou assistance à l'utilisateur dans des tâches complexes. Diverses approches peuvent être évoquées, qui peuvent s'avérer pertinentes dans le contexte de synthèse.

Un premier ensemble de systèmes vise à atteindre une autonomie de plus en plus grande, comme les systèmes multi-agents, dont l'approche est analysée dans [Ricordel and Demazeau, 2000]. Constitués d'un ensemble d'entités logicielles autonomes, communiquant entre elles, placées dans un environnement et collaborant pour atteindre un but, ces systèmes comptent sur l'apparition de comportements émergents, non codés dans les entités logicielles élémentaires. Mais cette émergence, même si elle permet parfois l'apparition de solutions originales et innovantes à des problèmes complexes, pose un problème de non reproductibilité potentielle des résultats, qui n'est pas souhaitable dans un contexte de recherche scientifique tel que la synthèse.

On peut citer aussi deux grandes familles de méthodes qui mimiquent le raisonnement humain. Le Raisonnement à Partir de Cas, ou «Case Based Reasoning» (CBR) en anglais [Aamodt and Plaza, 1994], exploite un raisonnement par analogie, en supposant que des problèmes similaires auront des solutions similaires. Mais il présuppose l'existence d'une base de cas connus et une certaine immuabilité du

domaine d'application. Or l'acquisition de cas passés est en pratique difficilement réalisable, en particulier dans une activité telle que la synthèse où les problématiques et le champ à étudier sont fluctuants. Une autre approche, extension du raisonnement à base de règles, s'applique plutôt à représenter les problèmes à résoudre et les méthodes permettant d'arriver à leur résolution, point de vue qui semble mieux en adéquation avec la notion de synthèse, où la notion de but ou tâche est centrale, et qui sera plus explicité dans le prochain paragraphe.

2.4.3.2 Problèmes et méthodes de résolution

La réalisation d'un document de synthèse dans un but particulier et pour un public donné est un problème complexe. Cette complexité suggère une analyse plus poussée de cette notion de but, de tâche utilisateur. Ces notions ont fait l'objet de nombreuses études par la communauté d'Intelligence Artificielle, menant à diverses normalisations et définitions de tâches et algorithmes associés, tels ceux développés autour de la méthodologie CommonKADS [Sacile, 1995] ou des méthodes de résolution de problèmes (Problem-Solving Methods ou PSM) [Fensel et al., 2001] pour lesquelles [Fensel et al., 2003] proposent un langage de spécification.

Mais ces méthodes s'attachent peu à formaliser l'effort de focalisation nécessaire à la résolution du problème, qui s'avère indispensable dès lors que de vastes espaces informationnels sont considérés. De plus, elles sont forcément limitées : tous les problèmes et toutes les pistes de résolution ne sont pas accessibles au concepteur humain et certains problèmes font atteindre les limites computationnelles des ordinateurs.

Par contre, elles suggèrent deux pistes de conception particulièrement pertinentes : décomposition du problème en sous-tâches unitaires et recours à des composants logiciels [Szyperski, 2003] indépendants et réutilisables, permettant chacun de résoudre un problème unitaire, combinés en fonction de la tâche particulière à résoudre. Dans ce contexte, [Crubézy and Musen, 2004] proposent un système implémentant des méthodes de résolution de problèmes sous forme de composants logiciels et usant d'ontologies pour représenter tout à la fois les connaissances du domaine et les éléments de raisonnement.

2.4.4 Représenter des connaissances

2.4.4.1 Introduction

Qu'il s'agisse de systèmes de Recherche d'Information ou de systèmes basés sur des paradigmes d'Intelligence Artificielle tels que présentés dans le paragraphe précédent, les logiciels actuels reposent de plus en plus sur des représentations du monde

qui leur soit intelligible.

Par exemple, dans le cadre du paradigme de Recherche d'Information en Contexte, le premier élément qui peut être enrichi par delà une représentation de type «sac de mots» est la collection de documents. Cet enrichissement passe généralement par une représentation au niveau conceptuel et non plus simplement statistique sur des termes. Dans le même esprit, les méthodes de résolution de problèmes s'appliquent sur des entités du monde qui doivent leur être accessibles, ce qui nécessite un formalisme adapté à une manipulation par des programmes informatiques tout en restant lisible par l'humain qui les conçoit.

Ces besoins de représentations de connaissances impliquent la mise en place de formalismes particuliers, au sein d'une infrastructure adaptée, telle que présentée dans le prochain paragraphe.

2.4.4.2 Taxonomies, thésaurus et ontologies

L'infrastructure couramment utilisée dans ce cadre de représentation des connaissances est fournie par le Web sémantique, proposition relativement récente du W3C [5], élaborée dans [Berners-Lee et al., 2001], et dont [Laublet et al., 2002] ou [Ding et al., 2002] présentent les caractéristiques principales.

Cette infrastructure vise à l'intégration d'ontologies au Web. Pour [Gruber, 1993], qui introduit cette notion de philosophie en informatique, il s'agit de représentations des concepts d'un domaine et des relations existant entre eux pouvant être utilisées par des agents en particulier logiciels. [Guarino, 1998], qui insiste sur les particularités méthodologiques et architecturales, en particulier pluridisciplinaires, de ces ontologies, présente quelques exemples de leur utilisation potentielle. Une définition consensuelle de cette notion reste, malgré son usage de plus en plus répandu, encore à définir, et au sens le plus large, de simples taxonomies ou graphes de termes, ou des thésaurus, peuvent être considérés comme des ontologies.

Cette intégration d'ontologies doit se baser sur des langages spécifiques orientés contenu. [Noy et al., 2001] font une présentation succincte de ces langages et des relations existant entre eux et montrent comment utiliser l'outil Protégé-2000 de l'Université de Stanford [6] pour saisir des informations. Pour la plupart, ces langages sont basés sur RDF, un langage de syntaxe de type XML. L'expressivité de RDF reste pourtant limitée, et des langages tels que OIL [Fensel et al., 2001] puis OWL, qui fournissent du vocabulaire additionnel et une sémantique formelle ont été introduits pour la représentation d'ontologies, en particulier sur Internet.

[5] http ://www.w3.org/
[6] http ://protege.stanford.edu/

Dans le cadre du Web sémantique, ces ontologies peuvent alors être exploitées par divers types d'entités logicielles. Ainsi [Soshnikov, 2002] présente une architecture multi-agent distribuée où des agents localisés accèdent à des fragments d'ontologies situés sur des nœuds du réseau différents pour réaliser des tâches impliquant des inférences. L'objectif d'une architecture de ce type est la mise en place d'agents logiciels qui assistent l'utilisateur dans ses tâches, par exemple la gestion de l'attribution d'organes en vue de transplantations dans [Cortés et al., 2000], ou la Recherche d'Information sur le Web pour [Ngu and Wu, 1997]. Des fonctionnalités distantes peuvent aussi être exploitées par le biais des services Web, infrastructure pour l'interaction entre applications bâtie comme une couche supplémentaire par dessus les protocoles Web existants et basée sur des standards XML qui est présentée dans [McIlraith et al., 2001] ou [Curbera et al., 2002].

L'intégration d'ontologies au Web implique alors d'évaluer comment enrichir le contenu afin de faciliter les tâches de l'utilisateur sans pour autant surcharger les documents avec des méta-données inutiles car inutilisées. Il s'agit aussi de maintenir constant un effort de mise à jour de ces méta-données, tout en se posant la question de leur potentielle universalité, ou au constraire de la spécificité d'une représentation pour une application particulière.

2.4.5 Des limites aux représentations d'entités

La synthèse a été abordée dans cette section sous un angle de représentation d'entités impliquées dans un processus complexe. Comme nombre de systèmes informatiques où la notion d'interaction prend une place de plus en plus prépondérante, des systèmes d'assistance à la synthèse se doivent d'individualiser leur relation à l'usager en se basant sur des représentations de l'utilisateur. Comme nombre de systèmes qui visent à assister un utilisateur engagé dans des tâches complexes, des systèmes d'assistance à la synthèse se doivent de mieux appréhender les opérations à réaliser par une représentation formelle des problèmes à résoudre. Comme nombre de systèmes manipulant de l'information, des systèmes d'assistance à la synthèse se doivent de se baser sur une image du monde passant par une représentation de connaissances.

Des représentations de l'utilisateur, du problème à résoudre et des connaissances apportent au système une flexibilité tout à la fois dans son interaction avec l'usager et dans ses raisonnements. Mais elles posent un ensemble de problèmes difficiles. Tout d'abord, elles impliquent la définition d'un formalisme spécifique qui soit compréhensible tout à la fois par des programmes informatiques et par des humains. Ensuite, elles posent le problème de la définition de la liste des entités à représenter ainsi que des relations entre elles et des éléments qu'il est pertinent de formaliser pour chaque entité. Se pose alors la question de la généricité d'une telle représentation ou de son lien étroit avec l'application considérée. Enfin elles requièrent un effort constant de rédaction, maintenance et évolution.

2.5 La synthèse : une activité aux multiples affiliations

La notion de synthèse a été introduite de manière succincte dans ce chapitre, en tant que réponse possible au double problème d'appréhension de données posé par la technologie des Tissue MicroArrays : l'appropriation préalable à une fouille d'un espace de données complexe et le replacement de données acquises en masse dans une démarche expérimentale classique. Les quelques caractéristiques primordiales de cette synthèse suggèrent sa parenté avec des techniques de diverses disciplines qui visent à appréhender des données, explorer des informations ou représenter des entités impliquées dans un processus complexe.

Dans le contexte de ces trois catégories d'appréhension des données, exploration d'informations et représentation d'entités, la représentation des entités impliquées apparaît comme un outil de plus en plus essentiel aux deux autres. Elle permet en effet une interactivité et un ancrage dans le champ applicatif étudié. Deux points de vue sur la synthèse peuvent alors être envisagés : un point de vue fouille de données ou un point de vue Recherche d'Information.

Le point de vue fouille de données, de part la multitude des méthodes et des modes de représentations visuelles des données qu'il fédère, semble tout à fait approprié. Mais il présente un certain nombre de limites. En particulier, les méthodes de fouille sont directement ancrées dans les données explicitement présentes dans la collection analysée. La prise en compte explicite d'une notion de but ou tâche a priori est difficile. Replacer des données dans une démarche expérimentale classique reste en général une activité manuelle.

Or, la plupart de ces limites d'un point de vue fouille de données semblent trouver une solution dans un point de vue Recherche d'Information par son entrée dans le système par le biais d'une requête. En effet, la requête, qui peut être sémantique, permet alors une interrogation de la collection à un niveau implicite, conceptuel. Cette requête permet l'expression plus ou moins élaborée d'un besoin. Le résultat de la Recherche d'Information regroupe des documents pertinents par rapport à la requête. Ceci peut alors être transposé comme informations pertinentes par rapport à une hypothèse à valider.

Le point de vue adopté sur la synthèse dans ma thèse est donc un point de vue Recherche d'Information. Mais ce point de vue Recherche d'Information n'est pas suffisant au sens strict du terme de Recherche d'Information, c'est-à-dire de simple correspondance entre représentations de l'information et requête pour constituer une liste ordonnée de documents pertinents.

La synthèse est une tâche difficile et, tout comme un besoin d'aller vers une

représentation moins naïve des entités impliquées s'est fait sentir dans les systèmes de Recherche d'Information actuels, la synthèse implique une prise en compte de l'utilisateur, de ses besoins, de sa requête, du corpus documentaire ou des résultats de la recherche à un plus haut niveau d'abstraction. Cette complexification sous-tend le recours à des concepts de représentation de l'utilisateur, du domaine d'étude, du problème à résoudre, de la collection de documents et du résultat de la synthèse, concepts issus des domaines de l'interaction homme/machine, des systèmes adaptatifs, de l'Intelligence Artificielle ou de la Visualisation d'Information. La synthèse paraît alors aussi comme multifacettes.

Ces considérations sur la notion de synthèse ont conduit à la construction d'un cadre plus formel pour cette notion, exposé dans le prochain chapitre. Cette proposition consiste tout d'abord en la mise en place d'un modèle de synthèse, inspiré des modèles de Recherche d'Information puis à l'explicitation des entités impliquées.

CHAPITRE

3 Bases conceptuelles pour la synthèse d'information

La notion de synthèse, telle que proposée comme solution aux problèmes d'appropriation de données TMA par l'utilisateur, a été introduite dans le précédent chapitre. Surtout, elle a été envisagée sous différents points de vue : appréhension des données, exploration d'informations et représentation d'entités impliquées dans le processus. Ceux-ci ont permis d'aborder la synthèse selon divers champs disciplinaires : fouille de données et Visualisation d'Information, sciences de l'information, et représentation de l'utilisateur, de problèmes et de connaissances. Un état de l'art rapide de ces différents domaines a permis d'identifier les relations de ces champs de recherche à la synthèse et a conduit à l'adoption d'un point de vue Recherche d'Information, augmenté de paradigmes d'autres disciplines. Ce point de vue sur la synthèse va être explicité plus en détails dans ce chapitre, en particulier par le biais de la construction d'un modèle de synthèse, qui va être replacé dans le contexte des modèles classiques de Recherche d'Information et dont les diverses composantes vont être analysées en détails.

3.1 Introduction

La notion de synthèse a été envisagée comme solution au problème général d'appréhension de gros volumes de données, qui a été évoqué dans le contexte particulier la technologie des Tissue MicroArrays. Une telle appropriation de la collection de données est en effet nécessaire, en préalable à une fouille de données ou pour se replacer dans une démarche expérimentale classique.

La synthèse est ici considérée comme un processus de collecte, agrégation, organisation, présentation d'informations hétérogènes dans un objectif précis, pour un public particulier. En tant que telle, elle implique de s'intéresser à diverses perspectives d'exploration de données, appréhension d'informations ou représentations d'entités, qui peuvent être rattachées à diverses communautés de recherche, tel que détaillé au chapitre précédent.

Une revue rapide des divers champs de recherche d'intérêt pour cette notion de synthèse a conduit à l'adoption d'un point de vue Recherche d'Information, augmenté de paradigmes issus d'autres domaines, tels que la fouille de données, la Visualisation d'Information, les systèmes adaptatifs ou l'Intelligence Artificielle. Mais si les diverses caractéristiques de la synthèse trouvent plutôt bien des propositions d'implantation par les différentes disciplines évoquées précédemment, l'une reste peu supportée dans un cadre de Recherche d'Information : la notion d'objectif ou de but est en effet encore rarement prise en compte.

Or cette notion de but est centrale pour la synthèse telle qu'elle est envisagée ici. En effet, en tant qu'outil permettant une appréhension de l'espace des données préalable à une fouille, elle se doit d'être dirigée, l'application de méthodes de fouille sans objectif sous-jacent conduisant rarement à la découverte de connaissances d'intérêt. En tant qu'outil permettant de replacer un gros volume de données dans une démarche expérimentale classique, elle est dépendante de l'hypothèse particulière à évaluer. Le point de vue adopté sur la synthèse est donc un point de vue centré sur l'usage, où la Recherche d'Information est envisagée comme une étape d'un processus plus complexe de résolution de problème.

Ce dernier point de vue requiert de se concentrer sur la notion de tâche, qui, dans le domaine des Tissus MicroArrays en particulier, ou dans celui de la recherche scientifique de manière plus générale, peut être considérée comme une autre transcription de l'étude que le chercheur veut réaliser. La multiplicité des études qui peuvent être réalisées à partir d'un même jeu de données fait alors de la synthèse une notion fédératrice pour une multitude de tâches complexes.

Il s'agit donc de proposer à l'utilisateur un système informatique assez flexible pour permettre tout à la fois :

- ⋆ l'expression de ces tâches multiples,

- ⋆ la résolution du problème de construction d'une vue synthétique apportant une assistance dans la réalisation de ces tâches,
- ⋆ la présentation de cette vue de manière adaptée aux tâches et à l'utilisateur.

La conception d'un tel système requiert tout d'abord une analyse plus poussée de la notion de synthèse dans ses diverses assertions. En particulier, la complexité sous-jacente suggère une limitation du champ couvert dans cette thèse au contexte applicatif des Tissue MicroArrays et à la synthèse de données.

L'objectif poursuivi est la construction d'un modèle de synthèse s'intégrant dans le contexte des modèles de Recherche d'Information existants. Les composantes de ce modèle devront être évaluées plus en détail afin d'ouvrir la voie à l'implantation d'un prototype le mettant en œuvre.

3.2 Point de vue adopté sur la synthèse

3.2.1 Introduction

Selon le point de vue adopté dans ma thèse, la synthèse est un processus de construction d'une vue compacte et personnalisée sur un domaine applicatif particulier, réalisée dans un but précis à partir de l'assemblage de données et informations hétérogènes.

Le chapitre précédent a permis d'évoquer diverses disciplines d'intérêt pour la synthèse, ce qui en fait une notion aux multiples facettes. La synthèse est aussi une notion fédératrice pour une multitude de problématiques différentes, dont l'objectif commun est la construction d'une vue organisée et argumentée sur un champ disciplinaire particulier. Enfin, la synthèse est une notion difficile, reposant sur des mécanismes cognitifs évolués, ce qui pose la question de la résolution automatisée de ce problème.

En particulier, la complexité esquissée ici suggère qu'il est utopique d'imaginer résoudre le problème de synthèse de manière générique, en couvrant l'ensemble des tâches de synthèse et l'ensemble des corpus documentaires possibles. La solution envisagée est alors de limiter le problème en le considérant dans un domaine applicatif particulier : le domaine des Tissue MicroArrays.

Les divers points de vue sur la synthèse, notion aux multiples facettes, aux multiples incarnations et à la résolution difficile, conduisant à une approche ancrée dans le domaine applicatif vont être évoqués de manière plus approfondie dans la suite de cette section.

3.2.2 Une notion complexe

3.2.2.1 Une notion aux multiples facettes

Comme vu précédemment, la synthèse est considérée comme une activité de collecte, agrégation, organisation, présentation d'informations hétérogènes dans un objectif précis, pour un public particulier. La synthèse est donc une opération complexe qui se décompose en un ensemble d'activités très diverses.

Tout d'abord, la synthèse implique des activités de définition de l'objectif poursuivi par le chercheur engagé dans une telle opération. Cette définition d'objectif peut être rapprochée de la formulation de requête de la Recherche d'Information. Cette formulation doit dépasser la simple liste de mots clés, pour permettre l'expression d'une tâche de synthèse complexe spécialisée pour une étude spécifique d'un domaine d'application particulier.

Deuxièmement, la synthèse requiert la mise en œuvre d'activités de collecte de documents et données pertinents. Collecter des informations pertinentes est le cœur des processus de Recherche d'Information, par l'application de fonctions de correspondance. Mais la simple mise en correspondance entre requête et documents ne suffit pas dans le contexte de la synthèse, et cette recherche d'informations pertinentes devient alors une étape du processus de synthèse.

Ensuite, la synthèse sous-tend des activités d'extraction d'informations utiles au sein des éléments pertinents sélectionnés. Cette extraction d'informations rappelle fortement le cœur de métier de la fouille de données, mais ne s'y limite pas, puisqu'elle pourrait aussi consister en la sélection plus précise de parties d'entités pertinentes. Un exemple pourrait être la recherche de passages en Recherche d'Information, qui propose la construction d'une liste d'extraits de documents pertinents plutôt qu'une liste de documents pertinents.

Puis, la synthèse induit des activités d'organisation conceptuelle des informations utiles, ce qui suggère une notion de classification telle qu'elle se retrouve en fouille de données ou dans les propositions de présentation de résultats de Recherche d'Information les plus avancées. Mais cette classification n'est pas forcément réalisée dans l'absolu, afin d'obtenir les groupes reflétant le mieux la structure du jeu de données : elle peut aussi être guidée par l'objectif de la synthèse.

Par la suite, la synthèse suppose des activités de disposition des éléments au sein d'un document qui reflète de façon structurelle l'organisation conceptuelle définie, problématique qui relève de l'analyse visuelle des données, et surtout de la Visualisation d'Information, dont les représentations sont plus évoluées. Mais tous les types de représentations ne sont sans doute pas adaptés à la problématique de synthèse, et le volume de données à présenter dans un espace réduit suggère de privilégier

les représentations les plus compactes, dans un souci d'adaptation tout à la fois au problème et à l'usager.

Ultérieurement, la synthèse inclut des activités de mise en forme du document final, mise en forme qui repose sur la construction d'un document multimédia complexe. Cette construction est de plus en plus personnalisée dans les applications actuelles et cette adaptation de la forme est le propre des systèmes adaptatifs.

Ensuite, la synthèse nécessite des activités de vérification de la qualité des productions réalisées au cours des autres phases. Cette notion de qualité se rapproche de la notion de pertinence de la Recherche d'Information, en y incluant non seulement une évaluation de la liste des éléments sélectionnés, mais encore une évaluation de leur organisation et du rendu de cette organisation. Mais cette mesure n'est pas suffisante, et un processus d'évaluation complet de l'ensemble du système d'assistance à la synthèse doit être envisagé, comme pour toute production issue de l'ingénierie logicielle.

Enfin, la résolution du problème de synthèse est une activité complexe, qui mobilise connaissances et méthodes variées : connaissances du domaine étudié, connaissances des procédures expérimentales spécifiques de ce domaine, préférences ou expériences personnelles et comportements typiques issus de la formation des chercheurs. Ceci sous-tend des représentations formelles de l'ensemble de ces entités, par le biais non seulement d'ontologies, mais aussi de profils utilisateurs et de représentations de problèmes, problématiques qui relèvent tout à la fois des systèmes adaptatifs et systèmes à base de connaissances de l'Intelligence Artificielle.

La synthèse est donc une activité multiple par les diverses facettes qu'elle présente, tel qu'illustré par la Fig. 3.1. Mais cette multiplicité ne se limite pas à cette dimension pluridisciplinaire et couvre aussi une dimension intrinsèque à la notion de synthèse : la multiplicité des problématiques qu'elle fédère, présentée plus en détails dans le prochain paragraphe.

3.2.2.2 Une notion fédératrice pour une multitude de problèmes

La synthèse, dans l'assertion considérée, est une opération intellectuelle visant à réunir les diverses parties d'un ensemble au sein d'une vue unifiée. Cette construction est réalisée dans un objectif particulier. Or cet objectif est loin d'être unique : en un sens, il n'y a pas une synthèse mais des synthèses.

En effet, dans un contexte de synthèse dans une visée scientifique, les chercheurs peuvent adopter des points de vue très variables sur les documents ou données disponibles, points de vue qui dépendent de l'objectif qu'ils ont en tête. La synthèse peut alors être considérée, de manière simplifiée, comme une transformation

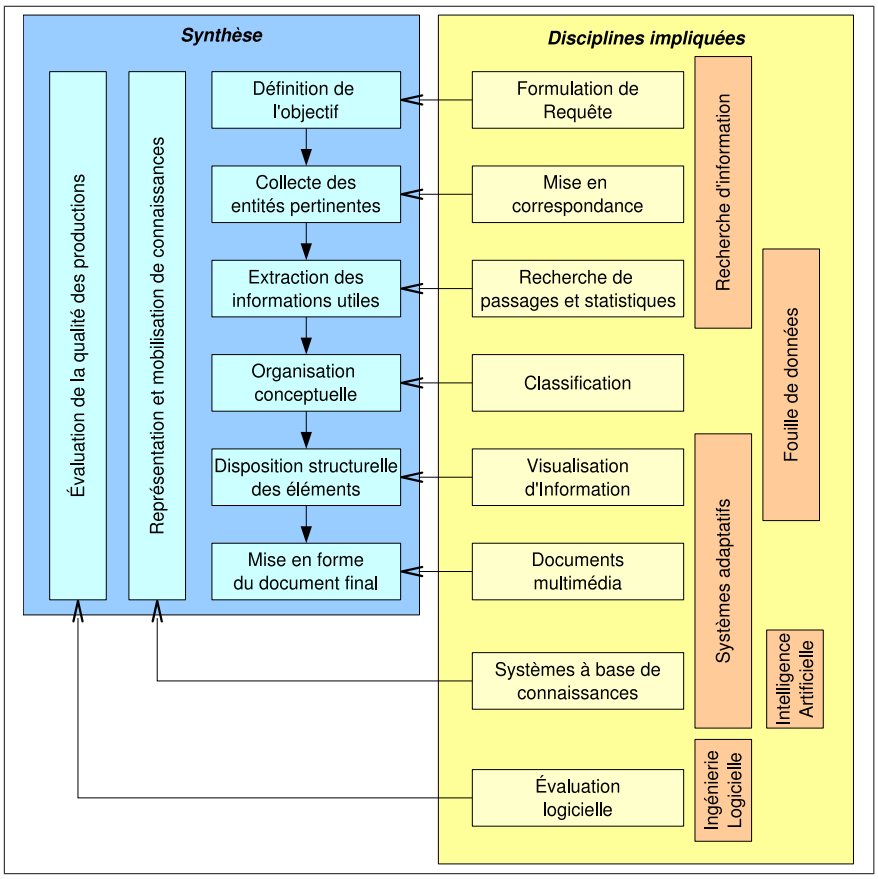


Fig. 3.1: La synthèse, une activité multifacettes - Les diverses activités successives (à droite de l'encadré «Synthèse») ou servant de support aux autres (à gauche de l'encadré «Synthèse»), qui sont impliquées dans la synthèse sont supportés par les disciplines exposées de l'encadré «Disciplines impliquées». Chaque activité peut être associée à des techniques dont le domaine d'appartenance est indiqué verticalement.

des informations disponibles. Chaque point de vue peut alors être regardé comme résultant d'une transformation particulière, correspondant à une tâche de synthèse précise. La multiplicité de la synthèse réside ensuite aussi au niveau des éléments subissant une transformation donnée, le résultat final dépendant des documents sur lesquels elle est appliquée.

Métaphoriquement, cette transformation peut être assimilée à l'effet d'une lentille en optique. La déviation des rayons lumineux par la lentille dépend de la forme de la lentille : sa taille, son épaisseur, son rayon de courbure. L'ensemble des formats de lentilles possibles représente alors la notion de synthèse en général. Une lentille aux caractéristiques particulières peut être assimilée à une tâche de synthèse spécifique. De plus, en optique, l'image obtenue après passage des rayons lumineux au travers de la lentille dépend de l'objet observé. Cette sensibilité à l'objet étudié peut être considérée comme une image de la multiplicité de la synthèse en rapport aux éléments étudiés.

Cette métaphore optique illustrant la multiplicité de la synthèse, est schématisée Fig. 3.2. Mais bien qu'elle rende compte de la double démultiplication du problème de synthèse, par les diverses tâches et divers corpus documentaires, elle reste limitée : une image optique reste très similaire à l'objet d'origine, alors que la synthèse implique une transformation complexe, par sélection et agrégation d'entités, qui déforme l'objet/l'espace informationnel observé au sein de son image/résultat de synthèse.

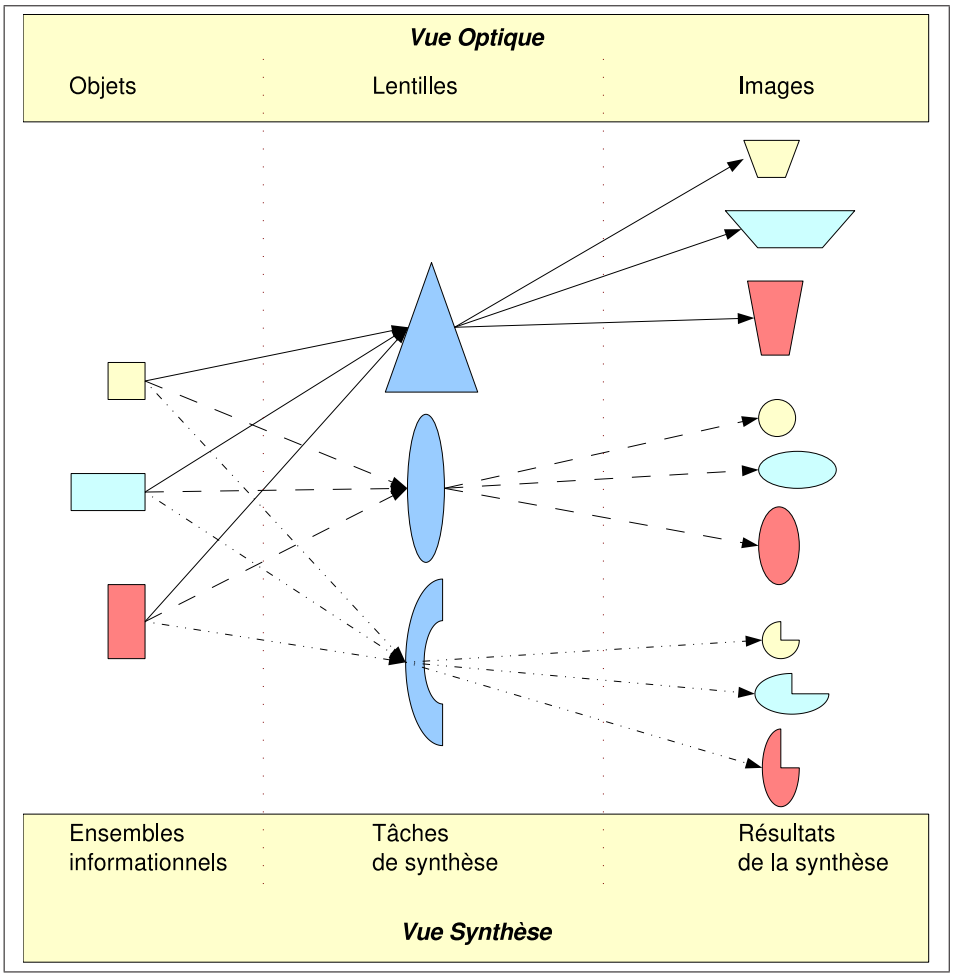


Fig. 3.2: La synthèse, fédération de multiples problèmes - Les divers ensembles informationnels étudiés sont représentés par des objets différents. Ils sont transformés par diverses tâches de synthèse, dont le fonctionnement est assimilé à celui d'une lentille optique. La transformation de chaque objet par chaque lentille conduit à une image résultat unique, de la même façon qu'une tâche de synthèse réalisée à partir d'un jeu de documents induit un document synthétique particulier.

La synthèse fédère donc un espace complexe de problèmes, par la multitude des tâches qu'elle regroupe et par la multitude de contextes d'application, domaines d'étude ou documents utilisés, qui peuvent être envisagés pour chaque tâche. Mais contrairement à l'optique où des lois bien connues régissent la construction de l'image d'un objet par une lentille, la synthèse est une notion dont le mode de résolution est mal défini car elle est difficile, comme exposé dans le prochain paragraphe.

3.2.2.3 Une notion difficile

Le processus de synthèse implique de nombreuses activités qui relèvent de diverses disciplines. Or, chacune de ces activités est en soit un problème complexe. Ainsi, par exemple, la sélection d'entités pertinentes par des méthodes de Recherche d'Information est un domaine de recherche toujours très actif car les résultats des systèmes de Recherche d'Information actuels sont satisfaisants mais pas parfaits : certains documents pertinents ne sont pas retrouvés alors que des documents hors sujet sont retournés.

De plus, il s'agit d'une notion fédératrice pour de multiples tâches, qui peuvent être réalisées dans différents contextes sur des collections documentaires variés. Cette multiplicité des instances du concept de synthèse suggère qu'il n'y a pas de solution unique au problème de synthèse, mais plutôt que chaque tâche peut être considérée comme un problème à part entière.

La synthèse est donc une opération difficile, et se pose alors le problème de sa résolution. Comme on l'a vu précédemment, la construction de systèmes informatiques réellement intelligents est encore loin de notre portée et la procédure classiquement utilisée en Intelligence Artificielle est de simuler l'intelligence. Afin d'apporter une assistance aux chercheurs qui réalisent une synthèse, il s'agit alors de simuler partiellement le processus «manuel», en proposant un support informatique à l'ensemble des activités impliquées.

La résolution d'un problème de synthèse induit tout d'abord d'apporter une assistance à la description du problème, soit de mettre en place un système de saisie de requête permettant la représentation non seulement des entités d'intérêt mais aussi de l'objectif poursuivi.

Le système doit ensuite être en mesure de construire un document de synthèse selon un processus qui doit refléter les opérations conduites par les chercheurs engagés dans une synthèse. Il s'agit alors de mener une combinaison d'activités, combinaison qui dépend du type de synthèse. Celle-ci peut être imaginée comme l'application d'un modèle de tâche générique, spécifique du type de synthèse considéré.

Cette combinaison d'activités est menée en prenant en compte les spécificités du chercheur et du thème étudié, orientée par la requête, afin de construire un document de synthèse. Ce document apporte une vue d'ensemble sur les documents ou les données et leur contexte.

La résolution du problème de synthèse telle qu'évoquée ici se rapproche beaucoup de la résolution d'un problème de Recherche d'Information, par son entrée par une requête, la sélection d'éléments pertinents d'un corpus, la construction de résultats en fonction de la requête. Mais la requête est une requête structurée décrivant une

tâche de synthèse spécialisée, le document de synthèse organisé dépasse la simple liste ordonnée par l'ajout de Visualisation d'Information à la Recherche d'Information classique et la sélection de documents pertinents n'est qu'une étape parmi d'autres de la résolution du problème de synthèse. Cette Recherche d'Information augmentée de ces divers concepts devient ce que j'appelle Recherche d'Information orientée tâche.

Mais une vue Recherche d'Information orientée tâche, si elle fournit des pistes pour aider à la résolution du problème de synthèse, n'est pas suffisante. En effet, le nombre de tâches de synthèse qui peuvent être considérées ainsi que la multiplicité des domaines d'application et la multitude de formes que peut prendre le corpus documentaire en font un champ trop vaste pour être traité de manière exhaustive dans le cadre d'une thèse.

Ancrer la question de synthèse dans le domaine applicatif, et dans le cas présent l'orienter selon la problématique d'appréhension des données Tissue MicroArrays, est la solution envisagée ici pour limiter l'espace d'interrogation. Les restrictions sur le problème de synthèse posées par le domaine des Tissus MicroArrays sont analysées plus en détails dans les prochains paragraphes.

3.2.3 Une approche ancrée dans la problématique expérimentale

3.2.3.1 Introduction

Ainsi qu'il vient d'être exposé, le problème de synthèse, si considéré dans un cadre générique, est très complexe, selon divers points de vue.

Tout d'abord, la notion de synthèse est fédératrice de nombreuses tâches très diverses. Par exemple, on peut considérer une revue contradictoire et exhaustive d'une thématique, la progression historique d'une idée, la répartition géographique d'un problème, la comparaison de plusieurs ensembles, etc.

Ensuite, la résolution informatisée de chacune de ces tâches implique de reproduire un processus manuel spécifique. Même si certaines activités sont identiques d'une tâche à l'autre, elles ne le sont pas toutes : une répartition géographique implique une classification par exemple par continent, puis pays, puis région, alors qu'une étude historique induit la construction d'une chronologie. Le champ des activités, qui peuvent être interprétées comme des problématiques à résoudre, est donc de plus en plus grand, au fur et à mesure qu'on multiplie le nombre de tâches de synthèse.

De plus, l'espace documentaire présente une variété très importante. Première-

ment, tout jeu de documents s'inscrit dans un domaine applicatif particulier. Or on peut envisager pléthore de domaines applicatifs, de la littérature à la mécanique quantique, de la cuisine à l'oncologie, etc. Pour chacun de ces domaines, nombre de corpus documentaires peuvent être envisagés. Ainsi, dans le cas de la recherche en oncologie, on peut considérer tous les articles recensés par PubMed[1] ou uniquement ceux publiés par une liste prédéfinie de revues de références ; parallèlement on peut prendre en compte uniquement les données acquises au sein d'un laboratoire, ou les associer à des données publiées par d'autres équipes sur Internet. Se pose alors un problème de nature des informations, du texte libre des articles à des données numériques en passant par des documents multimédia.

La résolution du problème de synthèse dans un contexte générique est donc difficilement envisageable, et la solution proposée est de limiter l'espace d'interrogation. Cette limitation de point de vue est ici dirigée par la problématique qui a été à l'origine de la définition du problème de synthèse : l'appréhension de données TMA, en tant qu'étape préalable à une fouille de données et en tant qu'outil permettant de replacer les données acquises par une technologie à haut débit dans une démarche expérimentale classique.

Ancrer la problématique de synthèse dans le contexte des Tissue MicroArrays permet une réduction du champ d'investigation, aussi bien en ce qui concerne la liste des tâches de synthèse, les activités à mettre en œuvre pour mener les tâches de synthèse à bien, du fait de problématiques particulières à résoudre, et l'espace documentaire à considérer ainsi qu'il sera présenté par la suite.

3.2.3.2 Des tâches spécifiques

Placer la problématique de synthèse dans le contexte applicatif des données Tissue MicroArrays implique tout d'abord de s'intéresser aux objectifs poursuivis par les chercheurs qui ont recours à la technologie, dans un cadre de recherche en oncologie. Ces objectifs peuvent être envisagés selon plusieurs perspectives.

Un premier point de vue est orienté par les éléments d'intérêt de l'étude et consiste en une transposition dans un contexte d'exploitation de données des types de conceptions de blocs TMA évoqués par [Kajdacsy-Balla et al., 2007]. Dans ces travaux, plusieurs types de plan de construction de TMA sont proposés :

- ⋆ «Outcomes-based TMA» : il s'agit de regrouper les individus ayant le plus d'informations de suivi dans leurs dossiers cliniques, afin d'évaluer les molécules en tant que marqueurs de pronostic,
- ⋆ «Progession-based TMA» : l'objectif ici est de montrer comment évoluent les tissus au cours du processus de transformation tumorale, en regroupant

[1] http ://www.ncbi.nlm.nih.gov/sites/entrez ?db=pubmed

des échantillons de tissus prélevés chez des donneurs sains, des tissus jugés sains chez des patients atteints de cancer, des tissus pré-cancéreux, des tissus tumoraux à divers stades, des tissus issus de récidive et métastases,

- ⋆ «Tumor-grade TMA» : cette configuration vise à présenter les divers stades du cancer,
- ⋆ «Tumor heterogeneity TMA» : avec des échantillons prélevés du cœur de la tumeur au tissu sain périphérique en passant par la frontière entre les deux, ce plan propose une étude des phénomènes d'infiltration,
- ⋆ «Consecutive cases TMA» : plan le plus courant dans la littérature, il consiste à regrouper l'ensemble du matériel biologique disponible,
- ⋆ «Specialty TMA» : ce terme regroupe l'ensemble des TMA construits dans un but précis autre que ceux exposés avant.

Bien que présentant l'avantage de décomposer la problématique d'exploitation des données TMA en plusieurs types d'études ou tâches, le recours à une telle organisation pose un certain nombre de problèmes.

Tout d'abord, la spécificité vis à vis du domaine applicatif est extrêmement forte : ces tâches ne peuvent aucunement être transposées hors du champ de la recherche en oncologie. De plus, l'existence du modèle «Specialty TMA» est problématique, puisqu'il couvre une multitude de tâches diverses qui ne sont pas recensées ni explicitées. Enfin, si ces plans suggèrent des critères de sélections d'individus et de présentation compacte des données sous forme de grille comme au sein d'un bloc TMA, ils fournissent peu de direction quant aux activités d'extraction d'informations, d'organisation conceptuelle ou d'organisation structurelle.

Il semble donc pertinent d'analyser les objectifs de la recherche en oncologie en conjonction avec les objectifs définis de l'appréhension des données TMA, c'est-à-dire un préalable à la fouille de données et une méthode pour se replacer dans une démarche expérimentale classique.

Or, une exploration préalable à une fouille de données doit permettre d'appréhender la structure de l'espace informationnel. Conjointement, la recherche en oncologie a été introduite comme visant à proposer des tests de dépistage (c'est-à-dire des éléments de diagnostic), des protocoles thérapeutiques (qui passent par une étude d'éléments de pronostic), en étudiant les mécanismes de transformation tumorale (soit une dynamique). Ces thèmes sont alors à envisager dans une perspective de tests d'hypothèses afin de permettre de se replacer dans le cadre d'une démarche expérimentale classique.

Des tâches permettant de rendre compte de ces problématiques de structure, pour préparer la fouille, et de diagnostic, pronostic et dynamique, pour se replacer dans une démarche expérimentale classique, doivent donc être mises en place.

En ce qui concerne la structure de l'espace, celle-ci doit permettre d'appréhender

la répartition des individus en divers groupes. Basée sur la valeur d'une variable, il s'agit d'une distribution ; construite automatiquement par des algorithmes de classification, elle peut être assimilée à la comparaison entre les groupes construits par le classifieur.

Pour le pronostic, il s'agit par exemple de montrer une corrélation entre le devenir des patients et des quantifications de marquage pour des molécules particulières, ce qui peut être envisagé par exemple sous la forme de l'évolution de l'espérance de vie en fonction d'un pourcentage de cellules marquées en une molécule.

Au niveau diagnostic, il peut s'agir de trouver des molécules dont le marquage est significativement différent entre tissus normaux et tissus tumoraux. Une telle étude peut être réalisée en comparant des pourcentages de cellules marquées entre divers types de tissus.

Enfin, pour montrer la dynamique de la transformation tumorale, alors que ne sont disponibles que des images à l'instant de l'ablation de la lésion pour chaque patient, la dimension temporelle se trouve remplacée par la notion de stade du cancer et une chronologie devient une construction multi-patients. La dynamique est alors illustrée par une comparaison entre groupes de patients à divers stades ou par l'évolution d'une mesure en fonction du stade.

Dans ce contexte, trois grandes catégories de tâches ont donc été envisagées, en collaboration avec les futurs utilisateurs du système, c'est-à-dire des biologistes et médecins, et dans un souci de généralisation à d'autres domaines applicatifs : comparaison, évolution et distribution.

3.2.3.3 Des problématiques spécifiques à résoudre

Ayant posé une définition succincte des tâches de synthèse considérée dans le cadre de l'exploitation des données TMA, il s'agit alors de déterminer plus précisément quels groupes d'activités sont impliqués dans ce contexte particulier et quelles problématiques spécifiques doivent être résolues dans le cadre de ces activités.

De manière générale, la synthèse a été présentée comme la combinaison d'activités de formulation d'une étude à réaliser, de sélection d'entités pertinentes, d'extraction d'informations intéressantes à partir des entités pertinentes, d'organisation conceptuelle des informations, d'organisation structurelle reflétant l'organisation conceptuelle, de présentation d'un document de synthèse, le tout réalisé en prenant en compte des connaissances du domaine applicatif et en prenant en compte la qualité des éléments produits. Dans le cadre des tâches de comparaison, évolution et distribution envisagées dans le contexte de l'appréhension des données TMA, ces diverses activités peuvent être décrites de manière un peu plus précise.

Ainsi, la sélection d'entités pertinentes implique le choix de patients intéressants dans le contexte de l'étude à réaliser, sur la base de leurs dossiers cliniques et des données histologiques disponibles. L'extraction d'informations reviendrait alors à choisir, au sein des dossiers cliniques et des données histologiques, les variables d'intérêt, comme un stade, un pourcentage de cellules marquées, etc. De manière générique, ces deux types d'activités peuvent être considérées comme des activités de sélection.

Ensuite, l'organisation conceptuelle doit refléter la problématique centrale de chacune des tâches. Ainsi, la comparaison induit la construction d'une hiérarchie de groupes d'individus, guidée par des algorithmes de classification ou des valeurs de variables. L'évolution sous-tend la définition d'un ou des individus représentatifs de chaque combinaison de valeurs pour deux variables. La distribution requiert le groupement d'individus selon la valeur d'une variable. Refléter structurellement ces organisations conceptuelles induit la définition d'une distance, et le placement des groupes et des individus de proche en proche selon cette distance. Intimement liées ces deux activités d'organisation conceptuelle puis structurelle peuvent être envisagées conjointement en tant qu'activités d'organisation.

Ensuite, la présentation du document de synthèse, qui est construit dans un but exploratoire, ne se limite pas à l'affichage d'une structure compacte issue de l'organisation des éléments, mais doit aussi permettre l'accès au contexte, soit aux dossiers cliniques complets et informations histologiques, dont les images de lames et spots. Ces activités d'affichage tout à la fois d'une vue de synthèse et de vues annexes sur les données constituent des activités de présentation.

Enfin, l'ensemble sous-tend des considérations qualité. Comme il a été exposé dans le Chapitre 1, l'acquisition de données expérimentales n'est jamais exempte d'erreurs ni d'approximations. Conjointement, les dossiers cliniques, et en particulier les informations de suivi, sont rarement complets : les archives papier n'ont pas été complètement informatisées, le patient a été perdu de vue par l'hôpital suite à un déménagement, etc. Cet état de fait implique l'intégration d'une démarche qualité au sein de chacune des trois grandes classes d'activités considérées : sélection, organisation et présentation.

La sélection est l'activité où la majorité des problématiques qualité interviennent. Les problèmes de données impliquent en effet la gestion des données manquantes, que ce soit en excluant de la sélection les patients dont les dossiers cliniques et informations histologiques sont incomplets, ou en inférant ces données à partir d'individus similaires. De plus, les données histologiques sont acquises à partir d'échantillons biologiques stockés en blocs de paraffine suite à un ensemble de traitements. Or, il existe d'éventuelles incidences du traitement paraffine sur l'immunomarquage. Ceci induit une prise en compte d'une validité spatio-temporelle de l'utilisation conjointe des échantillons. Au cours du temps, et d'une institution à l'autre, divers protocoles d'inclusion dans la paraffine, conduisant à des marquages variables, ont en effet été

utilisés.

Au niveau organisation, ce sont les limites posées par l'espace dans lequel peuvent être présentées les informations, au sein du document de synthèse, qui doivent être prises en compte. En effet, l'objectif est la construction d'une visualisation compacte, et l'inadéquation entre l'espace disponible et le volume d'informations sélectionnées peut induire soit l'exclusion, soit l'inclusion d'objets, selon le même type de critères qualité que la sélection.

Ainsi, chacune des tâches de synthèse envisagées peut être décomposée en trois grands ensembles de problèmes à résoudre, sélection, organisation et présentation, qui incluent entre autres des problématiques qualité.

3.2.3.4 Un espace documentaire particulier

Une autre spécificité du domaine TMA vis à vis de la synthèse est l'espace documentaire à considérer. En effet, envisagé dans un cadre générique, cet espace peut prendre différentes formes : des textes, des représentations de documents, des données, des documents multimédia, etc.

Cette diversité pose problème dans la résolution, chaque activité devant être déclinée de manière spécifique pour chaque forme de document. Par exemple, la sélection sur du texte libre induit des techniques de Recherche d'Information, alors que sur des éléments à la granularité très fine comme des items stockés en base de données, de simples systèmes de correspondance exacte, comme des requêtes SQL de type SELECT, suffisent.

Le contexte TMA induit comme sources d'informations les dossiers cliniques des patients et des données histologiques telles que des mesures de marquages et des informations concernant blocs, lames et spots, dont des images.

Cet espace documentaire est hétérogène, puisqu'il inclut des formes de documents variées : du texte libre (comptes-rendus des médecins dans les dossiers par exemple), des mots (un diagnostic par exemple), des valeurs numériques (une mesure de marquage), des éléments multimédia (images de spots). Il est aussi structuré à diverses échelles, du patient considéré dans son ensemble au niveau intracellulaire. Cette hétérogénéité, associée aux relations intriquées existant entre les divers éléments, est source de complexité.

Mais cette complexité peut être limitée. En particulier, l'existence d'une structure commune à tous les dossiers cliniques facilite l'appréhension de l'espace documentaire.

De plus, les éléments les plus complexes sont les textes et les images. Or, les informations essentielles sont en général représentées sous une forme autre : le texte n'intervient qu'en support ou explicitation de valeurs quantitatives ou qualitatives et les images sont associées à des annotations sous forme de mots clés. Textes et images en tant que tels peuvent donc être laissés de côté dans un premier temps, ne gardant que des données brutes.

3.2.3.5 Une tâche particulière en support à la conception

Se focaliser sur le contexte TMA pour appréhender le concept de Recherche d'Information orientée tâche a permis de limiter le champ d'investigation du problème. Ainsi, les tâches de synthèse à considérer sont des tâches de synthèse de données, qui sont envisagées comme relevant de trois catégories : comparaison, évolution et distribution. Ensuite, ces tâches sous-tendent des activités de sélection, organisation et présentation, qui incluent des problématiques qualité. Enfin, le corpus documentaire, bien que complexe, est bien structuré et laisse peu de place au texte libre.

Mais cette vision reste encore trop générale pour permettre une appréhension fine de la problématique de synthèse dans le cadre applicatif des TMA. Une méthode courante dans ces circonstances est de se baser sur un exemple, qui guide à l'arrière plan les réflexions. C'est la démarche que j'adopte ici.

Comme indiqué précédemment, la technologie des TMA se veut un outil de recherche en oncologie. Parmi les objectifs de cette recherche, on peut citer les problématiques de diagnostic, soit entre autres de détermination de molécules dont l'expression est significativement différente entre tissus sains et tissus pathologiques. C'est une problématique de ce type qui va être envisagée comme support à la suite de la réflexion.

Parmi les problèmes courants rencontrés en routine par les anatomopathologistes, on peut citer l'observation des pièces opératoires pour déterminer si toute la tumeur a bien été retirée. Se pose alors la question de la frontière entre la tumeur et le reste de l'organe, qui n'est pas forcément très nette. Cette question induit d'étudier s'il y a des différences significatives entre tissu tumoral et tissu adjacent à la tumeur. Ces différences peuvent être évaluées par comparaison entre tissus tumoraux et adjacents pour un groupe de patients homogène, par exemple ayant la même pathologie de cancer du côlon. L'élément à comparer, puisque la problématique est envisagée au niveau tissulaire, peut être le pourcentage de cellules marquées pour diverses molécules impliquées dans la transformation tumorale. Une prise en compte du compartiment cellulaire dans lequel est situé le marquage peut permettre une comparaison plus fine.

L'étude considérée peut alors être exprimée sous la forme : «comparaison du

pourcentage de cellules marquées entre les différentes molécules étudiées, entre tissus tumoraux et adjacents à la tumeur, en fonction de la localisation intracellulaire du marquage, chez les patients atteints d'un cancer du côlon».

La Fig. 3.3 illustre le type de résultat théorique qui pourrait être envisagé pour cet exemple. La comparaison entre les différentes molécules étudiées conduit à la constitution de quatre groupes, un par marqueur. Ensuite, la comparaison en fonction de la localisation du tissu par rapport à la tumeur induit le découpage de chaque groupe en deux sous-groupes, un pour le tissu adjacent à la tumeur et l'autre pour le tissu tumoral. Puis la comparaison en fonction de la localisation intracellulaire du marquage induit le découpage de chaque sous-groupe en trois ensembles, un par compartiment intracellulaire : membrane, cytoplasme, noyau. Enfin, les individus, au sein de chaque groupe de niveau le plus fin, présentent les mesures de pourcentage de cellules marquées correspondants.

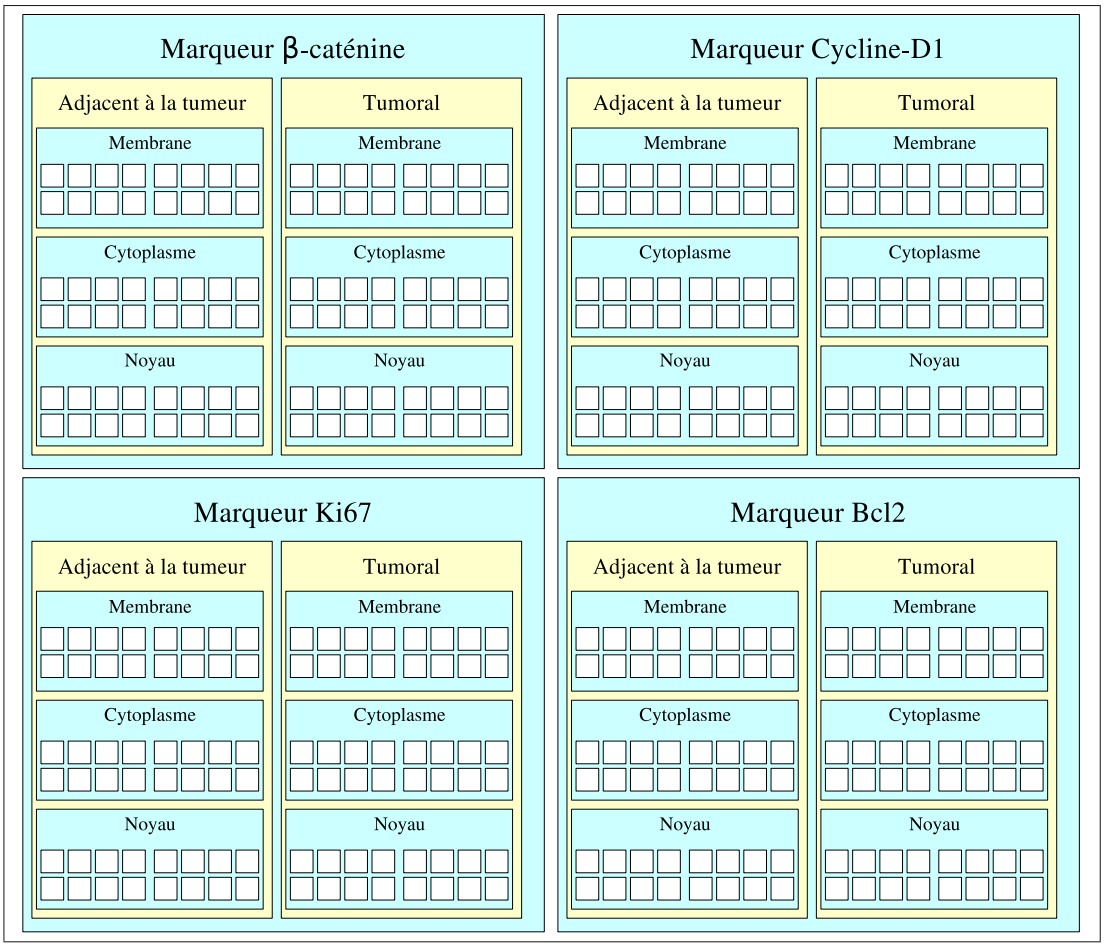


Fig. 3.3: Résultat théorique de l'exemple de comparaison - Les diverses axes de comparaison induisent la constitution de groupes imbriqués dans lesquels les individus (les petits carrés) présentent leur pourcentage de cellules marquées, objectif de la comparaison.

Cette base applicative étant posée, il s'agit alors d'évaluer comment l'angle Recherche d'Information orientée tâche envisagé pour la notion de synthèse s'ar-

ticule avec les divers modèles de Recherche d'Information existants, objectif de la prochaine section.

3.3 Modèle de synthèse

3.3.1 Introduction

Un nouveau paradigme de Recherche d'Information adapté à la synthèse, la Recherche d'Information orientée tâche, a été défini. Il faut alors le replacer dans le contexte des divers modèles existants. L'objectif n'est pas de définir un nouveau modèle de Recherche d'Information mais de replacer dans ce cadre mon point de vue sur la synthèse. La représentation adoptée ne se veut pas non plus un nouveau modèle mais plutôt une version simplifiée des divers modèles existants.

Les sciences de l'information visent à analyser le phénomène d'accès à l'information et s'intéressent à l'interaction entre l'homme et des sources d'information. Le processus de recherche peut alors être envisagé comme s'organisant donc autour d'un triptyque *Utilisateur* $\leftrightarrow$ *Problème* $\leftrightarrow$ *Information*.

Selon ce processus, l'utilisateur commence par formuler un problème informationnel sous la forme d'une requête. Cette formulation pose la question de l'adéquation de l'image mentale du problème avec son expression au sein de la requête. En fonction de celle-ci, l'information pertinente selon les critères du système est extraite du corpus informationnel et ce résultat est présenté à l'utilisateur. Celui-ci évalue alors la pertinence des informations transmises en fonction de ses besoins. L'ensemble est un processus dynamique, réitéré plusieurs fois, avec une dépendance forte entre épisodes de recherche.

Le modèle générique correspondant, présenté Fig. 3.4, inclut trois types de composantes : entités, interactions et évaluations.

Les entités consistent en les éléments du triptyque qui sont au cœur du modèle. L'utilisateur représente l'individu engagé dans le processus de recherche, le problème sa motivation pour la recherche et l'information la base sur laquelle est réalisée la recherche.

Les interactions lient les entités au sein du processus. Ainsi, la formulation consiste en la représentation du problème par l'utilisateur, la recherche en la mise en adéquation du problème avec l'information et le résultat en la présentation à l'utilisateur d'une sélection pertinente réalisée à partir de l'information.

Chacune des interactions pose un problème d'évaluation qualité. La formulation

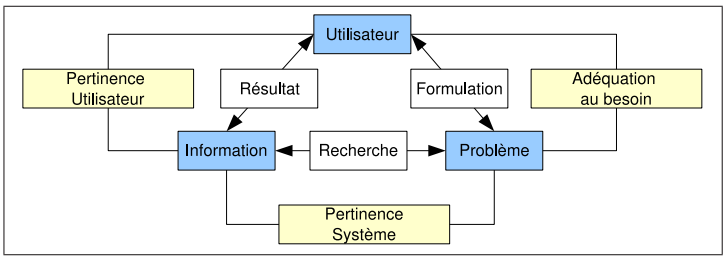


Fig. 3.4: Modèle générique de Recherche d'Information - Le processus de Recherche d'Information résulte des interactions entre trois entités : utilisateur, problème et information. Il implique diverses interactions : formulation de son problème par l'utilisateur sous forme de requête, recherche d'une correspondance entre problème et information, puis présentation des résultats extraits du corpus informationnel à l'utilisateur. Ces activités posent des problèmes d'évaluation qualité : adéquation au problème de la formulation de la requête par l'utilisateur, pertinence système des documents retournés jugée par les algorithmes de recherche, pertinence utilisateur évaluée par l'usager confronté aux résultats du système.

implique de s'intéresser à l'adéquation au besoin entre utilisateur et problème, la recherche sous-tend une notion de pertinence système entre problème et information, et le résultat induit une pertinence utilisateur entre l'usager et l'information.

Ce modèle peut être lu à différents niveaux d'interprétation et le fossé entre les diverses représentations du problème d'accès à l'information est large. De nombreux travaux visent à combler cette distance selon l'un ou l'autre des éléments du triptyque, mais la prise en compte des facteurs humains (cognitifs, comportementaux, sociaux...) au sein des systèmes de Recherche d'Information actuels est un problème complexe.

L'état de l'art du domaine des sciences de l'information conduit Section 2.3 permet toutefois de concevoir, sur une échelle du plus technique, algorithmique, proche du système, au plus abstrait, proche des problématiques cognitives, un ensemble de modèles de Recherche d'Information : modèles de Recherche d'Information classique ou modèles comportementalistes. Le modèle de synthèse, orienté tâche, se place à un niveau intermédiaire. Les prochains paragraphes vont permettre une exploration rapide de ces différents modèles.

3.3.2 Modèle de Recherche d'Information classique

Comme exposé Paragraphe 2.3.2.3, l'opérationnalisation du processus de Recherche d'Information par des systèmes informatiques a été basée sur un modèle de Recherche d'Information classique très simple, présenté Fig. 3.5.

Dans le cadre de ce modèle, le problème ou les besoins de l'utilisateur se limitent à des mots clés, éventuellement issus d'un vocabulaire contrôlé tel qu'une ontologie. Ces mots clés sont formalisés sous forme de liste au sein de la requête.

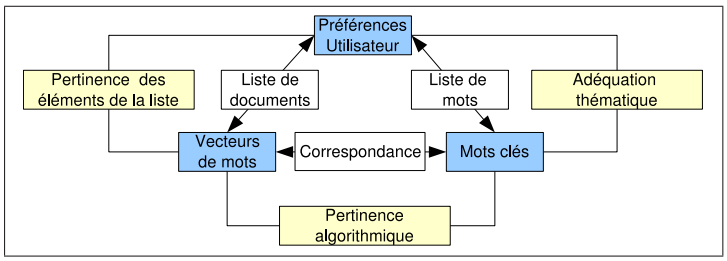


Fig. 3.5: Modèle de Recherche d'Information classique - Il consiste en une simple correspondance entre une requête sous forme de mots clés et des représentations de documents sous forme de vecteurs de mots, permettant la construction d'une liste de documents algorithmiquement pertinents, où l'utilisateur en tant que tel est absent ou représenté par des préférences.

L'information servant de base à la recherche est constituée par un corpus documentaire, en général représenté sous forme d'un ensemble de vecteurs de mots. Ces représentations de documents sont contruites lors d'une phase préalable d'indexation, généralement par extraction de termes signifiants (c'est-à-dire hors articles, déterminants, conjonctions, etc.) et lemmatisation (soit réduction du mot à sa forme canonique).

Les algorithmes de recherche consistent alors en une simple correspondance entre la liste de mots de la requête et les vecteurs de mots de la collection de documents. Ces algorithmes, basés sur des modèles booléens, vectoriels, probabilistes ou autres permettent la construction d'une liste de documents pertinents. Cette liste est éventuellement ordonnée en fonction d'une mesure de pertinence attribuée par le système à chaque document.

La notion d'utilisateur est en général absente ou limitée à des préférences ou un profil simples. Au sein de ce modèle, la concordance entre la requête et le problème de l'utilisateur devient une simple adéquation thématique. En effet, l'expression d'un besoin sous forme de simples mots ne permet que de cerner le sujet qui est d'intérêt pour l'utilisateur, sans manifestation d'intention.

La correspondance entre le problème utilisateur et le corpus documentaire est évaluée selon une pertinence purement algorithmique : les mots de la requête apparaissent dans les documents du corpus avec une fréquence statistique plus ou moins élevée.

Le jugement des résultats de la recherche par l'utilisateur reste une pertinence des éléments d'une liste. Cette pertinence est envisagée en général par les concepteurs des systèmes de manière individuelle pour chaque document. Or, lorsqu'il juge un document, l'utilisateur est en général influencé par les documents qu'il a déjà consulté. Mais cette notion de dépendance entre évaluations successives par un utilisateur est peu prise en compte.

Bien que très simple, ce modèle a permis le développement de nombreux systèmes

de Recherche d'Information opérationnels aux résultats satisfaisants. Mais il présente un certain nombre de limites, liées à cette simplicité même. En particulier, la prise en compte de l'utilisateur en tant qu'entité pensante, placé dans un contexte socio-économique et affectif, juge final de la qualité des résultats de la recherche, reste très succincte. Cette dimension est par contre au cœur des modèles comportementalistes.

3.3.3 Modèle de Recherche d'Information comportementaliste

Comme présenté Paragraphe 2.3.2.2, il existe pléthore de modèles comportementalistes pour le processus de Recherche d'Information. Bien que très variés, ils peuvent être envisagés comme rentrant dans le cadre du modèle simplifié présenté Fig. 3.6.

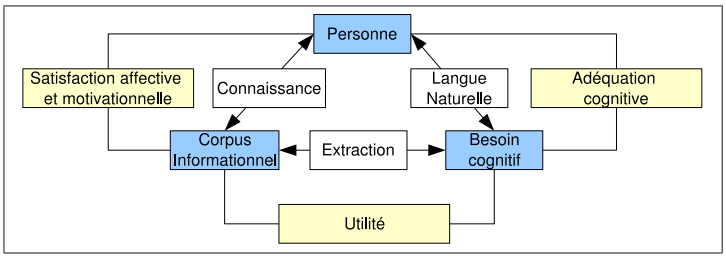


Fig. 3.6: Modèle de Recherche d'Information comportementaliste - Il donne une place prépondérante à l'utilisateur en tant que personne et se place à un haut niveau d'abstraction cognitive.

Ce modèle donne une place prépondérante à l'utilisateur, placé dans un contexte socio-économique, cognitif, affectif, qui n'est plus limité à sa représentation au sein d'un système informatique par un profil, mais vraiment considéré en tant que personne, dans toute la complexité de l'humain.

Cette personne est placée face à un besoin cognitif, c'est-à-dire une inadéquation entre sa représentation intellectuelle du monde et sa perception physique du monde, qui induit une nécessité de compléter ses connaissances afin de modifier sa représentation du monde en conséquence. Ce besoin, qualifié de cognitif, n'est pas forcément purement intellectuel et des contraintes externes (socio-économiques par exemple) ou une charge émotionnelle peuvent intervenir.

Ce besoin cognitif est exprimé en langue naturelle et confronté à un corpus informationnel, dont la forme peut être très variée, de documents en texte libre à des images et vidéos en passant par d'autres personnes. Un processus d'extraction permet la construction éventuelle de connaissances nouvelles. Cette extraction et intégration aux connaissances de la personne ne sont en effet pas systématiques : des contingences émotionnelles, religieuses, sociales, peuvent induire un rejet des informations extraites.

L'expression du besoin en langue naturelle pose un problème d'adéquation cognitive entre les deux. En effet, la formulation de la perception d'un besoin est en général un problème difficile lié à la distance entre la pensée et le langage.

La correspondance entre besoin et information est évaluée en termes d'utilité. Ici, l'utilité est considérée en terme d'adéquation des connaissances extraites à l'activité en cours et en terme de nouveauté par rapport à ce qui est connu ou déjà proposé par le système. Les connaissances extraites se doivent alors de mener à la satisfaction affective et motivationnelle de l'utilisateur.

Ce modèle très abstrait est proche du processus réel de Recherche d'Information tel que perçu par l'usager et permet ainsi une meilleure compréhension des mécanismes intellectuels en jeu. Mais ce haut niveau d'abstraction rend difficile la construction de systèmes de Recherche d'Information, du fait de la complexité des opérations et des entités impliquées. En définitive, il se heurte aux limites de l'Intelligence Artificielle, qui ne sait actuellement que contrefaire les raisonnements humains et non les reproduire.

3.3.4 Modèle de Synthèse

La synthèse est envisagée dans mes travaux selon un point de vue Recherche d'Information orientée tâche, où cette dernière notion prend une place centrale. Ce concept de tâche est envisagé selon les assertions de comparaison, évolution et distribution d'une synthèse basée sur des données, dans un domaine applicatif centré sur la recherche expérimentale.

La notion de tâche apporte à la définition du problème informationnel de l'utilisateur des dimensions supplémentaires par rapport aux simples mots clés du modèle de Recherche d'Information classique. Mais elle reste plus simple et formellement exprimable que le concept de besoin cognitif du modèle de Recherche d'Information comportementaliste.

Dans la même veine, les différentes composantes du modèle de synthèse, présenté Fig. 3.7, sont à mi-chemin entre leurs représentations dans le modèle classique et dans le modèle comportementaliste.

Ainsi, l'usager est envisagé comme un archétype utilisateur spécialisé par des préférences. Le besoin est représenté par une tâche spécialisée par une requête structurée. Le processus de synthèse est appliqué sur des documents structurés pour construire un document synthétique organisé. La formulation de la requête pose alors un problème d'adéquation à la tâche dans laquelle l'utilisateur est engagé. Le processus de synthèse pose un problème de pertinence situationnelle et le document de synthèse est jugé par l'utilisateur sur la base d'une pertinence interprétationnelle

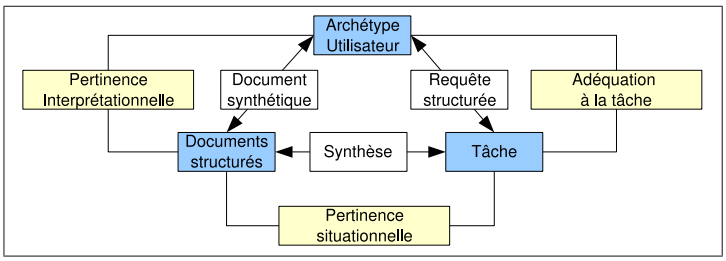


Fig. 3.7: Modèle de Synthèse - Il est conçu comme un intermédiaire entre le modèle classique, très algorithmique, et le modèle comportementaliste, abstrait et proche de l'utilisateur et de ses processus cognitifs.

pour l'étude qu'il cherche à mener.

Ce modèle doit permettre la construction d'un système de Recherche d'Information orientée tâche, ou système d'assistance à la synthèse, qui soit opérationnel. Une telle opérationnalisation sous-tend une analyse plus poussée des différentes composantes du modèle, analyse menée dans la prochaine section.

3.4 Composantes du modèle

3.4.1 Introduction

Le modèle de synthèse, afin d'être utilisé comme base conceptuelle pour la mise en place d'un système effectif de Recherche d'Information orientée tâche, doit être évalué plus précisément. La Fig. 3.8 en présente une vision plus détaillée.

Ainsi qu'il a été décrit précédemment, ce modèle consiste en une incarnation particulière, construite en gardant à l'esprit le contexte applicatif des Tissue MicroArrays, du triptyque *Utilisateur* $\leftrightarrow$ *Problème* $\leftrightarrow$ *Information* qui est au cœur de la représentation du processus de Recherche d'Information adoptée.

Ces trois entités sont liées par des interactions : requête structurée, synthèse et document de synthèse. Ces interactions sont sujettes à des problématiques d'évaluation : adéquation à la tâche, pertinence situationnelle et pertinence interprétationnelle.

L'objectif des prochains paragraphes est une description plus précise de ces diverses composantes du modèle, organisées selon les trois types d'éléments impliqués : entités, interactions et évaluations.

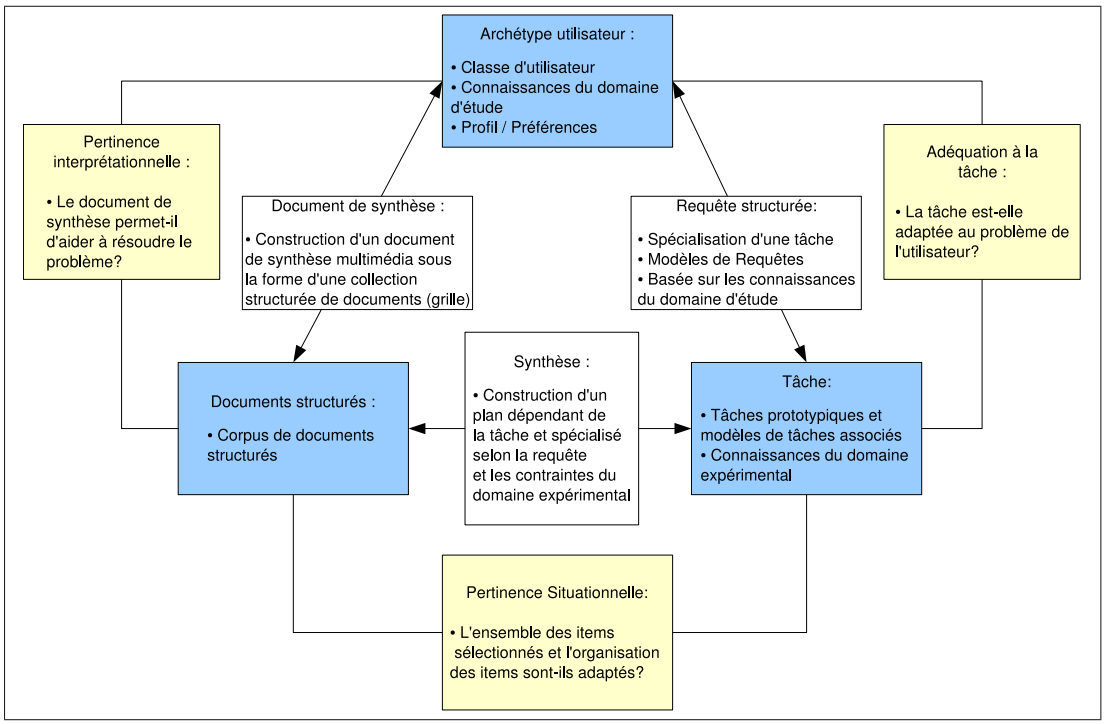


Fig. 3.8: Modèle de Synthèse détaillé - Les entités, interactions et évaluations sont présentées dans leurs incarnations spécifiques au modèle de synthèse.

3.4.2 Entités

3.4.2.1 Tâche

3.4.2.1.1 Introduction

Comme mentionné précédemment, le modèle de synthèse envisagé propose d'aller plus loin qu'une représentation par mots clés des besoins ou problèmes de l'utilisateur, en les formulant à un niveau d'abstraction supérieur, le niveau tâche, sans pour autant espérer atteindre une représentation parfaite d'un besoin cognitif. Cette notion de tâche est donc centrale au sein du modèle de synthèse proposé.

La synthèse a été montrée comme une notion fédératrice pour une multitude de tâches, et une exploration plus poussée de cette notion implique tout d'abord de s'intéresser à l'étendue du champ couvert par les tâches de synthèse, avant de considérer la représentation d'une tâche individuelle et sa résolution.

3.4.2.1.2 Taxonomie de tâches

Des taxonomies générales de tâches ont été proposées en relation avec les études menées d'un point de vue comportementaliste du processus de Recherche d'Information, par exemple par [Choo et al., 1999] ou [Wilson, 1999]. Bien que ces taxonomies fournissent un aperçu de la multiplicité des tâches à considérer dans le cadre générique de la Recherche d'Information, elles restent à un niveau trop général : si elles incluent une notion de synthèse, elles ne proposent pas de décomposition de cette notion. De plus, elles suggèrent peu de pistes pour rendre l'expression d'une tâche opérationnelle.

Or, le Paragraphe 3.2.2.2 a permis d'établir la multiplicité des tâches fédérées par la notion de synthèse. Dans le contexte applicatif des TMA, Paragraphe 3.2.3.2, trois catégories de tâches de synthèse de données ont été dégagées : comparaison, évolution et distribution. En considérant ces catégories de manière plus fine, un ensemble de tâches générales, qui seront nommées par la suite «tâches prototypiques», a été dégagé. Ces tâche prototypiques correspondent aux différents formats de lentilles de la métaphore optique de la Fig. 3.2, ou aux différents types d'études qu'un chercheur voudrait réaliser en utilisant les données TMA.

Une taxonomie non exhaustive de ces tâches prototypiques a été construite et elle est présentée Fig. 3.9. Au premier niveau de la hiérarchie, se retrouvent les catégories de tâche évoquées précédemment : comparaison, évolution, distribution.

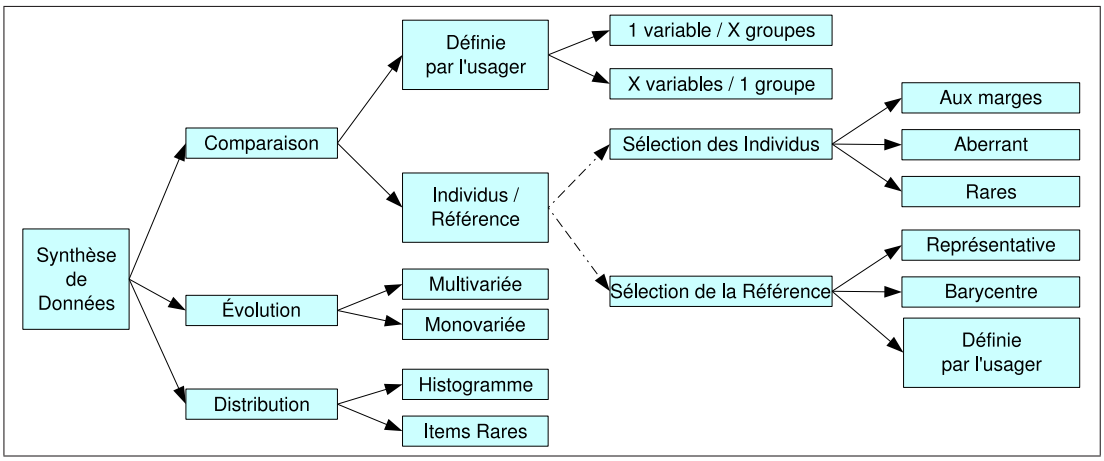


Fig. 3.9: Vue partielle de la taxonomie des tâches prototypiques - Les tâches de synthèse basées sur des collections de données, telles que celles contruites dans le contexte des TMA, sont envisagées comme relevant de trois catégories, elles-mêmes subdivisables en catégories plus fines : les «comparaisons» permettent d'apprécier la structure de l'espace informationnel étudié, les «évolutions» fournissent un aperçu sur les corrélations entre éléments et les «distributions» donnent accès à la répartition des individus.

En ce qui concerne les comparaisons, deux types peuvent être envisagés. Les critères de comparaison peuvent être définis par l'usager et permettre la comparai-

son de la valeur d'une variable en divers groupes d'individus (comme dans l'exemple du Paragraphe 3.2.3.5), ou la comparaison de plusieurs variables au sein d'un même groupe (par exemple les différentes composantes T, N et M du stade pTNM pour les patients atteints d'un cancer du côlon). Il peut aussi s'agir de la comparaison d'individus particuliers par rapport à une référence. Ceci implique le choix du type d'individus, aux marges d'un groupe, aberrants ou rares, qui sont toujours très instructifs en tant que cas particuliers dans un contexte de recherche. En ce qui concerne la référence de la comparaison, celle-ci peut consister en un individu représentatif, le barycentre d'un groupe de référence ou un individu défini par l'usager.

Au niveau évolution, celle-ci peut être envisagée comme monovariée (par exemple l'évolution du nombre de ganglions envahis par le cancer en fonction du nombre de ganglions observés lors de l'ablation de la tumeur) ou multivariée (par exemple le pourcentage de cellules marquées et l'intensité du marquage en un marqueur donné, en fonction du stade de la maladie).

Enfin, au niveau distribution, on peut s'attacher à la structure globale de la population considérée (par exemple la distribution par classes d'âge) ou spécifiquement à des individus rares (par exemple la distribution par classe d'âge des individus ayant une expression d'une molécule bien inférieure aux autres).

Il s'agit alors de s'intéresser plus précisément aux activités spécifiques impliquées par chacune de ces tâches prototypiques.

3.4.2.1.3 Des tâches composées

Ainsi que présenté Paragraphe 3.2.2.3 les tâches de synthèse sont des tâches complexes et doivent faire l'objet d'une analyse plus poussée pour permettre leur représentation dans un objectif de résolution.

Dans le domaine applicatif des TMA, comme évoqué Paragraphe 3.2.3.3, elles supposent tout d'abord des activités de sélection d'entités pertinentes (les patients) et extraction d'information (des données particulières issues du dossier clinique ou des données histologiques), qui peuvent être regroupées sous une même thématique de sélection.

Elles nécessitent ensuite une organisation conceptuelle (hiérarchie, groupes, etc.) puis une disposition structurelle (juxtaposition d'ensembles, dispersion selon plusieurs dimensions, etc.) des éléments, soit des opérations d'organisation.

Puis elles requièrent une mise en forme du document final (vue synthétique et vues annexes), soit une activité de présentation. Chacun de ces groupes d'opérations inclut des activités liées à la qualité de la production.

Chaque tâche prototypique peut ainsi être modélisée comme la composition de trois sous-tâches principales (sélection, organisation et présentation), qui elles-mêmes sont des tâches complexes et peuvent être décomposées en arbres de sous-tâches. Ces arbres de sous-tâches définissent ce qui sera appelé par la suite le «modèle de tâche» correspondant à chaque tâche prototypique.

À des fins d'illustration, on peut considérer la problématique évoquée Paragraphe 3.2.3.5 : la «comparaison du pourcentage de cellules marquées entre les différentes molécules étudiées, entre tissus tumoraux et adjacents à la tumeur, en fonction de la localisation intracellulaire du marquage, chez les patients atteints d'un cancer du côlon».

Il s'agit d'une tâche de comparaison de la valeur d'une variable entre des groupes d'individus définis par l'utilisateur (tâche située tout en haut de la taxonomie de tâches de la Fig. 3.9). Une vue simplifiée du modèle de tâche correspondant à cette tâche, qui sera par la suite évoquée comme tâche de «comparaison» est présenté Fig. 3.10. Cet exemple de modèle de tâche montre bien la décomposition en trois sous-problèmes de sélection, organisation et présentation, qui a été introduit dans le domaine des données TMA au Paragraphe 3.2.3.3. Chacune de ces catégories de sous-tâches se décompose alors en problèmes de niveau plus fin.

Ainsi, dans le cas d'une comparaison, la sélection induit tout d'abord une sélection des éléments pertinents, qui est réalisée par une sous-tâche élémentaire de sélection selon des critères définis par l'utilisateur. Ensuite, intervient une sélection de groupes. Celle-ci implique tout d'abord la constitution des groupes, par construction de l'ensemble des groupes possibles selon des critères de groupement définis par l'utilisateur, puis l'attribution des éléments sélectionnés aux groupes. Cette sélection de groupes fait intervenir une sous-tâche qualité, par une notion de représentativité : les individus intégrés au groupe doivent être représentatifs du groupe. L'ensemble du processus lui aussi induit des problèmes de qualité et en particulier de validité. Les objets intégrés doivent être valides, ce qui nécessite une gestion des données manquantes. Une validité spatio-temporelle doit aussi être prise en compte, par intégration d'individus dont le matériel biologique est stocké de manière similaire, c'est-à-dire d'individus d'une même période temporelle et soignés dans des institutions usant des mêmes procédures.

Ensuite, intervient l'organisation des groupes et éléments. Celle-ci implique tout d'abord la définition d'une géométrie, réalisée sous forme d'une grille, choix qui sera exposé plus en détails dans le Paragraphe 3.4.3.3, dont la taille est définie au sein d'une première sous-tâche élémentaire. Ensuite, il s'agit de déterminer les zones affectées à chaque groupe au sein de la grille. Puis, au sein de chaque zone est définie la position de chaque élément, en fonction de critères d'ordonnancement définis par l'usager. Ce processus pose des problèmes qualité d'adéquation de l'ensemble des groupes et éléments sélectionnés à la grille. En cas de manque de cases pour placer tous les objets, il faut étendre la grille ou exclure des objets de la sélection. En cas

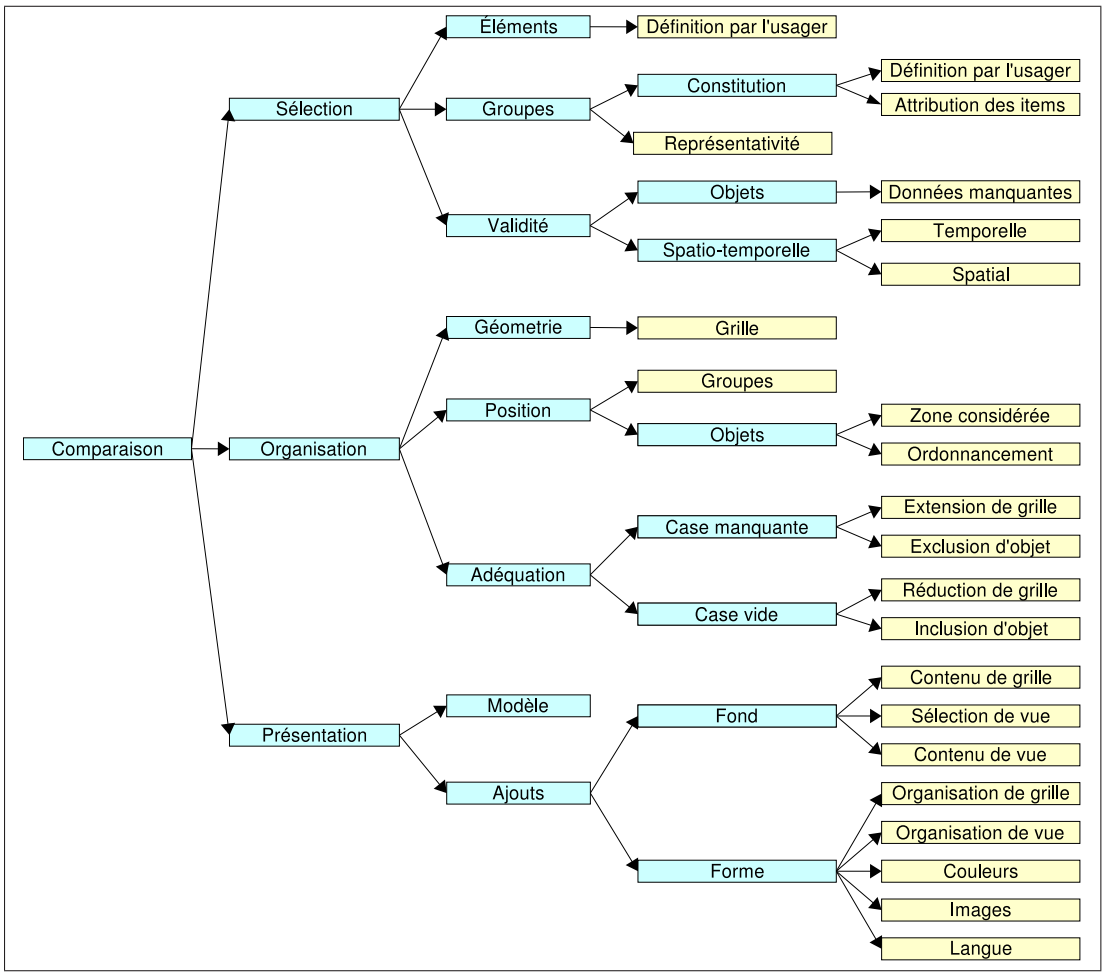


Fig. 3.10: Modèle de tâche simplifié - Une tâche de type «comparaison » est décomposée en un arbre de sous-tâches. Chaque rectangle nœud représente une catégorie de sous-tâches qui se décompose elle-même au niveau suivant, jusqu'à atteindre des nœuds feuille qui représentent des sous-tâches élémentaires.

de cases vides, on peut envisager une réduction de la taille de la grille ou l'inclusion d'objets supplémentaires.

Enfin, intervient la présentation du document de synthèse finale. Celle-ci a lieu par application d'un modèle, complémenté par des ajouts spécifiques. En ce qui concerne le fond, il s'agit des données affichées au sein de la grille, de la liste des vues complémentaires sur les données à proposer et du contenu de chaque vue. Au niveau forme, des sous-tâches élémentaires spécifiques peuvent diriger l'organisation des éléments présentés au sein de la grille, l'organisation des éléments au sein de chaque vue, les choix de couleurs, images affichées ou de langue.

3.4.2.2 Archétype Utilisateur

3.4.2.2.1 Introduction

De plus en plus, dans les modèles de Recherche d'Information, l'utilisateur n'est plus un élément fantôme qui n'existe que parce qu'il est la source de la formulation d'une requête. Il est considéré comme entité pensante individualisée, dans un contexte socio-économique, cognitif, affectif. Le mode de raisonnement qu'il emploie, son vocabulaire sont partiellement définis par sa formation, ses centres d'intérêt...

La prise en compte de cet utilisateur en tant qu'individu au sein du processus de synthèse implique de s'intéresser à sa représentation. Cette représentation, dans le cadre applicatif de l'appréhension des données TMA, implique de considérer les diverses catégories d'utilisateurs potentiels de l'application, des étudiants en biologie aux anatomopathologistes, en passant par les chercheurs en oncologie et les médecins.

La représentation de l'utilisateur passe en général par la définition d'un ensemble de modèles d'utilisateurs, la représentation de ces modèles et l'individualisation d'un modèle pour un utilisateur particulier, ainsi que décrit ici.

3.4.2.2.2 Notion d'archétype

Selon [Brusilovsky, 1996], un modèle utilisateur se doit d'inclure les connaissances de l'utilisateur, son but ou tâche, son expérience passée et ses préférences. Dans le cadre du modèle de synthèse, et contrairement aux systèmes pédagogiques dans lesquels les modèles utilisateur sont très utilisés, l'usager n'est pas caractérisé par un objectif d'apprentissage en tant que tel, puisqu'il va chercher à accomplir une tâche spécifique lors de chaque interaction avec le système. Cette notion de but étant indépendante du modèle utilisateur, se pose la question de la représentation des trois autres catégories d'informations : connaissances, expérience passée et préférences.

Comme il a été vu Paragraphe 2.4.2, la représentation de l'utilisateur est généralement tributaire de la modélisation d'un groupe d'individus similaires, modèle qui est ensuite personnalisé pour un individu particulier, ainsi que le note [Kobsa, 2001].

Selon ce mécanisme, dans le cadre du modèle de synthèse, est considérée une notion d'archétype utilisateur, qui représente une classe d'utilisateurs partageant des connaissances spécifiques sur le domaine applicatif et un mode de fonctionnement intellectuel similaire dans ce domaine. Ce mode de fonctionnement peut être assimilé à la notion d'expérience des modèles utilisateur classiques. Ces connaissances et modes de raisonnements particuliers constituent des modèles mentaux, dont l'existence a été étudiée chez les avocats dans [Sutton, 1994]. Ainsi, dans le domaine

d'étude des Tissue MicroArrays utilisés en oncologie, les différents archétypes utilisateur recouvrent des catégories d'usagers de type : technicien de laboratoire, étudiant en biologie de 1^{er} cycle, chercheur en biologie, anatomopathologiste, etc.

3.4.2.2.3 Représentation des archétypes

Ayant défini une liste d'archétypes utilisateur à prendre en compte, il s'agit alors d'évaluer comment les représenter. D'après l'assertion d'archétype envisagée, cette représentation inclut connaissances sur le domaine d'étude et modes de raisonnement particuliers.

En ce qui concerne les connaissances sur le domaine d'étude, celles-ci doivent être considérées comme spécifiques de l'archétype. Classiquement, au sein des modèles utilisateurs tels que présentés par [Brusilovsky, 1996], deux pratiques sont courantes par la représentation de connaissances de l'usager : le recours à un stéréotype prédéfini, correspondant à un niveau de connaissances particulier, en général novice ou expert, et le recours à un modèle de superposition («overlay model» en anglais), où le niveau de connaissances de l'utilisateur est défini pour chaque concept de la représentation de connaissances du domaine.

Dans le cadre du modèle de synthèse, une approche mixte est envisagée, avec un modèle de superposition défini pour chaque archétype. Ainsi, si l'on considère une représentation du domaine sous forme d'une taxonomie, cette approche revient à prendre en compte une taxonomie par archétype. Plus précisément, les différences de connaissances entre individus relevant d'un archétype ou d'un autre concernent majoritairement la précision et l'étendue de leurs connaissances. Dans le cas d'une taxonomie du domaine étudié, chaque archétype se voit donc associé à une partie plus ou moins complète de l'arbre de la taxonomie. La Fig. 3.11 illustre ce phénomène en comparant des taxonomies hypothétiques pour un anatomopathologiste et un étudiant en biologie de 1^{er} cycle. Le premier, ayant une plus grande expertise du domaine, possède un arbre aux branches plus fournies.

Au niveau des raisonnements particuliers, soit des comportements idiosynchratiques associés à chaque archétype, ceux-ci altèrent les opérations de chaque tâche de synthèse. Ceux-ci peuvent donc être considérés comme des contraintes supplémentaires posées sur la résolution de chaque sous-tâche élémentaire. Ainsi, si l'on considère l'exemple de la sous-tâche de gestion des données manquantes, deux méthodes de résolution (exclusion des individus ayant des données manquantes et inférence des données manquantes à partir d'individus similaires) sont possibles. Les archétypes utilisateurs permettent de privilégier l'une par rapport à l'autre. Par exemple, l'anatomopathologiste veut se baser sur des données réelles et fiables pour réaliser des publications, et va imposer l'exclusion des entités comportant des données manquantes. Par contre l'étudiant cherche uniquement à comprendre des mécanismes

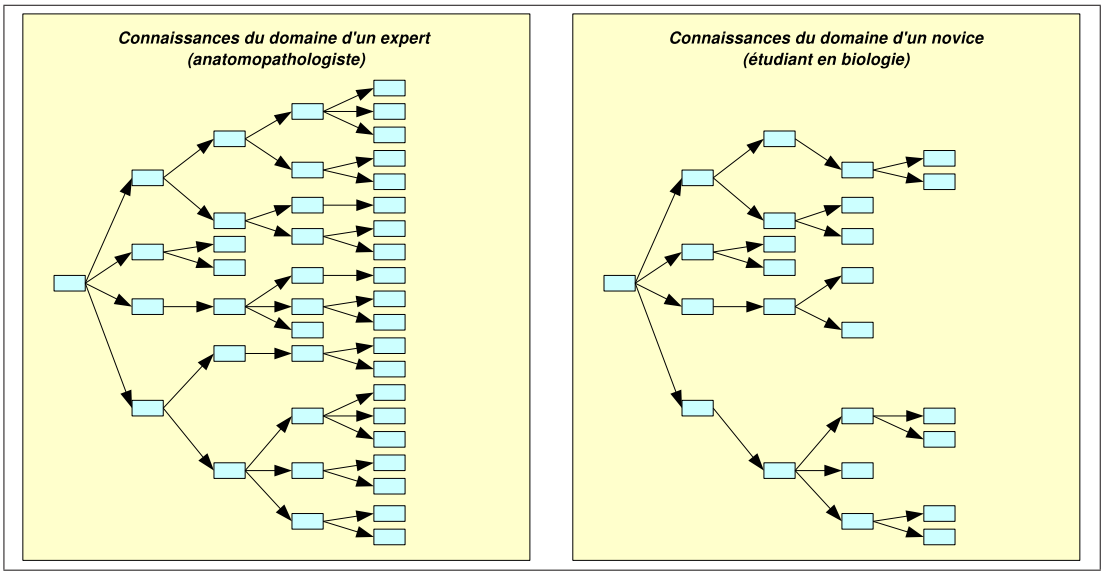


Fig. 3.11: Variation des connaissances du domaine selon l'archétype - L'anatomopathologiste est un expert du domaine, contrairement à l'étudiant. L'arbre des connaissances à sa disposition est plus fourni.

génériques et peut accepter des données inférées.

3.4.2.2.4 Personnalisation de l'interaction

La personnalisation de l'interaction avec le système passe ensuite par une individualisation des archétypes et d'autres éléments qui ne sont pas spécifiques d'un archétype.

A propos de l'individualisation d'un archétype, celle-ci peut consister tout d'abord en l'altération de la représentation des connaissances du domaine de l'archétype pour un individu particulier. Cette altération peut prendre deux directions.

Premièrement, elle peut concerner la représentation de la taxonomie du domaine d'étude. Par exemple, l'étudiant A peut avoir besoin d'accéder aux mêmes concepts que l'anatomopathologiste. En pratique, ceci correspond à l'ajout ou à la suppression de branches de la taxonomie du domaine, par rapport à l'arbre de l'archétype correspondant.

Deuxièmement, elle peut concerner les comportements idiosynchratiques de l'individu. Elle induit alors la mise en place de contraintes supplémentaires sur les tâches par rapport à celles incluses dans l'archétype. Ainsi l'étudiant A peut préférer une exclusion des données manquantes et l'étudiant B des inférences.

D'autres éléments de l'interaction font partie de ces caractéristiques qui se retrou-

vent dans nombre de logiciels et ne sont pas spécifiques à la synthèse, ni à la notion d'archétype. Ils relèvent d'un choix ou d'un goût individuel, tels que des thèmes de couleurs, des emplacements de menus. Ces préférences sont difficilement évaluables par l'analyse des interactions de l'usager avec le système, qui sont couramment stockées sous forme de traces d'exécution. Elles relèvent en général d'une saisie directe dans une interface de construction de profil spécifique.

3.4.2.3 Documents structurés

3.4.2.3.1 Introduction

La synthèse, comme tout processus de Recherche d'Information, se base sur un corpus documentaire pour effectuer la recherche d'éléments pertinents. Classiquement, en Recherche d'Information, l'utilisation de ce corpus a lieu suite à un processus d'indexation, conduisant à des représentations de documents, en général sous forme de vecteurs de mots. On pourrait aussi l'envisager de manière directe, par extraction de connaissances à partir de texte libre.

Ces deux façons d'envisager le corpus documentaire présentent toutefois des limites. La représentation sous forme de vecteurs de mots fait perdre le sens des documents et ne permet plus qu'une exploitation statistique. L'utilisation d'une collection de documents en langue naturelle pose un certain nombre de problèmes, telles que l'impossibilité d'une compréhension d'un texte par un système informatique ou les difficultés associées à l'extraction d'informations sur un concept.

L'existence de telles limites suggère le recours à une représentation de documents intermédiaire, tout à la fois plus complexe que le «sac de mots» courant en Recherche d'Information mais moins complexe que la langue naturelle. Ces considérations vont être présentées dans la suite.

3.4.2.3.2 Des limites des représentations documentaires courantes

L'utilisation d'un corpus documentaire sous forme de textes en langue naturelle ou sous forme de vecteurs de mots pose un certain nombre de problèmes dans le cadre d'une synthèse scientifique.

Tout d'abord, les outils de traitement automatique des langues actuels restent encore très basiques et ne sont pas en mesure de permettre une compréhension des textes par des logiciels informatiques. Les systèmes de Recherche d'Information courants réalisent donc un traitement majoritairement statistiquement au niveau mot ou groupe de mots et les représentations de connaissances permettent surtout

une gestion des problèmes de polysémie, homonymie, synonymie, etc. Or la synthèse, réalisée sur de textes en langue naturelle, nécessite une compréhension, une interprétation, une construction de sens.

Ensuite, une tâche élémentaire de synthèse est en général réalisée en relation avec un ou des concepts du domaine d'étude, comme par exemple, dans le domaine des TMA, la localisation de la tumeur ou un diagnostic. De plus, comme on l'a vu Paragraphe 3.4.2.1, les opérations possibles autour de ces concepts ne se limitent pas à une simple sélection comme le fait la Recherche d'Information, puisqu'une tâche de synthèse inclut des problématiques d'organisation et de présentation. Or l'extraction d'informations associées à ces concepts à partir de textes, par des méthodes proches du traitement de la langue ou de la Recherche d'information, est une opération difficile et sujette à erreurs. Mais ces erreurs ne sont pas acceptables dans le cadre de recherche scientifique qui est celui de la synthèse telle qu'envisagée dans le domaine applicatif des TMA.

Ces limites suggèrent le recours à un mode de représentation des documents qui soit intermédiaire entre le vecteur de mots et le texte libre.

3.4.2.3.3 Une solution : des documents structurés

La représentation de documents sous une forme structurée, par exemple au format XML, avec une structure connue (soit un schéma XML fixé à l'avance) permet de résoudre les problèmes rencontrés avec le texte libre ou une représentation sous forme de vecteurs de mots.

En effet, la structure du document fournit des indices supplémentaires sur leur sens, qui permettent de s'affranchir d'une compréhension fine dans le contexte de synthèse et de dépasser une simple juxtaposition de termes, ainsi que le présente [Groß-Hardt, 2002].

De plus, afin de résoudre une tâche élémentaire, il faut être en mesure d'identifier des éléments de connaissances du domaine et les valeurs qui leur sont associées dans chaque document (exemple : un diagnostic et sa valeur d'adénocarcinome Lieberkühnien). La résolution de ce problème est facilitée par le recours à des documents structurés dont la structure est connue. Dans ces documents, les éléments de structure peuvent être associés à des concepts et leur contenu à une valeur ou un texte dans lequel rechercher des informations, et l'adéquation entre représentations de connaissances et représentations de documents est facilitée.

Le choix d'une représentation de documents sous une forme structurée est aussi particulièrement adéquate dans un cadre de recherche scientifique, et en particulier dans le domaine des TMA. En effet, les informations associées à la pratique de cette

technique peuvent souvent être représentées sous forme d'ensemble de faits simples : nombres (un âge, une mesure, etc.), termes (nom d'un organe, d'une maladie, etc.) ou textes courts (commentaires d'une image, d'un diagnostic, antécédents d'un patient, etc.). Ces faits sont aussi en général facilement organisables au sein d'une structure fixe : le dossier clinique d'un patient souffrant d'un cancer du côlon comportera toujours le même jeu d'informations basiques, quel que soit le patient ou le mode de stockage du dossier. Ainsi, la Fig. 3.12 présente une vue simplifiée d'un tel dossier, qui se décompose en informations d'état-civil, diagnostic, données cliniques et thérapeutiques.

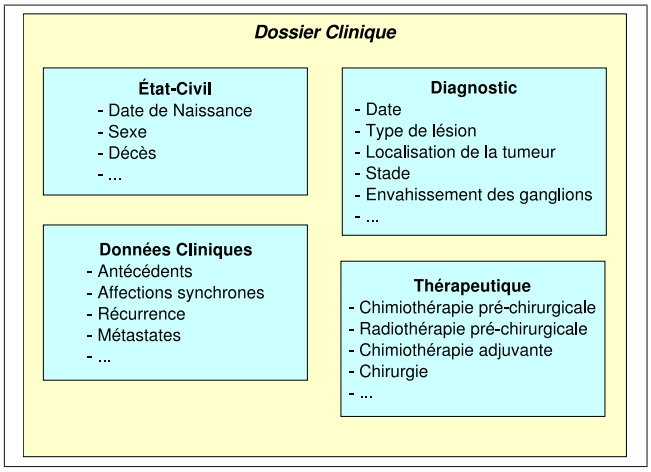


Fig. 3.12: Structure d'un dossier clinique - Le dossier clinique d'un patient atteint d'un cancer du côlon se décompose en quatre grandes sections, elles-mêmes structurées en catégories, etc.

3.4.3 Interactions

3.4.3.1 Requête Structurée

3.4.3.1.1 Introduction

Dans les systèmes de Recherche d'Information courants, la requête est constituée d'une liste de mots clés dont tous les éléments ont tous la même valeur vis à vis du problème de sélection de documents pertinents.

Dans le cadre de la synthèse, l'ajout de la notion de tâche à la Recherche d'Information rend inadaptée cette formulation sous forme de liste et nécessite une altération de la forme de la requête, et en particulier sa structuration. La multiplicité des tâches de synthèse implique une multiplicité des structures de requête, conduisant à la notion de modèle de requête.

Ce problème de formulation et structuration de la requête en un modèle est

analysé dans les paragraphes suivants.

3.4.3.1.2 Un problème de formulation

La formulation de la requête doit permettre la spécialisation d'un modèle de tâche pour une étude particulière. Chaque ensemble de mots clés doit donc pouvoir être associé à une ou des sous-tâches élémentaires du modèle de tâche pour les spécialiser. Une liste de mots clés n'est alors plus suffisante, comme illustré Fig. 3.13.

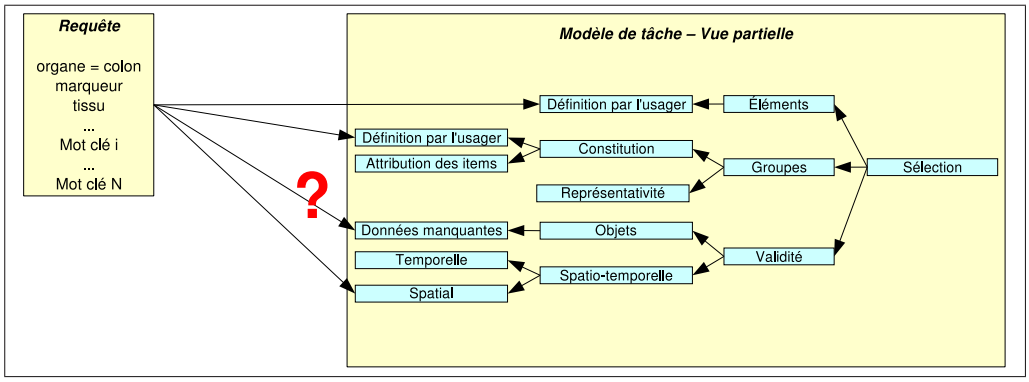


Fig. 3.13: Problème de spécialisation du modèle de tâche - Le recours à une requête sous forme de liste de mots clés, à gauche, pose un problème de spécialisation du modèle de tâche (présenté sous une forme partielle limitée à la problématique de sélection) à droite. En effet, il est impossible de déterminer avec une telle requête quels mots clés devront être utilisés pour spécialiser une sous-tâche particulière : l'organe côlon permet-il la sélection d'éléments, ou la constitution de groupes ?

Les éléments de la requête ayant tous la même valeur vis à vis du modèle de tâche considéré, il est impossible de déterminer quel mot clé utiliser pour spécialiser une sous-tâche élémentaire particulière. Par exemple, l'organe côlon permet-il de sélectionner des patients, ou de les grouper ? La spécialisation d'un modèle de tâche induit donc une altération de la formulation de la requête.

3.4.3.1.3 Une altération de la requête qui devient structurée

Afin de résoudre le problème de spécialisation du modèle de tâche par la requête, une solution possible est de donner à différents éléments de la requête une fonction différente vis à vis du modèle de tâche. Ce type de point de vue est introduit par exemple dans [Mäkelä et al., 2005] pour la Recherche d'Information dans le cadre du Web sémantique. Dans ces travaux, la requête est construite selon diverses facettes, permettant la présentation des résultats selon des vues différentes dépendant des facettes de la requête.

Dans le cadre du problème de synthèse, la requête est alors considérée comme un ensemble de n-uplets de rôles/valeurs où chaque tuple joue un rôle différent vis à vis

de la tâche de synthèse considérée. Ce principe est illustré schématiquement par la Fig. 3.14.

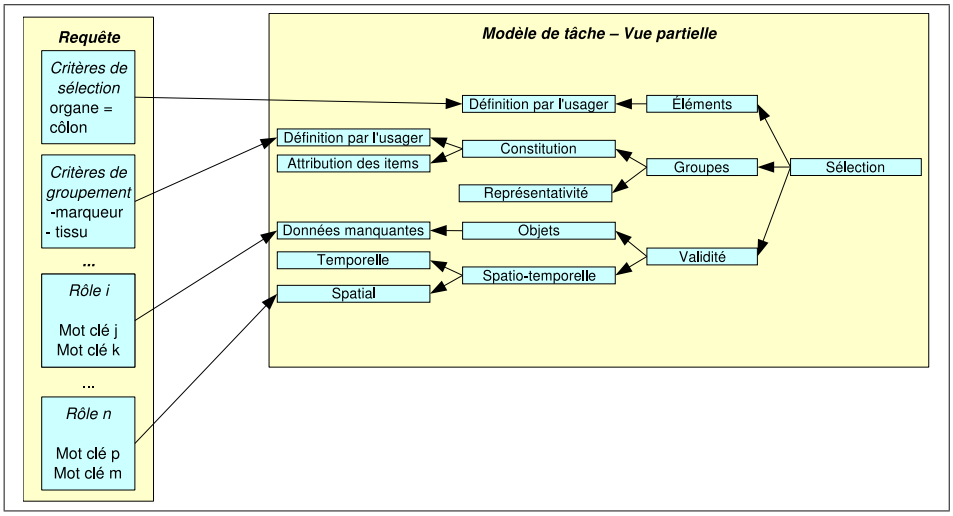


Fig. 3.14: Principe d'une requête structurée en rôles - La requête est décomposée en rôles, chacun spécialisé par des mots clés. Les rôles sont associés à des sous-tâches élémentaires de la tâche prototypique, permettant de diriger la spécialisation du modèle de tâche avec le contenu du rôle correspondant.

Chaque requête est décomposée en un ensemble de rôles et chaque rôle se voit spécialisé par un ensemble de mots clés lors de la formulation de la requête par l'utilisateur. Chaque rôle est associé à une ou des sous-tâches du modèle de tâche. La spécialisation du modèle de tâche pour une étude spécifique consiste alors en une association de chaque sous-tâche élémentaire avec les valeurs du rôle de la requête correspondant. Ainsi, les mots clés avec pour rôle «critères de sélection» (dans notre exemple, l'organe côlon) sont associés à la sous-tâche de définition des éléments sélectionnés, alors que ceux qui ont comme rôle «critères de groupement »(dans l'exemple le marqueur, le tissu tumoral ou adjacent, etc.) sont utilisés par la sous-tâche de constitution des groupes.

3.4.3.1.4 Une notion de modèle de requête

Les modèles de tâches varient en fonction de la tâche prototypique. Les rôles de la requête sont donc eux aussi dépendants de la tâche prototypique. Chaque tâche prototypique est en conséquence associée à un modèle de requête définissant ces rôles.

Chaque modèle de requête consiste donc en une liste de rôles que l'utilisateur doit spécialiser pour représenter l'étude qu'il veut réaliser. Une telle liste, si présentée de manière non organisée, serait difficile à appréhender par l'usager. Or, les divers rôles de la requête remplissent des fonctions différentes vis à vis du modèle de tâche

correspondant. Certains permettent une représentation générale de l'étude, tel son titre ou le domaine applicatif associé. D'autres permettent la description de l'étude en tant que telle, soit une définition des besoins synthétiques de l'usager. D'autres enfin précisent le contexte expérimental particulier de l'étude.

Un modèle de requête inclut donc trois ensembles de rôles distincts : généralités, besoins et contraintes expérimentales. Un exemple de modèle de requête pour une tâche de type «comparaison» est proposé Tab. 3.1.

Tab. 3.1: Exemple de modèle de requête - Il s'agit d'un modèle de requête pour une tâche de type «comparaison», présentant les trois catégories de rôles : généralités, besoins et contraintes expérimentales, ainsi que leurs contenus particuliers pour une requête de ce type.

Élément du modèle	*Description*
Généralités :	
- Tâche	Catégorie de tâche, ici comparaison
- Titre	Titre de l'étude correspondant à la requête
- Description	Description plus précise de l'étude
- Domaine	Domaine applicatif auquel se rapporte la requête
Besoins :	
- But	Élément cible, c'est-à-dire élément dont les valeurs doivent être comparées
- Critères d'inclusion	Critères guidant la sélection des items pertinents, soit termes classiques de Recherche d'Information
- Critères de groupement	Critères guidant l'organisation des items, soit critères définissant la composition des groupes à comparer
- Critères de tri	Critères guidant l'ordonnancement des items au sein d'un groupe
Contraintes expérimentales :	
- Langue	Langue à utiliser pour l'affichage
- Couleurs	Code couleur à utiliser au sein du document de synthèse pour représenter des valeurs de variables
- Géométrie	Contraintes portant sur la structure dans laquelle les informations du document de synthèse sont affichées
- ... - Application des critères de sélection	Choix de la méthode d'application des critères de sélection, entre une application stricte ou approximative
- Gestion des données manquantes	Choix de la méthode de gestion des données manquantes entre une exclusion des individus aux données incomplètes et une inférence des données manquantes
- Représentativité des groupes	Choix de la méthode de vérification de la représentativité des groupes, entre une maximisation de la variabilité ou une maximisation de l'homogénéité
- ...	

Les ensembles de rôles les plus complexes relèvent des parties «Besoins» et «Contraintes Expérimentales», qui vont être explicités plus avant ici.

En ce qui concerne les besoins de l'utilisateur, les éléments à décrire se rapportent à l'étude à réaliser. Il s'agit dans l'exemple considéré de la comparaison d'une vari-

able entre un ensemble de groupes et sous-groupes. Il faut donc définir la variable à comparer, avec le rôle «But», et les critères de constitution des groupes, avec le rôle «Critères de groupement». Cette comparaison est réalisée sur une population donnée, qui est restreinte par des «Critères d'inclusion». Enfin, au sein de chaque groupe de niveau le plus fin, les individus doivent être organisés, et en particulier ordonnés, selon les valeurs des éléments définis comme «Critères de tri».

Pour les contraintes expérimentales, certaines, comme les Couleurs, permettent de poser des contraintes supplémentaires correspondant aux goûts de l'utilisateur. D'autres concernent plutôt le mode de raisonnement à utiliser dans le cadre de l'étude considérée pour une sous-tâche élémentaire particulière, par exemple la gestion des données manquantes par exclusion ou inférence.

Formuler une requête revient alors à sélectionner une tâche de synthèse et à spécialiser le modèle de requête correspondant avec des concepts du domaine applicatif pour construire une requête spécifique de l'étude à mener.

3.4.3.2 Synthèse

3.4.3.2.1 Introduction

Le processus de synthèse en tant que tel consiste en l'exécution d'une tâche spécialisée pour une étude particulière, afin de construire un document de synthèse. Il implique donc l'exécution de bouts de programmes correspondant chacun à une sous-tâche élémentaire. Les paramètres de ces entités logicielles sont les éléments de spécialisation de la sous-tâche dans le modèle de tâche spécialisé.

Ce processus induit la résolution d'un ensemble de problèmes.

Tout d'abord, il nécessite une représentation des éléments de résolution des tâches élémentaires, à des fins documentaires et pour aider à leur utilisation. Ensuite, il requiert le choix d'une méthode de résolution pour chacune des sous-tâches élémentaires et la spécialisation de chaque méthode de résolution choisie pour l'étude en cours. Ceci peut être assimilé à l'instanciation d'un objet d'une classe en programmation orientée objet, le modèle de tâche correspondant à une classe et la tâche spécialisée à une instance de la classe. Enfin, il nécessite un ordonnancement des éléments de résolution en vue de l'exécution, du fait de dépendances entre sous-tâches élémentaires.

Ces divers aspects vont faire l'objet de la suite du paragraphe.

3.4.3.2.2 Représentation des éléments de résolution des tâches élémentaires

De façon générale, le modèle de tâche associé à chaque tâche prototypique consiste en un ensemble organisé de sous-tâches élémentaires. Afin de résoudre un problème de synthèse, il faut associer à chacune des sous-tâches élémentaires au moins une méthode de résolution, correspondant à un bout de code.

Il devient alors nécessaire de construire une représentation de chacune des méthodes de résolution possibles, indiquant quelle sous-tâche élémentaire la méthode permet de résoudre et quel élément logiciel permet cette résolution effective. Ces caractéristiques évoquent fortement le concept des méthodes de résolution de problèmes, présent en Intelligence Artificielle, tel que décrit par [Fensel et al., 2001].

Ce concept inclut la définition de représentations pour chaque méthode individuelle, celles-ci étant organisées au sein d'une librairie de méthodes de résolutions. Chaque méthode est aussi associée à un élément logiciel l'implémentant, par exemple sous forme de composant logiciel, comme dans [Crubézy and Musen, 2004].

Il convient alors de proposer un mode de représentation pour chaque méthode de résolution, dans l'esprit du langage de spécification proposé par [Fensel et al., 2003]. Le Tab. 3.2 présente les éléments principaux devant intervenir dans une telle représentation.

Tab. 3.2: Représentation simplifiée d'une méthode de résolution de problème - Cette représentation simplifiée décrit les principales caractéristiques d'une telle représentation.

Élément de la représentation	*Description*
Nom	Nom de la tâche correspondant à la méthode décrite
Description	Description à des fins documentaires de la méthode de résolution
Composant	Élément logiciel correspondant
Cibles	Structures dans lesquelles écrire les résultats de l'exécution
Paramètres	Structures dans lesquelles lire les éléments de spécialisation de l'exécution

Il s'agit ensuite d'organiser ces représentations au sein d'une librairie de méthodes de résolution de problèmes, ou librairie de composants. Or, chaque composant est associé à une sous-tâche élémentaire d'un modèle de tâche. Il est donc légitime de considérer que l'ensemble des composants utilisables dans le cadre d'une tâche prototypique particulière soient organisés au sein d'une taxonomie identique à celle du modèle de tâche correspondant. Le même raisonnement étant valide pour toutes les tâches prototypiques, la taxonomie de la librairie de composants devient alors une jointure de l'ensemble des taxonomies des modèles de tâches.

3.4.3.2.3 Choix de la méthode de résolution

L'exécution du processus de synthèse, en fonction d'une tâche particulière décrite dans une requête, implique tout d'abord le choix du composant à exécuter pour chaque sous-tâche élémentaire du modèle de tâche correspondant.

En effet, on dispose de deux taxonomies à l'organisation similaire : une taxonomie de sous-tâches élémentaires pour le modèle de tâche et une taxonomie de composants au sein de la librairie de composants. Mais la correspondance un à un entre les deux a lieu au niveau sous-tâche élémentaire (rectangles hachurés de la Fig. 3.15), et il peut exister plusieurs méthodes de résolution ou composants pour chaque sous-tâche élémentaire. Par exemple, la sélection d'entités pertinentes par rapport à des critères de sélection peut être réalisée selon une méthode de correspondance exacte (en Recherche d'Information, ceci correspondrait à des documents contenant tous les mots de la requête) ou approximative (en Recherche d'Information, ceci correspondrait à des documents contenant au moins un mot de la requête).

Se pose alors le problème de la sélection du composant à exécuter, tel qu'illustré Fig. 3.15 pour la partie sélection d'un modèle de tâche de type comparaison. Par exemple, une question d'importance pour un domaine d'étude tel que celui des Tissue MicroArrays est la gestion des informations manquantes. Comme le montre la Fig. 3.15, la librairie de composants intègre divers heuristiques définissant des méthodes pour gérer ce problème spécifique, comme l'exclusion des individus contenant des données manquantes de la liste des items sélectionnés ou l'inférence des valeurs manquantes à partir d'individus similaires.

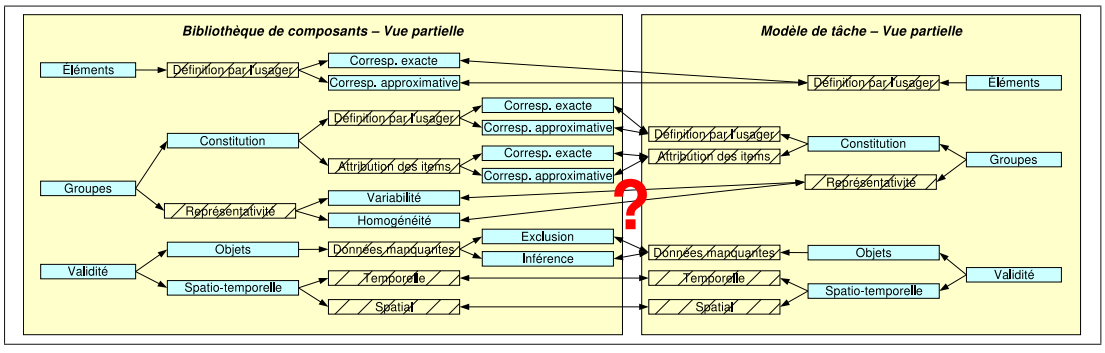


Fig. 3.15: Problème de choix de la méthode de résolution - Plusieurs composants ou méthodes de résolution peuvent être disponibles pour résoudre une même sous-tâche élémentaire (les sous-tâches élémentaires sont présentées dans des rectangles hachurés), ce qui pose un problème de choix de la méthode à utiliser.

La sélection du composant à utiliser pose alors un double problème.

Tout d'abord, les divers composants disponibles pour résoudre une sous-tâche élémentaire ne sont pas nécessairement interchangeables. Ils peuvent avoir des paramètres différents, des méthodes différentes, un format de sortie différent. Le

choix du composant à utiliser pose alors un problème de composition de composants, de compatibilité entre interfaces, entre entrées et sorties de composants qui seront exécutés successivement. Ce problème de composition intervient dès qu'un ensemble d'entités logicielles doit être composé pour accomplir une tâche plus complexe, et devient d'autant plus critique quand la composition est dynamique, comme c'est le cas ici. Cette problématique fait l'objet de nombreux travaux, par exemple dans le domaine de la composition de services Web, où [Rao and Su, 2004] réalisent une revue des approches possibles. Entre autres alternatives, une solution simple est alors de faire l'hypothèse que tous les composants correspondant à la même sous-tâche élémentaire ont une interface identique, éliminant ainsi le problème de composition.

Ensuite, intervient un problème complexe de fusion d'informations, pour déterminer la méthode de résolution à choisir dans le cadre d'une étude particulière. Or, des contraintes sur ces méthodes de résolution sont spécifiées à divers niveaux au sein du système de synthèse. La résolution du problème de fusion d'informations peut alors être envisagée sous forme d'heuristiques prenant en compte des contraintes telles que des préférences de l'utilisateur (comportements idiosynchratiques introduits Paragraphe 3.4.2.2), la requête (contraintes expérimentales présentées Paragraphe 3.4.3.1) ou le domaine d'application.

En effet, les divers domaines applicatifs possibles ont des spécificités qui doivent parfois être prises en compte pour la résolution des tâches. Les tâches de synthèse sont donc influencées par des caractéristiques du domaine d'application, qui seront qualifiées par la suite d'«expérimentales».

Les connaissances expérimentales peuvent être considérées comme la partie des connaissances du domaine qui est mobilisée au cours de la réalisation d'une tâche de synthèse. Les connaissances expérimentales sont hétérogènes et mal définies ; elles se basent sur des heuristiques pour orienter l'exploration des données d'un domaine applicatif particulier et fournissent des protocoles pour diriger les expériences. Ainsi, dans le domaine TMA où l'objectif est une validation d'hypothèses par des données expérimentales, les correspondances exactes vont être privilégiées par rapport aux inférences.

3.4.3.2.4 Spécialisation de la méthode de résolution

Le problème du choix de la méthode de résolution pour chaque sous-tâche élémentaire du modèle de tâche ayant été introduit, le modèle de tâche conduit à une représentation opérationnalisée de la tâche où chaque sous-tâche élémentaire est associée à une seule et unique méthode de résolution. Se pose alors le problème de la spécialisation de cette tâche opérationnalisée pour l'étude à réaliser.

Le Paragraphe 3.4.3.1 décrit comment un modèle de tâche peut être spécial-

isé à partir d'une requête structurée. Mais la requête n'est pas la seule source de paramètres pour les composants à utiliser pour résoudre une tâche de synthèse. Des spécificités du domaine applicatif, sous forme de contraintes expérimentales, et des préférences de l'utilisateur peuvent aussi intervenir, ainsi qu'illustré Fig. 3.16. Il s'agit là encore d'un problème de fusion d'informations entre de multiples sources.

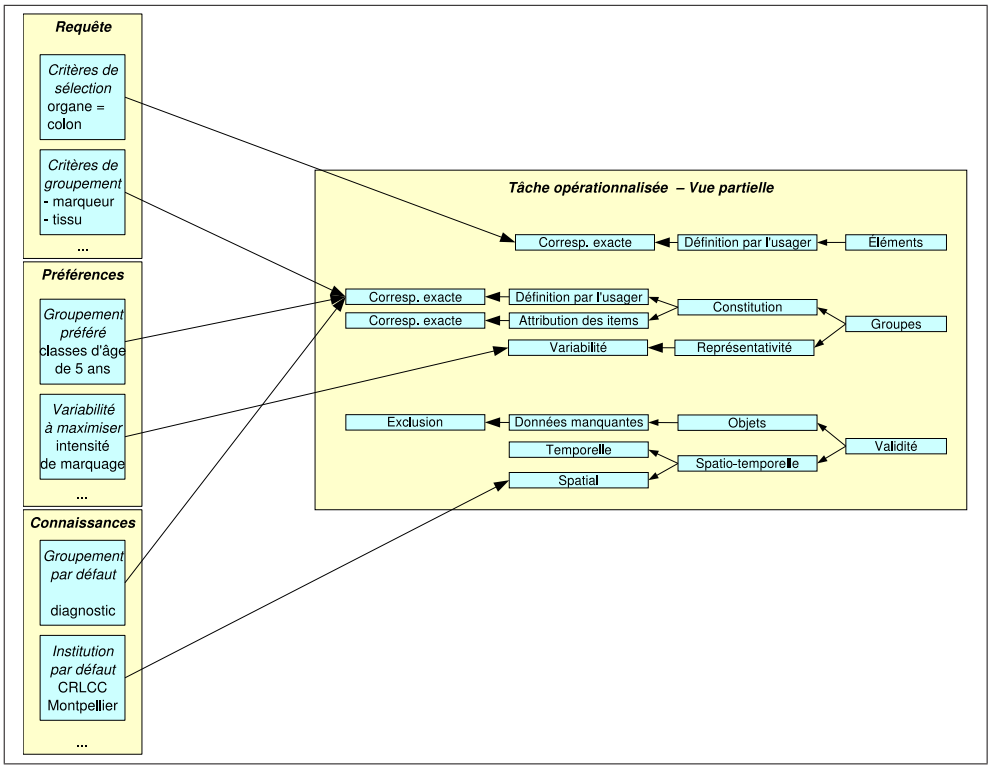


Fig. 3.16: Spécialisation d'une tâche - Des informations issues de la requête, des connaissances du domaine d'étude et des préférence utilisateur sont utilisées pour spécialiser la tâche opérationnalisée.

La question qui se pose alors est le choix de la source pour la spécialisation quand plusieurs sont possibles, comme par exemple pour le composant de constitution des groupes selon des critères définis par l'usager. En effet pour cette sous-tâche élémentaire, la requête indique, comme critères de groupement, marqueur, localisation du tissu par rapport à la lésion et localisation intracellulaire, les préférences utilisateur, des classes d'âge de 5 ans, et les connaissances du domaine, un diagnostic.

Choisir des critères de groupement implique alors de mettre en place des heuristiques de sélection entre les informations issues de la requête, des préférences utilisateur ou des connaissances du domaine d'étude. Cette sélection et l'association des éléments de spécialisation pertinents à chaque sous-tâche élémentaire du modèle de tâche permet la construction d'une représentation de tâche spécialisée.

3.4.3.2.5 Ordonnancement des éléments

Une fois définie une tâche spécialisée, la résolution du problème de synthèse requiert l'exécution du composant choisi pour chaque sous-tâche élémentaire avec comme paramètres les éléments de spécialisation associés. Cette exécution implique la définition d'un ordre d'exécution des composants.

En effet, chaque composant est défini comme ayant un ou plusieurs paramètres et une ou plusieurs cibles. Un paramètre d'un composant peut être défini à partir de la requête, des préférences de l'utilisateur ou des connaissances du domaine d'étude, mais aussi à partir du résultat ou cible d'un autre composant. Il existe donc des relations de dépendance entre composants, qu'il faut prendre en compte dans l'ordre d'exécution, ou plan de synthèse.

Le processus de construction de ce plan de synthèse est décrit Fig. 3.17 pour la partie sélection d'une tâche de comparaison.

La tâche spécialisée décrit l'ensemble des composants à exécuter pour résoudre le problème de synthèse particulier considéré. Parallèlement, la librairie de composants indique pour chaque composant la source de ses paramètres et cibles. La mise en corrélation de ces deux ensembles d'informations permet de construire un graphe de dépendances spécifique de l'étude à réaliser. Ce graphe de dépendances est alors utilisé pour déterminer un ordre d'exécution ou plan de synthèse, dont l'exécution permet la construction du document de synthèse.

3.4.3.3 Document de synthèse

3.4.3.3.1 Introduction

L'objectif du processus est la construction d'un document de synthèse permettant une visualisation compacte et complète des informations pertinentes pour l'étude menée par le chercheur. Cette construction dépasse la simple sélection et implique une organisation conceptuelle qui doit être reflétée par une organisation structurelle.

Les tâches de synthèse telles qu'envisagées dans la taxonomie du Paragraphe 3.4.2.1 incluent de manière sous-entendue la notion d'organisation structurelle. Par exemple, les tâches de comparaison induisent la constitution de groupes au sein d'une classification hiérarchique. La question qui se pose alors est le choix d'une structure qui permette de refléter cette organisation conceptuelle.

Les réflexions menant à ce choix et la description de la structure choisie vont être traitées ici.

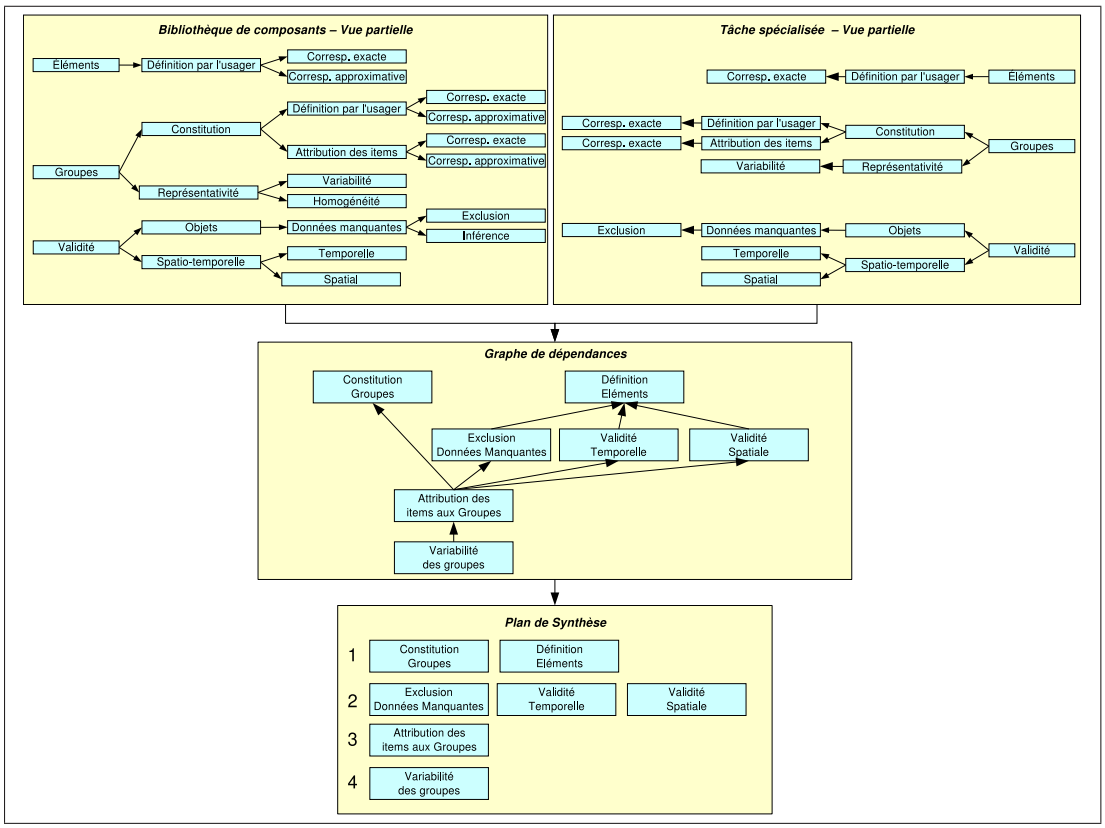


Fig. 3.17: Problème d'ordonnancement - A partir de la tâche spécialisée et de la librairie de composants, qui indique les paramètres et cibles de chaque composant, il faut construire un graphe de dépendances entre composants. Ce graphe de dépendances permet la définition d'un ordre d'exécution ou plan de synthèse.

3.4.3.3.2 Choix d'une structure pour le document de synthèse

La définition d'une structure pour le document de synthèse, structure permettant de mettre en valeur l'organisation conceptuelle des informations quelle que soit la tâche prototypique considérée, implique tout d'abord de déterminer une dimensionnalité de la visualisation. En effet, ce problème d'affichage doit être considéré comme un problème de Visualisation d'Information et des techniques de ce domaine doivent être mises en œuvre, de la même façon que [Shneiderman, 2002] suggère leur usage pour la fouille de données.

L'affichage du résultat d'une synthèse peut alors être imaginé en mode textuel, en 2D ou en 3D. Le Paragraphe 2.2.3.2 a permis une exploration de cette problématique de dimensionnalité «idéale» en Visualisation d'Information, à travers les évaluations d'interfaces menées par diverses équipes de chercheurs telles que celle de [Heidorn and Cui, 2000]. Ces travaux laissent entendre que dans l'état actuel des choses, les structures 2D sont le meilleur compromis entre l'expressivité de la représentation et l'effort cognitif à fournir pour sa compréhension et manipulation

par l'utilisateur.

Le problème suivant à résoudre est celui de la compacité de la présentation, nécessaire pour permettre la présentation d'un maximum d'informations dans un minimum d'espace. Dans ce contexte de représentation à deux dimensions, le concept de grille a été montré, toujours dans le Paragraphe 2.2.3.2, comme une solution pertinente.

De plus, dans le domaine applicatif des TMA, la notion de grille est porteuse d'un sens supplémentaire, puisqu'elle est en adéquation avec l'objet physique construit par la technologie. Cette correspondance entre la représentation virtuelle d'un jeu de données dans un but d'appréhension de données et la structure tangible d'une lame TMA devrait faciliter l'interprétation du document de synthèse pour les utilisateurs habitués au domaine applicatif.

Se pose alors la question de l'adéquation d'une représentation sous forme de grille avec les résultats des diverses tâches prototypiques envisagées dans le cadre de la synthèse de données TMA. Or, celles qui sont considérées ici, d'après la taxonomie établie Paragraphe 3.4.2.1, se décomposent en trois catégories : comparaison, évolution et distribution.

Il s'agit alors d'évaluer comment représenter les unes et les autres dans un contexte de grille, exercice mené Fig. 3.18. Pour chacune des trois catégories de tâches, il faut définir une distance et placer les individus de proche en proche selon cette distance.

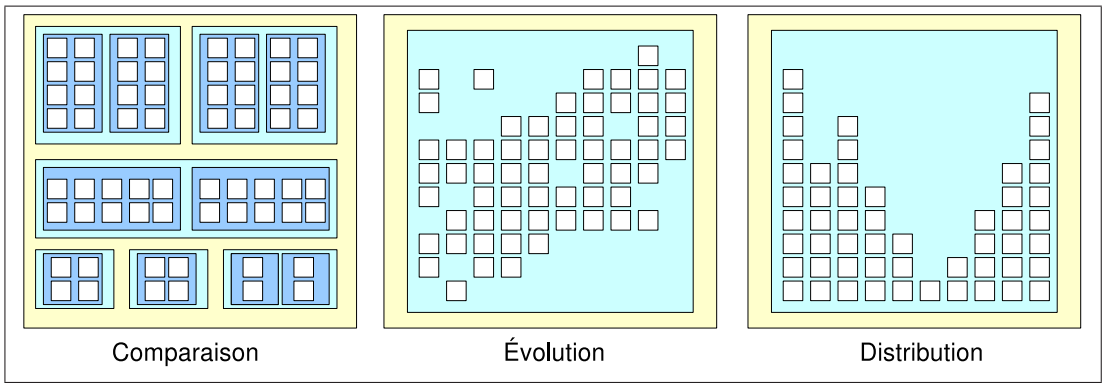


Fig. 3.18: Adéquation des tâches de synthèse avec une présentation sous forme de grille - Les trois grandes catégories de tâches de synthèse peuvent être supportées par une présentation des résultats au sein d'une grille.

En ce qui concerne les tâches de comparaison, la constitution d'une hiérarchie de groupes conduit à la mise en place d'un ensemble de «paquets», qui doivent se voir attribuer une aire de la grille. La direction qui est donnée pour placer les groupes ne correspond alors pas à l'un des bords de la grille mais plutôt à un placement du bord vers le centre. Au niveau des individus, la distance est définie par la variable choisie comme critère de tri et le placement se fait ligne par ligne.

Au niveau de l'évolution, plusieurs distances sont mises en place et correspondent chacune à une direction de la grille : de gauche à droite pour la première, de bas en haut pour la seconde, etc.

Enfin, pour la distribution, une seule distance est prise en compte, selon une seule direction de la grille, de gauche à droite. Les individus situés à la même distance sont alors «empilés».

La forme de grille paraît donc bien adaptée à la présentation du cœur de documents de synthèse, mais cette forme n'est pas suffisante et le problème du fond doit être abordé.

3.4.3.3.3 Aperçu du document à construire

Le document de synthèse est envisagé comme construit autour d'une structure de grille complexe où les éléments d'intérêt sont organisés, intégrés et agrégés. Au lieu d'une liste où l'organisation est guidée par la pertinence individuelle des éléments, la grille permet une agrégation complexe guidée par une tâche. Chaque item est alors considéré comme situé dans le contexte d'autres éléments avec lesquels il peut partager des propriétés ou au contraire dont il diffère.

Utiliser une grille permet une visualisation simple et compacte des résultats d'une recherche, offrant une vue relationnelle sur la collection et mettant en exergue des tendances majeures ainsi que des phénomènes inattendus (régularités, ruptures, points communs ou événements rares).

Mais cette représentation n'est pas suffisante pour permettre une appréhension dirigée du corpus informationnel et cette structure doit aussi être vue comme un document complexe organisant l'accès à la collection étudiée selon un point de vue particulier (la tâche de synthèse considérée).

Pour permettre cet accès, la grille doit être associée à des vues contextuelles dépendant du domaine d'étude, afin de permettre la construction de nouvelles connaissances ou pour servir de support à des publications, tel que présenté Fig. 3.19. Ainsi, dans le domaine des Tissue MicroArrays, des vues sur le patient, par son dossier clinique, et les informations histologiques, par les images et annotations ou mesures associées, sont jointes à la grille de synthèse.

De plus, l'ensemble peut être augmenté, à des fins documentaires et informatives, par des références à des études similaires, des liens bibliographiques, ou des informations sur les molécules étudiées, qu'il s'agisse de leur séquence en acides aminés ou de la séquence du gène correspondant.

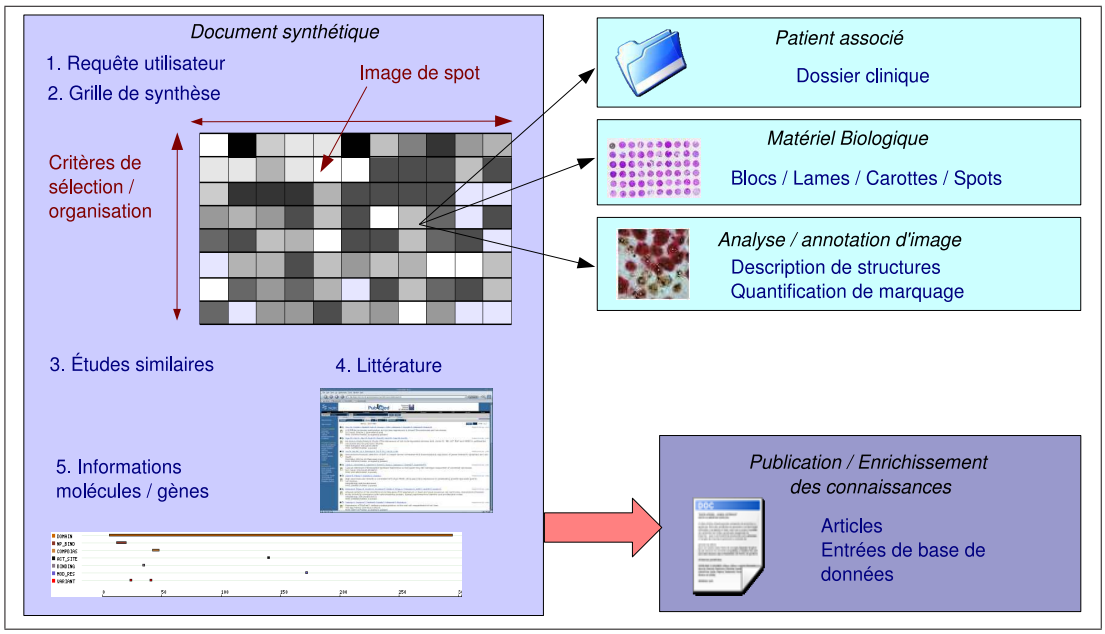


Fig. 3.19: Aperçu du document de synthèse - Le document de synthèse est centré autour d'une grille dont l'organisation structurelle reflète l'organisation conceptuelle guidée par des critères de sélection et organisation. Cette grille sert de point d'accès à des vues sur le domaine d'étude, ici le domaine des Tissue MicroArrays. Elle est aussi supportée par des informations contextuelles, par exemple bibliographiques, et est utilisée pour construire de nouvelles connaissances ou illustrer des publications.

3.4.4 Évaluations

3.4.4.1 Adéquation à la tâche

3.4.4.1.1 Introduction

Ainsi qu'il a été vu Paragraphe 2.3.3, la problématique d'évaluation des résultats est l'une des composantes sur lesquelles s'est contruite la Recherche d'Information, par la notion de pertinence. La synthèse, qui inclut des activités de sélection, est envisagée comme Recherche d'Information orientée tâche et se doit d'inclure de telles préoccupations.

Dans ce contexte d'évaluation, que ce soit pour les systèmes de Recherche d'Information en particulier, ou les logiciels en général, l'une des premières problématiques à prendre en compte est d'évaluer si le système proposé est en phase avec ce que les usagers veulent réaliser en l'utilisant. Dans le cadre de la synthèse, il s'agit d'une évaluation de l'adéquation à la tâche.

Évaluer l'adéquation à la tâche revient à déterminer si la formulation de requête structurée permet l'expression d'une tâche de synthèse telle qu'elle est envisagée par

l'utilisateur, soit de façon générale une adéquation au besoin. La question qui se pose alors est la forme que doit prendre l'évaluation de cette adéquation à la tâche.

3.4.4.1.2 Forme de l'évaluation

Ce qui doit être évalué dans le cadre de cette adéquation à la tâche est à quel point la notion de tâche de synthèse coïncide entre ses deux assertions différentes, celle de l'utilisateur et celle du système. Il s'agit donc, d'un point de vue pratique, d'une évaluation de la formulation de la requête.

Du point de vue de l'utilisateur, la tâche de synthèse se veut représentative d'un besoin réel. Or, comme le montre [Mizzaro, 1998], ce besoin réel est inaccessible, même à celui qui l'éprouve. C'est alors le besoin perçu qui est mis en jeu dans l'évaluation. Du côté système, la tâche de synthèse est représentée par une requête structurée. Il s'agit alors d'évaluer la correspondance entre ces deux représentations, ce qui implique de s'intéresser tout à la fois à la forme de cette évaluation et à une description plus fine de ce qui doit être évalué.

En ce qui concerne la forme de l'évaluation, la présence d'une composante humaine pose problème. En effet, la prise en compte de facteurs humains exclut la définition de mesures objectives calculées mathématiquement. Le seul moyen d'accéder aux perceptions de l'utilisateur est le recours à des interviews et questionnaires dans le cadre d'études utilisateur.

Il faut alors déterminer plus précisément ce que ces études utilisateur doivent montrer. Or, l'adéquation à la tâche telle qu'envisagée peut être appréhendée selon plusieurs dimensions.

Tout d'abord, il faut déterminer si la notion de tâche de synthèse pour l'utilisateur et la notion de tâche de synthèse pour le système correspondent au même concept, soit une adéquation de la tâche au sens stricte. Ainsi, il s'agit d'estimer si, dans l'absolu, les tâches prototypiques proposées correspondent effectivement à de vrais problèmes d'appréhension de données, tels qu'ils sont perçus par les utilisateurs. De plus, il n'est pas évident que ces tâches prototypiques soient adaptées dans le contexte applicatif considéré. Enfin, la formulation d'une requête, permettant la spécialisation d'une tâche prototypique, se base sur une taxonomie du domaine d'étude dont l'organisation doit être évaluée.

Ensuite, il faut évaluer si l'ensemble des tâches prototypiques proposées par le système couvre bien l'ensemble des tâches de synthèse envisagées par l'usager, c'est-à-dire une notion de complétude. Cette complétude concerne d'une part les tâches prototypiques proposées et d'autre part le vocabulaire du domaine tel que présenté dans la taxonomie.

Tab. 3.3: Dimensions de l'évaluation de l'adéquation à la tâche - Le dimensions envisagées pour l'évaluation de l'adéquation à la tâche sont présentées ici avec des exemples d'affirmations pour lesquelles l'usager devra donner un niveau d'approbation dans le cadre d'un questionnaire.

Dimension de l'évaluation	*Exemples d'affirmations*
Adéquation	- Les tâches proposées, telles que je les comprends, correspondent à de vrais problèmes d'appréhension de données. - Les tâches proposées (comparaison, évolution et distribution) sont adaptées aux problématiques courantes du domaine. - L'organisation taxonomique du vocabulaire est adéquate.
Complétude	- Tous les types de tâches que je voudrais réaliser sont supportées. - Le vocabulaire proposé est complet par rapport au domaine d'étude.
Expressivité	- La transposition d'un problème que je voudrais explorer en une requête ne pose pas de problème. - La syntaxe de la requête permet de décrire les tâches avec précision. - La syntaxe de la requête permet d'exprimer toutes les tâches de comparaison ou d'évolution que je voudrais réaliser.
Extensibilité	- Il est facile d'explorer les données en modifiant la requête.
Navigabilité	- L'organisation des éléments de l'interface de saisie est claire. - La saisie de la requête est facile à réaliser.

Puis il faut estimer si la formulation de requête telle que supportée par le système permet à l'usager d'exprimer son problème, c'est-à-dire l'expressivité de la requête. Cette expressivité peut être tout d'abord envisagée en terme de transposition d'un problème biologique sous forme de requête structurée, dont la charge intellectuelle doit être estimée. Ensuite, elle peut être considérée en terme de précision de la formulation. Enfin, elle rejoint un peu la notion de complétude, car cette expressivité doit aussi évaluer la capacité de la requête à décrire toutes les études possibles.

De plus, la reformulation joue un rôle dans la synthèse de même que dans la plupart des systèmes de Recherche d'Information modernes, ce qui induit la mesure d'une extensibilité.

Enfin, la formulation d'une requête implique une interaction entre l'usager et le système, qui apporte une problématique d'interface utilisateur, sous forme d'une notion de navigabilité. Cette navigabilité peut être considérée d'un point de vue général d'organisation de l'interface, et d'un point de vue particulier de saisie de requête.

Le questionnaire utilisé pour l'évaluation de l'adéquation à la tâche dans le cadre d'une étude utilisateur se doit de couvrir ces thématiques. Ainsi, le Tab. 3.3 propose quelques exemples d'affirmations, correspondant à chacune de ces dimensions. Le niveau d'approbation d'un utilisateur pour chaque affirmation peut être évalué dans le cadre d'une étude utilisateur.

3.4.4.2 Pertinence situationnelle

3.4.4.2.1 Introduction

La pertinence situationnelle telle qu'envisagée ici consiste en une évaluation du processus de synthèse en tant que telle, soit une évaluation objective du document construit en fonction de la tâche de synthèse et du corpus documentaire. En tant que telle, elle est à rapprocher de la notion de pertinence système des systèmes de Recherche d'Information.

Mais, de la même façon que la sélection d'items pertinents au sein d'une collection de documents en fonction d'une requête n'est qu'une étape parmi d'autres du processus de synthèse, des mesures de type précision et retour ne peuvent pas seules répondre au problème d'évaluation de la construction d'un document de synthèse. D'autres dimensions de ce problème d'évaluation système du processus doivent être prises en compte, tel que présenté par la suite.

3.4.4.2.2 Une évaluation multidimensionnelle

La synthèse est considérée comme un processus de sélection d'éléments pertinents, organisation conceptuelle des entités sélectionnées puis organisation structurelle reflétant cette organisation conceptuelle et enfin présentation de ces éléments au sein d'un document synthétique. Ces catégories d'activités doivent être reflétées dans une pertinence système.

Or les mesures classiques en Recherche d'Information, telle que la précision et le retour, ne permettent une évaluation objective que de l'efficacité de la sélection de chaque document, considéré individuellement. Il faut donc ajouter de nouvelles dimensions à ces mesures, correspondant aux activités complémentaires induites par la synthèse par rapport à la Recherche d'Information : organisation conceptuelle, organisation structurelle et présentation.

Ainsi, la première dimension supplémentaire doit évaluer l'organisation conceptuelle des entités sélectionnées. Ceci consiste donc à mesurer la pertinence de la classification des éléments qui est faite en fonction de la tâche de synthèse considérée. La seconde doit mesurer l'organisation structurelle, ou du moins dans quelle mesure l'organisation structurelle reflète l'organisation conceptuelle. Il s'agit donc d'une mesure de la qualité du placement des entités au sein de la grille du document de synthèse. Enfin, la troisième doit estimer la qualité du rendu du document de synthèse. Il s'agit donc d'une mesure de la qualité de l'affichage.

Ensuite, la pertinence situationnelle dépasse sans doute la conjonction de ces

quatre ensembles de mesures, de la même façon que le tout dépasse souvent la somme des parties. L'objectif de la synthèse est d'être un support à la construction de nouvelles connaissances, en permettant l'appréhension d'un espace documentaire. Une évaluation objective indiquant dans quelle mesure cet objectif est atteint semble nécessaire, bien que difficile à concevoir.

Enfin, la synthèse inclut des sous-tâches dont l'objectif affiché est d'améliorer la qualité de l'ensemble du processus, telles que la gestion des données manquantes ou de la validité des données. Une mesure des performances de ces processus qualité s'impose.

La pertinence situationnelle, conjonction de l'ensemble des mesures esquissées ici, apparaît donc comme multidimensionnelle, et, bien qu'envisageable sous forme d'une mesure objective, difficile à mettre en œuvre.

3.4.4.3 Pertinence interprétationnelle

3.4.4.3.1 Introduction

La pertinence interprétationnelle telle que proposée dans le modèle de synthèse consiste en un jugement de valeur du document de synthèse par l'usager en fonction de son problème de synthèse tel qu'il l'a exprimé dans la requête structurée. En tant que telle, cette pertinence interprétationnelle pose le même type de problématiques que l'adéquation à la tâche, puisqu'elle implique des facteurs humains.

Dans ce cadre, son évaluation ne peut être envisagée, comme l'adéquation à la tâche, que dans le cadre d'une étude utilisateur.

3.4.4.3.2 Composantes de la pertinence interprétationnelle

De même que l'adéquation à la tâche, la pertinence interprétationnelle peut être considérée selon plusieurs axes.

Tout d'abord, la clarté du document de synthèse peut être évaluée. Celle-ci peut être envisagée en sens d'intuitivité. Cette intuitivité concerne tout à la fois la relation entre la requête et le document de synthèse, qui doit être facilement compréhensible, et l'interprétation du document de synthèse, qui doit rester simple.

De plus, le contenu du document en tant que vecteur de sens doit être pris en compte, ce qui correspond à une notion d'informativité. Cette informativité concerne tout d'abord la forme du rendu, une grille, qui doit être pertinente. Ensuite, elle

Tab. 3.4: Dimensions de la pertinence interprétationnelle - Les dimensions envisagées pour la pertinence interprétationnelle sont présentées ici avec des exemples d'affirmations pour lesquelles l'usager devra donner un niveau d'approbation dans le cadre d'un questionnaire.

Dimension de la pertinence interprétationnelle	*Exemples d'affirmations*
Intuitivité	- La relation entre la requête et le résultat affiché est facilement compréhensible. - L'interprétation du résultat de la requête ne pose pas de problème.
Informativité	- La forme du rendu (tableau) est pertinente. - Le résultat proposé est porteur de sens. - Le résultat proposé apporte de nouvelles informations par rapport au problème posé.
Utilité	- Le document de synthèse affiché m'aide à répondre au problème exprimé dans la requête. - Le résultat pourrait être directement utilisé par exemple dans une publication.
Suggestivité	- Les résultats affichés m'aident à envisager de nouvelles requêtes. - Les résultats affichés suggèrent d'autres analyses avec d'autres outils.
Navigabilité	- La navigation au sein des informations (grille et fiches associées) est facile.

relève du contenu, qui doit tout à la fois avoir un sens perceptible par les utilisateurs et apporter de nouvelles informations, sinon la construction du document de synthèse est vaine.

Ensuite, l'usage du document doit être pris en compte, ce qui peut être considéré comme son utilité. Cette utilité peut être envisagée dans l'absolu, en tant qu'aide à la résolution du problème posé par la requête, et de manière pratique en tant qu'illustration de nouvelles connaissances par exemple au sein d'un article de journal.

Par la suite, le rôle que le document peut jouer dans la genèse de nouvelles études doit être considéré, c'est-à-dire une évaluation de sa suggestivité. Cette suggestivité concerne tout à la fois de nouvelles requêtes posées au sein de l'outil, ou l'utilisation d'autre outils, tels que des systèmes de fouille de données.

Enfin, il s'agit d'une structure interactive, ce qui pose là aussi des questions de navigabilité, entre la grille et les vues associées.

Ces divers points de vue sur la pertinence interprétationnelle peuvent être évalués par le biais d'un questionnaire dans le cadre d'études utilisateur. Ainsi, le Tab. 3.4 propose quelques exemples d'affirmations, correspondant à chacune de ces dimensions. Le niveau d'approbation d'un utilisateur pour chaque affirmation peut être là encore évalué dans le cadre d'un questionnaire.

3.5 Un modèle à la mise en œuvre non triviale

La synthèse a été abordée dans ce chapitre sous l'angle d'une Recherche d'Information orientée tâche. En tant que telle, il s'agit d'une activité qui peut être considérée comme relevant de la Recherche d'Information, augmentée d'un ensemble de paradigmes d'autres domaines, soit une activité multifacettes. Il s'agit aussi d'une notion fédératrice pour un ensemble de problèmes variés. Ces deux axes aux dimensionnalités élevées en font un problème difficile, dont la résolution informatisée relève de l'Intelligence Artificielle, par la mimique du processus réalisé manuellement par les chercheurs.

Considérer la synthèse selon cette vision de Recherche d'Information orientée tâche suggère la définition d'un modèle de synthèse, de la même façon que les divers points de vue sur la Recherche d'Information sont supportés par des modèles. Ce modèle de synthèse est basé sur un triptyque *Utilisateur* $\leftrightarrow$ *Problème* $\leftrightarrow$ *Information*, qui sert de base ici pour les diverses interprétations de la Recherche d'Information. Il apparaît surtout comme un modèle intermédiaire entre modèle opérationnel, proche du système, de l'algorithmique, et modèle comportementaliste, proche de l'utilisateur et de ses processus cognitifs, de haut niveau d'abstraction. Enfin, une analyse plus fine des diverses composantes du modèle permet d'ouvrir la voie à son opérationnalisation.

Ainsi, le modèle de synthèse inclut un ensemble d'entités. La notion de tâche prototypique recouvre les divers types d'études possibles, organisées au sein d'une taxonomie. Chaque tâche prototypique est associée à un modèle de tâche qui la décompose en une hiérarchie de sous-tâches élémentaires, potentiellement influencé par des contraintes expérimentales. L'utilisateur est considéré en terme d'archétypes, représentant une classe d'usagers, leurs connaissances sur le domaine d'étude et modes de raisonnements particuliers, qui sont individualisés par des préférences. Le corpus informationnel consiste en documents structurés.

Ensuite, ces entités sont liées par des interactions. L'expression d'une tâche particulière par un utilisateur est réalisée par l'intermédiaire d'une requête structurée, basée sur un modèle de requête spécifique d'une tâche prototypique. Ce modèle de requête est structuré en éléments génériques, besoins et contraintes expérimentales. Il est décomposé en rôles dont chacun est associé à une sous-tâche élémentaire du modèle de tâche correspondant. Le processus de synthèse, construction d'un document selon une tâche particulière sur une collection de documents donnée, implique la représentation des éléments de résolution correspondant à chaque sous-tâche élémentaire, leur sélection quand plusieurs sont disponibles pour résoudre un même problème, et leur ordonnancement. Le document de synthèse, présentation à l'utilisateur des résultats de la synthèse réalisée sur un corpus documentaire particulier, est envisagé sous forme d'une grille documentaire fournissant des points d'accès à des vues particulières sur la collection.

Enfin les diverses interactions doivent faire l'objet d'une évaluation, dans la tradition des systèmes de Recherche d'Information. Une première évaluation concerne l'adéquation des tâches de synthèse, telles qu'exprimées dans les requêtes structurées, au besoin, tel qu'il est perçu par l'utilisateur. Impliquant des facteurs humains, elle ne peut être atteinte que par des études utilisateur. Ensuite, la qualité du processus de synthèse doit être mesurée, par une pertinence situationnelle. En tant que pertinence système, celle-ci peut faire l'objet d'un calcul. Mais ce calcul doit prendre en compte d'autres dimensions que les simples retour et précision de la Recherche d'Information. Enfin, la valeur du résultat présenté sous forme de document de synthèse doit être caractérisée. Estimée par un juge humain, l'utilisateur, elle ne peut elle aussi être déterminée que par des études utilisateur.

Surtout, cette exploration des diverses composantes du modèle de synthèse a permis d'identifier un certain nombre de difficultés, qui devront être abordées d'un point de vue technique pour opérationnaliser le modèle :

- ⋆ Choix de la méthode de résolution d'une sous-tâche élémentaire : ce choix repose sur une hypothèse d'identité d'interface entre composants résolvant la même sous-tâche élémentaire, pour éviter la gestion d'un problème de composition. Ce choix repose aussi sur une fusion entre informations issues des connaissances expérimentales, de la représentation de l'utilisateur et de la requête. Il s'agit donc d'un problème complexe qui induit la mise en place d'un système de sélection de la méthode de résolution,
- ⋆ Spécialisation du modèle de tâche : cette spécialisation du modèle pour une étude particulière, conduisant à la construction d'une instance de tâche, repose elle aussi sur une fusion d'informations entre connaissances du domaine, représentation de l'utilisateur et requête structurée. Là encore, il s'agit d'un problème complexe qui implique la prise en compte d'un ensemble de contraintes potentiellement contradictoires,
- ⋆ Coordination de l'exécution d'une instance de tâche : la construction du document de synthèse correspondant à une instance de tâche particulière implique l'exécution d'un ensemble de composants qui partagent des informations et sont dépendants entre eux. Il s'agit alors de mettre en place un système de coordination de l'exécution de chaque composant, permettant tout à la fois l'ordonnancement de l'exécution et l'échange de données. Il s'agit d'un problème de planification qui n'est pas trivial,
- ⋆ Résolution des sous-tâches élémentaires : le processus de synthèse induit l'exécution d'un ensemble de composants, résolvant chacun une sous-tâche élémentaire, dont certains doivent exécuter des opérations difficiles. Par exemple, pour les tâches de comparaison, l'organisation des individus et des groupes au sein d'une grille est un problème épineux.

La réalisation d'un prototype de système d'assistance à la synthèse, basé sur le modèle de Recherche d'Information orientée tâche, doit permettre l'évaluation opérationnelle du modèle et implique de se confronter à cet ensemble de difficultés qui ont été identifiées. C'est l'objet du prochain chapitre.

CHAPITRE

4 Opérationnalisation du modèle de synthèse par un prototype de système

Le chapitre précédent a permis la définition de bases conceptuelles pour un système d'assistance à la synthèse. La synthèse a été abordée comme une activité multifacettes, fédérant un ensemble de problèmes difficiles à résoudre. Ces considérations sur la synthèse ont conduit à la voir comme un problème de Recherche d'Information orientée tâche. La définition d'un modèle de synthèse a permis de replacer ce processus dans le cadre des modèles de Recherche d'Information courants. L'analyse plus fine des composantes de ce modèle a été envisagée comme un préalable à une implémentation d'un prototype. Cette implémentation fait l'objet de ce chapitre et implique tout d'abord la définition du contexte technologique, la transposition des éléments du modèle d'un cadre conceptuel vers un cadre opérationnel et la construction d'un prototype manipulant ces éléments en mettant en pratique le processus décrit.

4.1 Introduction

Le contexte applicatif des Tissue MicroArrays a permis l'identification d'un problème d'appréhension d'un gros volume de données par un chercheur. Ce problème intervient en préalable à une fouille de données ou pour replacer les informations disponibles dans une démarche expérimentale classique. La synthèse, en tant que solution proposée, a été abordée plus en détails au chapitre précédent. En particulier, des bases conceptuelles pour la synthèse ont été posées.

Mais tout modèle reste lettre morte s'il ne fait pas l'objet d'une validation pratique, soit d'une confrontation à la réalité. Cette validation, dans le cadre de la conception logicielle, passe par la construction d'un prototype de système opérationnalisant les propositions incluses dans le modèle, ainsi que le test de ce prototype.

L'opérationnalisation du modèle de synthèse au sein d'un prototype fonctionnel implique tout d'abord de poser le cadre dans lequel le prototype sera réalisé. Ce cadre peut être envisagé selon deux axes.

Tout d'abord, le modèle de synthèse proposé ne consiste qu'en une base conceptuelle. Celle-ci doit servir de support pour la mise en place d'une architecture conduisant au développement du prototype. Bien que la synthèse soit envisagée comme Recherche d'Information orientée tâche, le cœur de métier de la Recherche d'Information, basé sur une correspondance entre requête et documents, n'est qu'une étape parmi d'autres au sein du processus de synthèse. L'architecture envisagée est alors celle d'un système d'information composite, agrégant un ensemble de fonctionnalités dont des fonctionnalités de Recherche d'Information.

Ensuite, le prototype, envisagé dans un premier temps comme dédié au domaine des Tissue MicroArrays, est intégré dans un contexte de projet particulier, qui pose un certain nombre de contraintes techniques qu'il faut prendre en compte pour le développement. De plus, ce choix d'un domaine applicatif, s'il a permis de rendre le modèle de synthèse applicable, en limitant le champ couvert par le modèle, ne doit pas être trop limitatif. Une certaine généricité doit être permise, en se reposant sur des représentations de connaissances du domaine par exemple, pour limiter la dépendance du prototype au domaine applicatif.

Ces éléments de contexte étant mis en place, il s'agit ensuite de se baser sur l'architecture logique du système pour analyser plus en détails les différentes fonctions intervenant dans le système afin d'établir leur architecture logicielle, et les éléments manipulés au cours du processus afin de fixer leur forme exacte. Cette conception se doit de s'attacher aux difficultés qui ont été identifiées lors de la description du modèle : choix d'une méthode de résolution pour une sous-tâche élémentaire, spécialisation d'un modèle de tâche, coordination de l'exécution d'une instance de tâche

et résolution de sous-tâches élémentaires complexes. Le contexte de développement étant celui d'un prototype, une solution doit être proposée pour chacun de ces verrous techniques, même si elle reste simple.

Ces différentes composantes de cadre de développement et exploration plus fine du fonctionnement du prototype, en particulier en ce qui concerne les entités impliquées, font l'objet des prochaines sections.

4.2 Cadre de développement

4.2.1 Architecture logique

4.2.1.1 Introduction

Le développement d'un prototype de système d'assistance à la synthèse requiert la définition d'une architecture logique articulant une vue des fonctions internes du système, ainsi que des entités manipulées. La représentation envisagée ici est une représentation de haut niveau qui vise surtout à décrire un processus cohérent qui rende compte du modèle de synthèse, sans préjuger de l'implémentation réelle qui sera construite par la suite.

Cette architecture logique, présentée sous forme schématique Fig. 4.1, peut être décomposée en deux fonctions principales : l'instanciation d'une tâche et l'exécution de cette instance de tâche. Ces deux fonctions peuvent ensuite être décomposées en fonctions de niveau plus fin.

L'instanciation d'une tâche peut elle-même être découpée en deux sous-fonctions : la saisie de la requête et la spécialisation du modèle de tâche en tant que telle. La saisie ou formulation de la requête implique le remplissage d'un modèle de requête, choisi par l'utilisateur, en fonction de l'étude à réaliser. Cette formulation repose sur une représentation des connaissances du domaine d'étude. Ensuite intervient une spécialisation du modèle de tâche. Ce modèle est rendu opérationnel par association de chaque sous-tâche élémentaire du modèle avec un composant, en tenant compte de contraintes expérimentales issues de l'archétype de l'utilisateur et de connaissances expérimentales de niveau domaine. Le modèle est ensuite rempli avec des informations issues de la requête, des préférences de l'utilisateur et des connaissances du domaine, pour constituer une instance de tâche.

L'exécution de l'instance de tâche passe par l'exécution des composants correspondant à chaque sous-tâche élémentaire induite par la tâche de synthèse. Au niveau sélection, ces composants permettent l'extraction des éléments pertinents à partir

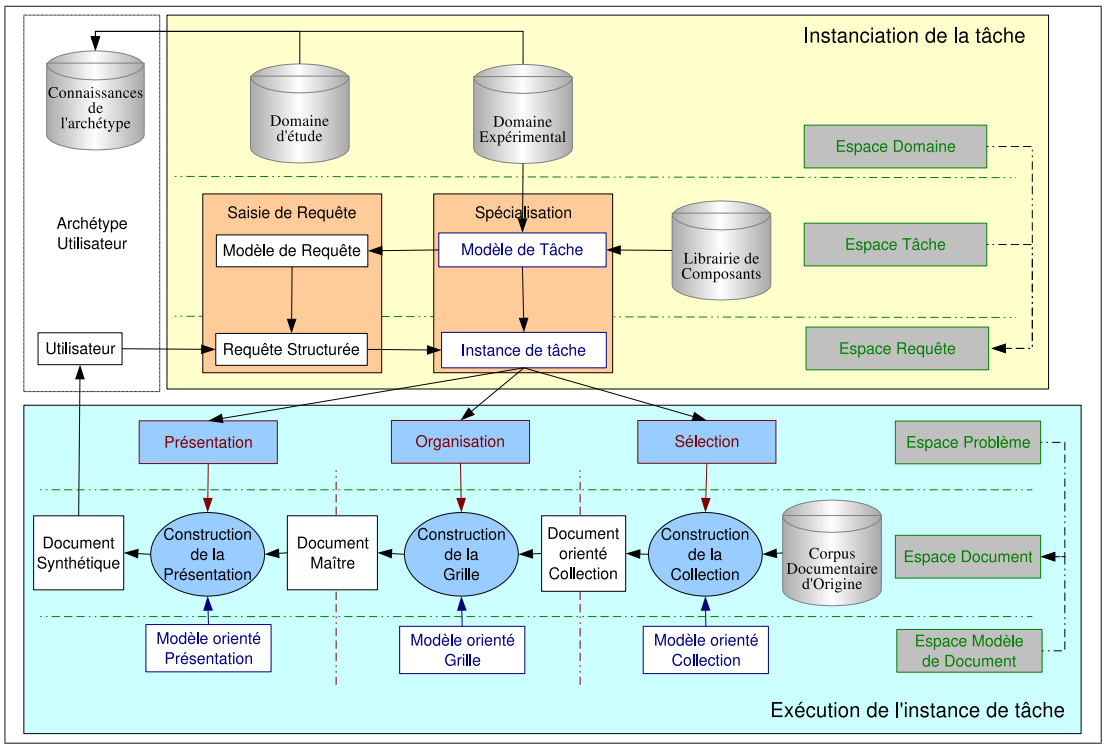


Fig. 4.1: Architecture logique - Le système d'assistance à la synthèse est décomposé en deux fonctions principales : l'instanciation d'une tâche et l'exécution de cette instance de tâche. Ces fonctions peuvent être ensuite décomposées en opérations plus fines et impliquent la manipulation d'un ensemble d'entités.

d'un corpus documentaire d'origine, pour construire un document orienté collection selon un modèle prédéfini. Au niveau organisation, le document orienté collection est transformé en document maître selon un autre modèle documentaire. Enfin, au niveau présentation, des paramètres et règles d'affichage spécifiques sont appliquées, pour proposer un document synthétique final.

Le bon déroulement de ces deux ensembles de fonctionnalités implique une prise en compte de l'utilisateur en tant qu'individu unique, en particulier pour permettre l'utilisation de ses préférences. Ceci sous-tend la mise en place, conjointement au système de synthèse en tant que tel, d'un système de gestion des utilisateurs, incluant tout à la fois un processus d'authentification et de gestion des préférences.

Les éléments de l'architecture sous-jacents à ces deux fonctionnalités sont décrits plus en détails dans le prochain paragraphe.

4.2.1.2 Décomposition de l'architecture

Comme il vient d'être indiqué, le processus de synthèse est envisagé comme un processus en deux phases, qui mène de la saisie d'une requête structurée par un utilisateur à l'affichage du résultat de la synthèse.

Étant donnée une requête structurée, l'instanciation d'une tâche implique la jointure de deux espaces au sein d'un troisième :

- ⋆ Espace Domaine : cet espace inclut les connaissances du domaine d'étude ainsi que les contraintes expérimentales, dont une fraction est accessible à chaque archétype utilisateur. Il fournit ainsi une image complète des entités du domaine applicatif à manipuler,
- ⋆ Espace Tâche : cet espace inclut la librairie de composants, c'est-à-dire des descriptions de résolution pour chaque sous-tâche élémentaire. Il regroupe les modèles de tâches, soit les représentations génériques décrivant comment résoudre une tâche prototypique, c'est-à-dire une catégorie de problèmes de synthèse particuliers. Il associe à chaque modèle de tâche un modèle de requête permettant la saisie. Il propose ainsi une vision des possibilités de manipulation d'entités issues du domaine applicatif,
- ⋆ Espace Requête : cet espace contient la requête structurée, issue de la spécialisation d'un modèle de requête par un utilisateur, ainsi que l'instance de tâche, construit par mise en relation d'une requête, de préférences et de connaissances. Il matérialise donc la jointure entre des éléments de l'Espace Domaine et de l'Espace Tâche.

Parallèlement, l'exécution de l'instance de tâche, issue de l'étape précédente, induit le même type de jointure, mais en mettant en relation des espaces différents :

- ⋆ Espace Problème : cet espace contient l'ensemble des composants spécialisés, correspondant à l'étude particulière considérée, organisés au sein d'une structure hiérarchique. Ces composants sont classés au niveau logique entre activités de sélection, d'organisation et de présentation. Il constitue donc une description du processus de synthèse à mettre en œuvre pour une étude particulière,
- ⋆ Espace Modèle de Document : cet espace regroupe des modèles de documents fournissant un squelette pour une représentation de l'information après chaque étape, une étape correspondant à la résolution d'un des sous-problèmes impliqués (sélection, organisation et présentation). Il s'agit donc d'un ensemble d'indications sur les éléments à construire,
- ⋆ Espace Document : cet espace présente le processus de synthèse en tant que tel, ainsi que les documents impliqués en cours de traitement, qu'ils soient source d'information ou résultat d'un traitement, de la collection d'origine au document synthétique, en passant par des documents intermédiaires. Il matérialise donc la jointure entre Espace Problème et Espace Modèle de Document.

L'ensemble fait intervenir une notion d'utilisateur, afin de permettre la personnalisation de l'interaction, ce qui induit la mise en place conjointe d'un système de gestion des utilisateurs, permettant l'individualisation des sessions au sein du système et la prise en compte de préférences.

Après cette vision descriptive des ensembles d'entités impliquées, il s'agit alors de s'intéresser au cadre technologique dans lequel cette architecture doit s'inscrire.

4.2.2 Cadre technologique

4.2.2.1 Introduction

La réalisation d'un premier prototype de système basé sur l'architecture qui vient d'être présentée s'inscrit dans un objectif à plus long terme. Le prototype doit permettre de valider la proposition qui a été faite, tout en posant des bases pour un futur système plus complet.

En conséquence, la conception du prototype doit prendre en compte un certain nombre de contraintes extérieures qui influencent le développement. Tout d'abord, la réalisation d'un système d'assistance à la synthèse s'intègre dans le cadre d'un projet de plate-forme associée à la technologie des Tissue MicroArrays. Ensuite, les futurs utilisateurs ont exprimé un certain nombre de souhaits et enfin, des choix personnels interviennent.

Ces ensembles de contraintes et choix techniques vont être exposés dans les prochains paragraphes.

4.2.2.2 Contexte du projet TMA-Explorer

La fin du Chapitre 1 a permis de mettre en exergue un des problèmes rencontrés dans l'informatisation de la technologie des TMA. Si nombre d'outils existent pour fournir une assistance à la plupart des étapes de la technique, il n'existe pas, à ma connaissance, de plateforme complète couvrant tout le processus, de la conception des blocs TMA à l'exploitation des données. La mise en place d'une plateforme intégrée dédiée à la technique des TMA est l'objectif de l'équipe RFMQ, avec le projet TMA-Explorer.

Les travaux présentés ici se placent dans le cadre de ce projet et le prototype de système d'assistance à la synthèse doit pouvoir s'intégrer au sein des outils déjà développés dans le cadre de cette plateforme. Cette intégration au sein de TMA-Explorer suggère tout à la fois un certain nombre de choix technologiques et la

réutilisation d'éléments déjà conçus ou développés.

La plate-forme a été conçue comme une application Web trois-tiers, accessible sur l'intranet du laboratoire, et à terme sur Internet. L'ensemble des données cliniques et histologiques, ainsi que les références aux images associées, sont stockées dans une base de données PostGreSQL, que j'ai conçue en collaboration avec Daniel Bret, durant son stage de dernière année d'ingénieur CNAM au sein de l'équipe [Bret, 2006]. Daniel Bret a aussi réalisé une application Web de gestion de données sous forme de pages JSP, tournant sur un serveur Apache Tomcat. Des ontologies du côlon et du sein pathologiques ont aussi été réalisées par le Dr. Joëlle Simony-Lafontaine, dans le cadre de sa thèse [Simony-Lafontaine, 2007].

Ces éléments impliquent d'utiliser la base de données existante comme corpus documentaire, et d'inclure une représentation formalisée partielle des ontologies du côlon et du sein pathologiques déjà réalisées à la représentation des connaissances du domaine. Ils induisent aussi le recours au langage JAVA pour le développement et à une architecture Web client/serveur.

4.2.2.3 Contraintes exprimées par les utilisateurs

Conjointement à ces contraintes liées à l'intégration du prototype à la plateforme TMA-Explorer, les futurs utilisateurs ont eux aussi exprimé des prérequis qui ont une influence sur les choix technologiques. En particulier, dans le cadre de la recherche scientifique, les préoccupations d'échanges d'informations et de réutilisation de résultats sont prédominants.

Or, dans le milieu scientifique, le langage XML a pris une place de premier plan en matière d'échange de données. Les modèles de données, sous forme de DTD ou schémas XML, permettant de représenter des entités spécifiques d'une discipline, comme SBML, pour représenter les réseaux moléculaires [Hucka et al., 2003], ou CellML, pour représenter des modèles mathématiques de fonctions biologiques [Lloyd et al., 2004], sont de plus en plus répandus. Dans le domaine des Tissue MicroArrays, le Tissue MicroArrays Data Exchange Specification [Berman et al., 2003], permet de décrire des blocs TMA et leur contenu.

De plus, le format XML est suffisamment simple pour que tout document XML puisse être compris par un lecteur humain sans grande expertise informatique. Il est aussi très structuré, ce qui facilite l'extraction d'informations par un parser simple, pour réutiliser le contenu dans d'autres applications.

Le recours à XML pour représenter le maximum d'entités impliquées dans le processus de synthèse paraît donc naturel.

4.2.2.4 Choix personnels

4.2.2.4.1 Introduction

L'objectif de développement est celui d'un premier prototype fonctionnel permettant d'évaluer les concepts qui ont été définis. Dans ce cadre, et en tenant compte des contraintes exposées précédemment, il s'agit de mettre en place un système simple, se basant sur des technologies simples.

Ce choix de simplicité va être traduit par la suite dans diverses composantes du système, telles que le format de représentation de connaissances, la technologie à utiliser pour les interfaces, ou l'architecture logicielle sous-jacente aux sous-tâches élémentaires.

4.2.2.4.2 Format de représentation de connaissances

Le système tel qu'envisagé inclut un ensemble de représentations pour des éléments aussi divers que des connaissances, des archétypes utilisateurs, des méthodes de résolutions problèmes, des modèles et des documents. Des spécifications et langages de représentations sont disponibles pour certaines : ontologies au format RDF [Noy et al., 2001] pour les connaissances du domaine ou UPML pour les méthodes de résolution de problèmes [Fensel et al., 2003].

Mais l'intégration d'une représentation à un format spécifique pour chaque catégorie d'entités implique un développement d'un système de lecture/écriture spécifique pour chaque format, ce qui devient rapidement difficile à gérer dans le cadre d'un prototype. J'ai donc fait le choix d'une représentation systématique au format XML, avec un schéma différent par type d'entités, schéma défini spécifiquement pour n'y inclure que les éléments essentiels dans le cadre du prototype.

4.2.2.4.3 Technologie pour les interfaces

Le système est proposé comme ayant une interface Web, ce qui implique de choisir une technologie pour la réalisation de cette interface. D'une part, les éléments existants de la plateforme suggèrent le recours à des pages JSP et à un traitement au sein d'un serveur Apache Tomcat. D'autre part, il a été choisi de représenter au maximum les éléments intervenant dans le système sous forme XML, ce qui inclut non seulement les modèles de requêtes et requêtes structurées, mais aussi les squelettes de formulaires permettant la saisie d'une requête à un format spécifique.

Cette dernière considération suggère le recours à la technologie XForms[1], standard basé sur XML actuellement développé par le W3C pour remplacer les formulaires HTML actuels, qui présente l'avantage d'une séparation de la structure et des données et d'une représentation des données saisies dans un formulaire directement au format XML.

Le problème malheureusement posé par cette technologie est sa jeunesse et son support dans les divers navigateurs Web existants, tels que Internet Explorer[2], Mozilla Firefox[3], Opera[4] ou Safari[5], reste encore très limité, même par l'intermédiaire de plug-ins. Par contre, plusieurs projets logiciels libres proposent des frameworks permettant le traitement des fichiers XForms au niveau serveur et la génération de code JAVA et JavaScript reproduisant ce qui aurait dû être généré par le client Web. Parmi ces frameworks, j'ai choisi d'utiliser Orbeon Forms[6], qui, après une exploration des solutions possibles, m'a paru le plus abouti.

4.2.2.4.4 Architecture logicielle sous-jacente aux sous-tâches élémentaires

La résolution de chaque sous-tâche élémentaire d'un modèle de tâche a été précédemment proposée sous forme de composants logiciels, architecture décrite par exemple dans [Szyperski, 2003]. Dans le contexte d'un développement en JAVA, l'architecture à composants classique est celle des JavaBeans[7], de Sun MicroSystems. Mais ce type d'architecture, qui implémente des assertions avancées de la notion de composants, paraît bien trop complexe, en rapport avec les besoins limités du prototype développé.

Par exemple, la communication par messages implique la définition d'un système de contrôle de cette communication. Par exemple, chaque composant doit être en mesure de déterminer quels messages lui sont destinés. De plus, tous les composants peuvent avoir une interface différente. Ceci nécessite une définition de chacune de ces interfaces et pose de gros problèmes de contrôle de la composition des composants, en particulier lorsqu'elle est dynamique, comme dans le cas du système envisagé.

Dans le cadre du prototype, j'ai fait le choix de mettre en place une architecture personnelle très simple, où tous les composants héritent d'une même classe abstraite commune, ce qui les dotent d'une même interface et de propriétés communes correspondant aux entrées/sorties. Ce qui est variable se limite alors aux traitements

[1] http ://www.w3.org/MarkUp/Forms/
[2] http ://www.microsoft.com/windows/products/winfamily/ie/default.mspx
[3] http ://www.mozilla-europe.org/fr/products/firefox/
[4] http ://www.opera.com/
[5] http ://www.apple.com/safari/
[6] http ://www.orbeon.com/
[7] http ://java.sun.com/products/javabeans/

réalisés.

Un tel choix limite les possibilités de réutilisation de composants tierce partie ou fige le système de communication, mais de telles contraintes semblent négligeables, étant donné que l'objectif n'est pas la construction d'une application complète à partir de briques de base qui seraient les composants, mais la composition dynamique d'entités logicielles correspondant chacune à la résolution de sous-tâches élémentaires très particulières. Étant donnée la spécificité de ces sous-tâches élémentaires, chercher des composants existants permettant de les résoudre, alors que les composants librement disponibles sont en général orientés vers des processus métier courants dans les entreprises, serait irréaliste. L'évolution du système n'induit alors pas la sélection et l'intégration de composants existants, mais passe par le développement spécifique de composants dérivant d'un modèle de base pour résoudre un problème unitaire particulier.

Un autre choix restrictif par rapport au fonctionnement de beaucoup d'architectures à composants courantes est le mode de communication des composants. Celle-ci a lieu en général par messages ou événements, ce qui induit la mise en place d'un système de contrôle des échanges d'informations complexe, en général prévu pour fonctionner à travers un réseau.

Or, dans le cadre d'un système d'assistance à la synthèse, chaque composant peut avoir plusieurs paramètres, qui correspondraient à la réception de plusieurs messages, et induiraient la mise en place, de façon interne au composant, d'un système complexe de rétention de messages et d'activation du composant uniquement quand tous les messages ont été reçus. De plus, les divers composants sont pour l'instant envisagés comme localisés au sein d'un même serveur et les besoins de communication se limitent à un partage de résultats de traitements. La communication par l'intermédiaire d'un tableau noir réduit la complexité du système de communication tout en satisfaisant les besoins d'échanges de données.

En effet, un système basé sur le principe du tableau noir, tel que présenté par [Corkill, 1991] par exemple, inclut tout d'abord un ensemble de sources de connaissances, c'est-à-dire des entités logicielles permettant de résoudre le problème, ici les composants. Ensuite, le tableau noir est une zone de stockage de données contenant des données en entrée, des solutions partielles et autres, et les sources de connaissances interagissent par des modifications apportées dans le tableau noir. Enfin, un système de contrôle prend les décisions d'exécution et d'accès aux ressources. Ici, ce processus externe est en charge du contrôle de l'activation, et n'initialise un composant que quand tous ses paramètres sont définis dans le tableau noir, ce qui le dote de capacités d'ordonnancement ou planification.

Une telle architecture présente l'avantage d'éviter d'inclure un contrôle d'activation au sein de chaque composant, tout en satisfaisant le besoin d'échange d'informations et permettant un contrôle dynamique de l'exécution.

4.2.3 Un contexte bien défini pour le développement

Cette première section a permis de poser un cadre fonctionnel et technique en vue du développement d'un prototype de système.

Au niveau fonctionnel, elle a présenté une architecture logique reflétant les concepts introduits par le modèle de synthèse, architecture qui a permis la décomposition du système en quelques fonctions ou étapes du traitement principales : saisie de requête, instanciation de la tâche et exécution de l'instance de tâche, en vue de la présentation du document de synthèse résultant à l'utilisateur. Les entités impliquées à chaque phase ont aussi été recensées. La mise en place d'un système de gestion des utilisateurs est aussi apparue comme nécessaire.

Au niveau technique, elle a permis de lister les choix technologiques sous-jacents au développement. L'intégration du prototype au sein du projet de plateforme TMA-Explorer implique la réutilisation d'éléments existants comme la base de données et l'incorporation du prototype au sein d'une application Web développée en JAVA. Le souci d'échange et réutilisation des données des utilisateurs suggère le recours massif à un format XML. Le contexte d'un prototype qui se veut très simple induit le choix personnel de représentations de toutes les entités impliquées au même format, la construction d'une interface utilisant la technologie XForms et le framework Orbeon Forms et le développement spécifique d'un framework composants.

Ce contexte étant posé, il s'agit maintenant de s'intéresser aux détails pratiques du fonctionnement du prototype et de représentation des entités impliquées dans le système. L'exploration de l'architecture logicielle du prototype va être réalisée dans le reste de ce chapitre selon deux axes.

Tout d'abord, la construction d'un document de synthèse implique la résolution d'une tâche particulière dans un domaine applicatif précis. Ces notions de tâche et domaine étant centrales, leurs représentations vont être évoquées en premier lieu.

Puis le processus de synthèse induit la manipulation de ces deux concepts principaux par les différentes fonctions identifiées dans l'architecture logique, fonctions qui vont être détaillées par la suite : gestion des utilisateurs, saisie de requête, instanciation de la tâche, exécution de l'instance de tâche, présentation du document de synthèse.

A des fins d'illustration, cette présentation est envisagée dans le cadre d'une étude concrète réalisée dans le domaine des Tissus MicroArrays par un utilisateur hypothétique. Le même exemple d'étude que celui du chapitre précédent va être utilisé : la «comparaison du pourcentage de cellules marquées entre les différentes molécules étudiées, entre tissus tumoraux et adjacents à la tumeur, en fonction de la localisation intracellulaire du marquage, chez les patients atteints d'un cancer du

côlon».

4.3 Des notions centrales à manipuler

4.3.1 Tâche

4.3.1.1 Introduction

Le Paragraphe 3.4.2.1 a permis une description générale de la notion de tâche telle qu'envisagée dans le modèle de synthèse. La synthèse, dans le contexte applicatif des TMA, regroupe un ensemble de problèmes, ou tâches prototypiques, organisés au sein d'une taxonomie, relevant de trois catégories principales : comparaison, évolution et distribution. Chaque tâche prototypique est aussi envisagée comme une tâche composée, conduisant à la définition d'un modèle de tâche sous forme d'une hiérarchie de sous-tâches élémentaires. La Fig. 3.10 a permis de donner un aperçu conceptuel d'un tel modèle de tâche pour une tâche de type comparaison.

Se pose alors le problème de la représentation fonctionnelle de ce modèle de tâche. L'analyse des besoins pour cette représentation conduit alors à une formalisation qui va être présentée ici.

4.3.1.2 Prérequis et contraintes pour un modèle de tâche

Le cadre technologique établi ainsi que les spécificités du modèle de tâche introduites au chapitre précédent posent un ensemble de contraintes sur le modèle de tâche, aussi bien en ce qui concerne sa forme que le fond.

Au niveau forme, le cadre technologique qui a été défini induit une représentation au format XML. Un modèle de tâche est donc un fichier XML.

En ce qui concerne le fond, chaque tâche prototypique est envisagée comme une hiérarchie de sous-tâches élémentaires, qui doit être reflétée par le modèle de tâche. Le format XML se prête bien à la représentation de telles hiérarchies, sous forme d'une structure de balises imbriquées. Au niveau le plus fin, doivent alors être représentées les sous-tâches élémentaires.

Chaque sous-tâche élémentaire est une entité unique, qui doit être spécialisée pour permettre la résolution d'un problème de synthèse particulier, et qui conduit à un résultat partiel spécifique. Ceci pose un problème de définition de l'interface et

des entrées et sorties de chaque sous-tâche élémentaire. Dans le cadre du prototype, elles sont considérées comme prédéfinies en ce qui concerne leur nombre, leur nature et leur format.

De plus, le traitement correspondant à chaque sous-tâche peut être réalisé par plusieurs composants. Au sein du prototype, les prérequis sur les entrées et sorties sont considérés comme satisfaits par tous les composants permettant de résoudre une sous-tâche particulière. Ceci permet de faire abstraction du problème de composition, qui intervient dès que des entités logicielles unitaires doivent être composées pour accomplir une tâche plus complexe. Ainsi, la définition d'un en-tête (interface et entrées/sorties) pour chaque sous-tâche élémentaire évite la mise en place d'un système complexe de contrôle de cohérence entre interfaces et résultats de composants d'une part et descriptions de sous-tâches élémentaires d'autre part.

Il s'agit alors de définir un format de description pour les diverses composantes essentielles d'une sous-tâche élémentaire. Une sous-tâche étant unique, celle-ci requiert un identifiant. De plus, la communication entre composants étant envisagée par le biais d'un tableau noir, le contrôle de l'exécution est reporté hors des composants et les seules informations pertinentes restent les entrées et les sorties.

Les entrées consistent en une liste de paramètres. Pour chaque sous-tâche, leur nombre et leur rôle dans le traitement est connu, par contre leur contenu est variable, puisqu'il dépend de la tâche de synthèse en cours. De plus, la source de la valeur de chaque paramètre est elle aussi variable.

En effet, certains paramètres ont des valeurs spécifiées de manière obligatoire au niveau de la requête, puisqu'ils correspondent à une spécification de l'étude à réaliser. On peut par exemple citer des critères de sélection, dans l'exemple d'étude considéré dans ce chapitre : la localisation de la tumeur dans le côlon. D'autres paramètres consistent en le résultat d'autres sous-tâches élémentaires, et se trouvent nécessairement au sein du tableau noir. Par exemple, la sous-tâche qui attribue des individus à des groupes utilise la liste des individus sélectionnés et la liste des groupes constitués qui ont été écrites dans le tableau noir au cours de traitements préalables.

D'autres paramètres, par contre ne sont pas nécessairement définis dans la requête. Ainsi, les contraintes géométriques sur la grille, qui induisent par exemple une réduction de la population sélectionnée, sont optionnelles dans la requête, et des sources alternatives pour ce paramètre doivent être définies. Celles-ci consistent en général en un élément des préférences de l'usager ou en connaissances expérimentales.

D'autres paramètres, enfin, doivent pouvoir être définis à partir d'une combinaison de sources. Par exemple, la sous-tâche élémentaire de gestion des données manquantes comprend un paramètre correspondant à la liste des éléments du dossier clinique ou des données histologiques qui sont utilisés dans la requête. Cette liste est

constituée par combinaison de plusieurs parties de la requête structurée : critères de sélection, de groupement, de tri.

En ce qui concerne les sorties, celle-ci consistent nécessairement en une écriture du résultat du traitement au sein du tableau noir, et ne nécessitent qu'une indication sur la zone du tableau noir dans laquelle chaque résultat sera écrit.

4.3.1.3 Structure du modèle de tâche

Les diverses contraintes qui viennent d'être décrites ont conduit à la définition d'un format de fichiers XML pour un modèle de tâche, dont un exemple partiel et une description plus précise sont fournis en Annexe A. La Fig. 4.2 fournit une vue schématique sur l'organisation fonctionnelle de ce fichier.

Celui-ci inclut tout d'abord une hiérarchie de catégories, reflétant la décomposition de la tâche prototypique jusqu'au niveau des sous-tâches élémentaires. Au niveau le plus fin est décrite chacune des sous-tâches élémentaires. Chacune de ces sous-tâches élémentaires est identifiée par son chemin dans la hiérarchie et surtout un nom. Elle présente ensuite des indications pour ses entrées et sorties. Au niveau entrées, chaque paramètre est identifié par un nom et indique la source à utiliser pour sa spécialisation. Cette source peut consister en une liste de sources alternatives ou une liste d'éléments à combiner.

4.3.2 Connaissances du domaine applicatif

4.3.2.1 Introduction

Ainsi qu'il a été présenté au chapitre précédent, les connaissances du domaine jouent un rôle central dans les traitements réalisés dans le cadre d'une tâche particulière. Celles-ci peuvent être divisées en deux catégories.

Les connaissances du domaine d'étude représentent les concepts qui peuvent être manipulés dans le cadre d'une synthèse et sont instanciés au sein des dossiers cliniques et données histologiques. En tant que telles, elle servent de base à la définition des connaissances des archétypes utilisateur et à la formulation de requête.

Les connaissances expérimentales permettent la définition de contraintes sur les sous-tâches élémentaires. Ces contraintes sont de deux types : contraintes sur le type de traitement à réaliser pour une sous-tâche particulière ou contraintes sur les paramètres du traitement. Ces connaissances expérimentales peuvent aussi être envisagées à plusieurs niveaux : de manière générique, au niveau du domaine appli-

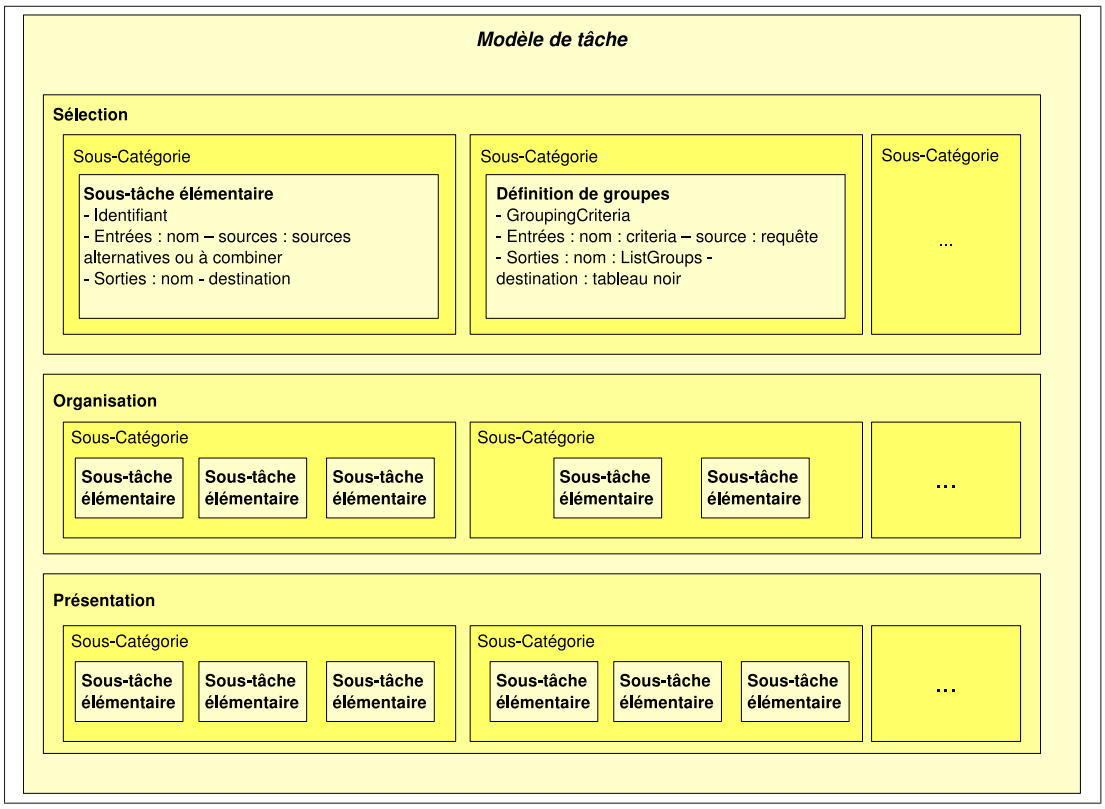


Fig. 4.2: Décomposition fonctionnelle d'un modèle de tâche - Chaque fichier XML de modèle de tâche est décomposé en trois parties principales correspondant aux trois catégories de problèmes à résoudre. Chacune contient une hiérarchie de catégories jusqu'au niveau sous-tâche élémentaire. Chacune de ces sous-tâches élémentaires est alors décrite, et en particulier les éléments de spécialisation, c'est-à-dire les entrées et sorties.

catif ; de manière personnalisée, en tant que comportements idiosynchratiques d'un groupe d'usagers ou d'un utilisateur particulier ; ou de manière spécifique, dans le cadre d'une étude particulière. Il s'agit donc d'un ensemble de contraintes à prendre en compte lors de l'instanciation d'une tâche.

La représentation formelle de ces connaissances est donc de première importance et va être détaillée ici.

4.3.2.2 Connaissances du domaine d'étude

4.3.2.2.1 Taxonomie

En ce qui concerne le domaine d'étude, la représentation de connaissances doit recouvrir tout à la fois les connaissances cliniques sur les patients et les problématiques biologiques, en particulier des informations anatomiques. Au niveau dossier médical, les systèmes de gestion informatisés sont de plus en plus nombreux, et des systèmes

tels que celui de [Patel et al., 2006], qui placent cette problématique dans le contexte des banques de tissus, contexte qui est celui des données associées aux TMA, sont de bonnes sources d'inspiration pour une représentation de connaissances.

Au niveau biologique, alors qu'il existe une ontologie de l'anatomie humaine pour des organes sains [Rosse and Mejino, 2003], les ontologies de tissus pathologiques ne sont pas très satisfaisantes et des ontologies du côlon et du sein pathologiques ont été construites par le Dr. Joëlle Simony-Lafontaine, dans le cadre de sa thèse [Simony-Lafontaine, 2007]. La taxonomie correspondante pour le côlon est présenté Fig. 4.3.

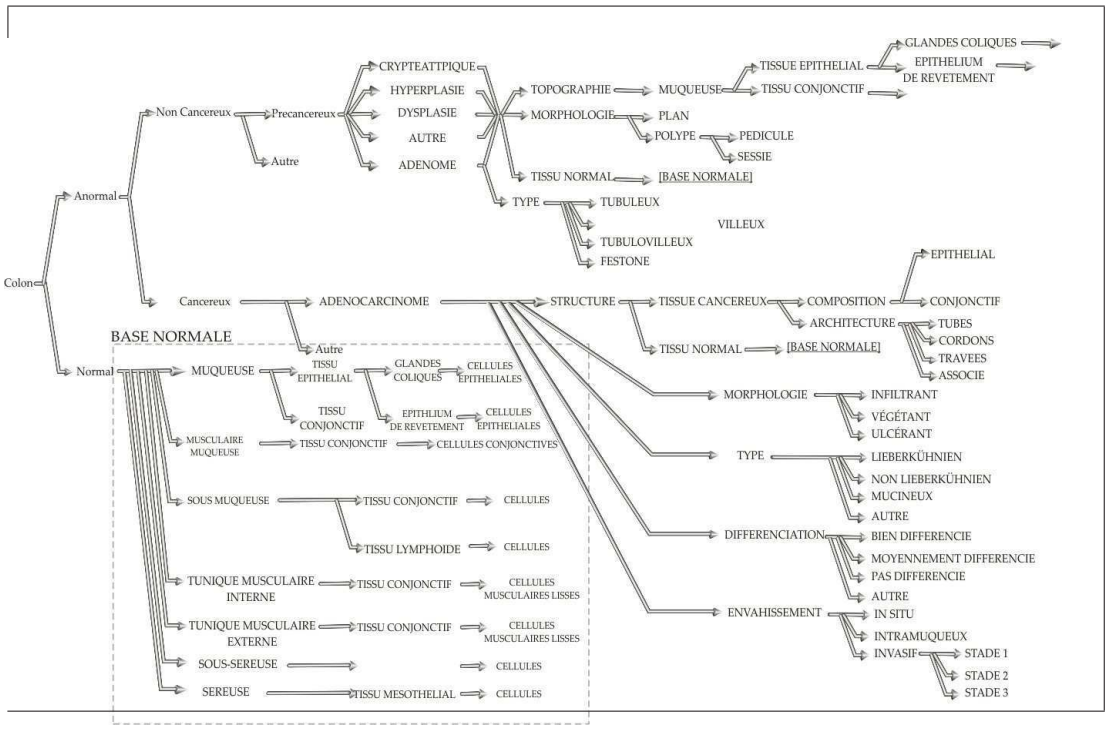


Fig. 4.3: Taxonomie de l'ontologie du côlon - Construite par le Dr. Joëlle Simony-Lafontaine, cette taxonomie reflète les diverses pathologies du côlon, de l'échelle de l'organe à l'échelle cellulaire.

La mise en relation des connaissances qui viennent d'être décrites et la base de données existante au sein du projet TMA-Explorer a permis la définition d'une taxonomie du domaine d'étude.

Cette taxonomie a été envisagée dans un objectif purement applicatif : elle se limite à organiser logiquement, au sein d'une hiérarchie de profondeur raisonnable, des entités présentes dans la base de données qui ont été jugées comme essentielles pour le prototype. Cette taxonomie, reflétée matériellement par une arborescence de répertoires au sein du système de fichiers du serveur, est présentée Fig. 4.4.

Les différents concepts présents au sein de cette taxonomie doivent alors être représentés de manière satisfaisante dans le cadre du prototype, ainsi que présenté dans le prochain paragraphe.

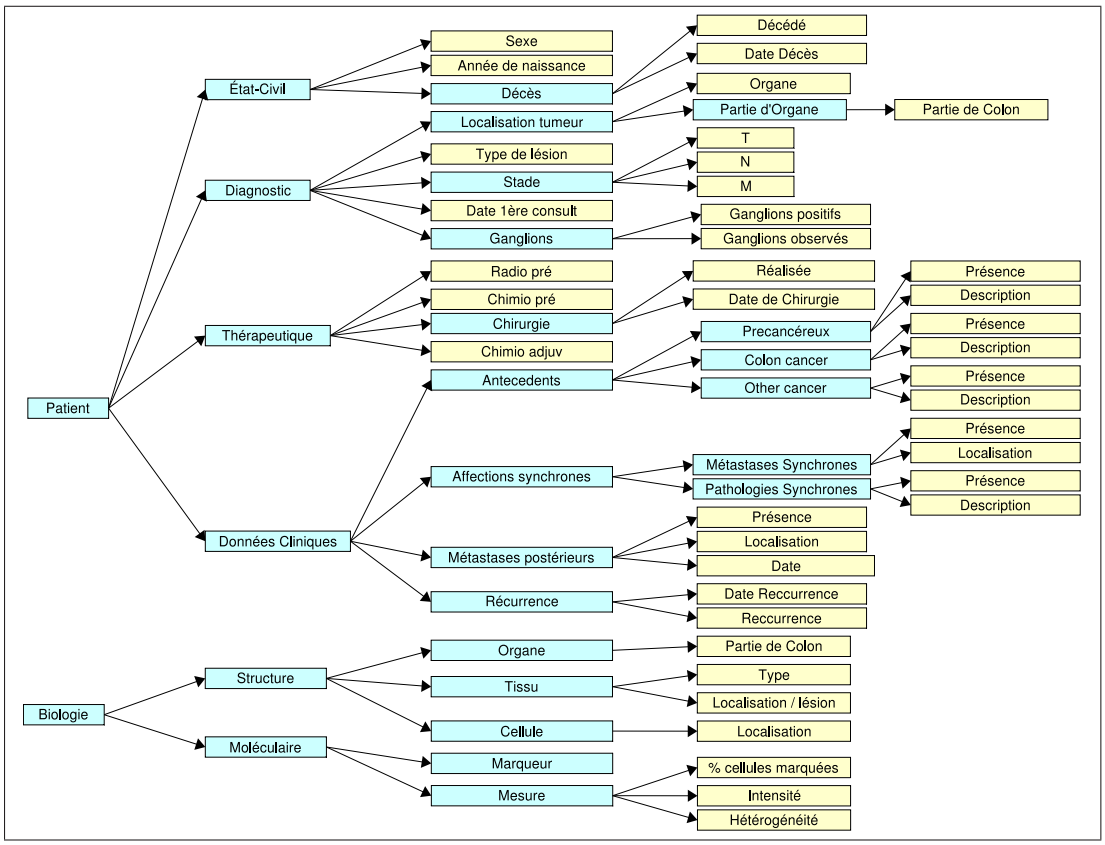


Fig. 4.4: Taxonomie du domaine d'étude - Les nœuds sont représentés physiquement par des répertoires, et les feuilles des fichiers XML au format spécifique. Cette taxonomie couvre les champs essentiels de la base de données.

4.3.2.2.2 Représentation de concepts

Les concepts correspondant à des catégories de la taxonomie du domaine d'étude ne requièrent pas une représentation plus étendue qu'un nom et une description, puisqu'ils jouent un simple rôle d'organisation. Par contre, les concepts feuille sont beaucoup plus complexes, car ce sont ceux qui sont manipulés par le processus de synthèse. En particulier, ils peuvent avoir des natures très différentes, impliquant un traitement différent dans le cadre du processus de synthèse. De plus, ils correspondent à des éléments en base de données, ce qui implique la définition d'une correspondance avec un champ particulier dans la base.

Au niveau de la nature des éléments impliqués, les différences d'échelle (du patient à la cellule), de type (nombres, textes, images) ont déjà été évoquées. Dans le cadre du processus de synthèse, c'est surtout le type qui est problématique.

En effet, la caractérisation d'un pourcentage de cellules marquées, soit une valeur numérique entre 0 et 100, ne peut pas être réalisée de la même façon que la description d'un type de lésion, qui correspond à une liste prédéfinie de valeurs textuelles.

Tab. 4.1: Diverses natures de concepts du domaine d'étude - Ce tableau répertorie les diverses natures de concepts telles qu'envisagées au sein du prototype. En plus d'un exemple, il fournit des contraintes sur les opérateurs et valeurs possibles pour chaque type de concept.

Nature du concept	*Exemple*	*Opérateurs possibles*	*Valeurs possibles*
Nœud	État-Civil	=;!=	liste prédéfinie de valeurs textuelles correspondant aux concepts de niveau plus fin
Quantitatif	Pourcentage de cellules marquées	>; <; =>; <=; =;!=	valeurs numériques comprises entre un minimum et un maximum
Qualitatif	Type de lésion	=;!=	liste prédéfinie de valeurs textuelles
Booléen	Décédé	=;!=	vrai ou faux
Date	Date de naissance	>; <; =>; <=; =;!=	dates comprises entre un minimum et un maximum
Texte libre	Commentaire	contient()	n'importe quel fragment de texte
Image	Image d'un spot	=;!=	nom du fichier image

Dans le premier cas, on peut définir des caractéristiques de type valeur minimum, valeur maximum, ou taille de classe ; dans le second, seule la liste des valeurs possibles, éventuellement ordonnée, a un sens.

De plus, les opérations possibles sont différentes. Ainsi, si l'on considère une sous-tâche élémentaire de sélection selon un critère donné, si le critère est le pourcentage de cellules marquées, des contraintes telles que «inférieur à 50» ou «supérieur ou égal à 20» ont un sens. Par contre, elles sont sans objet pour le type de lésion.

Ces considérations impliquent tout d'abord d'évaluer une liste des diverses natures de concepts, et des contraintes qu'elles posent sur des éléments tels que les opérateurs et valeurs possibles. Dans le cadre de la synthèse, cette évaluation a permis la construction du Tab. 4.1.

Dans le cadre du prototype, aucune «vraie» date n'est présente, toutes les dates étant définies en tant qu'années, qui peuvent être considérées comme des éléments quantitatifs. Les éléments booléens peuvent être considérés comme des éléments qualitatifs à deux valeurs, vrai et faux. En ce qui concerne le texte libre, la mise en place d'un opérateur de type «contient()» induit le recours à des méthodes de Recherche d'Information en tant que telles, qui sont très complexes à mettre en œuvre et ne sont pas l'objet de ma thèse. Ces types de concepts n'ont donc pas été pris en compte de manière spécifique au sein du prototype.

Il s'agit alors de proposer, pour chaque type de concept, une représentation particulière. Ainsi qu'il a été défini dans le cadre de développement, celle-ci doit être réalisée au format XML et induit la description des éléments nécessaires et suffisants pour leur traitement.

Ainsi, pour tous les éléments il faut définir une valeur par défaut, à utiliser quand seul le concept est mentionné sans choix d'une valeur spécifique. Pour les concepts qui ne sont pas des nœuds, il s'agit aussi d'indiquer une correspondance avec un champ de base de données.

Ensuite, les éléments quantitatifs doivent indiquer des valeurs minimum et maximum possibles, ainsi que des tailles de classes possibles, sur la base desquelles seront construits des groupes. Les éléments qualitatifs, quant à eux, doivent spécifier la liste de leurs valeurs possibles.

Un exemple de fichier XML, pour un élément qualitatif, la notion d'organe, est présenté en Annexe B.

4.3.2.3 Connaissances expérimentales

4.3.2.3.1 Introduction

Comme indiqué précédemment, les connaissances expérimentales peuvent être envisagées sous deux formes : contraintes sur les paramètres de certaines sous-tâches élémentaires ou contraintes sur le type de traitement à réaliser. Celles-ci ont été définies comme connaissances expérimentales par analogie avec les méthodes mises en place dans le contexte applicatif de construction de TMA réels, et pourraient être envisagées comme des éléments de définition de protocole expérimental.

Classiquement, et en particulier dans les publications de sciences expérimentales telles que la biologie, ces indications de protocole sont considérées en matière de matériel (les éléments utilisés, par exemple échantillons biologiques, réactifs, appareils, etc.) et méthodes (les procédures expérimentales mises en place, en terme de plan d'expérience, par exemple liste et durées des bains pour une coloration). Ici «matériel» correspondrait aux contraintes sur les paramètres et «méthodes» aux contraintes sur les méthodes de résolution.

Ces deux types de connaissances expérimentales vont donc être explorées plus en détails ici.

4.3.2.3.2 Connaissances de type «Matériel»

Dans le processus réel de fabrication de TMA, le matériel inclut entre autres des considérations comme la taille de l'aiguille utilisée pour prélever des carottes ou la taille du bloc de paraffine receveur, qui conditionnent la taille de la matrice lignes/colonnes du bloc TMA à construire. Des indications sur l'épuisement du bloc

donneur, qui empêche son utilisation dans un nouveau bloc car aucune nouvelle carotte ne peut physiquement être prélevée, ou sur l'âge du matériel, permettant de prendre en compte une influence du traitement subi par le bloc sur le marquage, relèvent aussi de ce type de problématique.

Par analogie, dans le cadre de la synthèse appliquée au domaine des TMA, sont considérés à ce niveau des éléments tels que la taille de la grille du document de synthèse, la langue ou le code couleur à utiliser au sein de la grille.

En ce qui concerne le fonctionnement du prototype, ces éléments interviennent en tant que paramètres de sous-tâches élémentaires particulières. Ainsi, la largeur et la longueur de la grille sont utilisées par une sous-tâche élémentaire qui définit l'adéquation de la grille au nombre total d'individus sélectionnés.

Se pose alors le problème de la représentation de tels éléments. Bien qu'ils soient conceptuellement différents des éléments inclus dans la taxonomie du domaine d'étude, ils n'en diffèrent guère par nature et peuvent être représentés de la même façon, c'est-à-dire au sein d'une hiérarchie de répertoires contenant des fichiers XML de description d'éléments feuille.

4.3.2.3.3 Connaissances de type «Méthodes»

En ce qui concerne les éléments de type méthode, le problème est plus complexe. En effet, ceux-ci sont définis comme des contraintes sur le traitement à réaliser dans le cadre d'une sous-tâche élémentaire, ce qui implique de décrire plus précisément ce qui est entendu par là, avant de proposer un mode de représentation.

Au sein du prototype, chaque sous-tâche élémentaire est résolue au niveau logiciel par un composant. Mais cette affirmation ne présuppose pas l'unicité du composant associé à chaque sous-tâche, et la relation peut être envisagée comme une relation de un à plusieurs. Or le processus de synthèse a été conçu au sein du prototype comme résultant de l'exécution d'un composant pour chacune des sous-tâches élémentaires du modèle de tâche.

Se pose alors un problème de choix de composant à exécuter pour chacune des sous-tâches élémentaires, déjà illustré au chapitre précédent par la Fig. 3.15. Ce problème, en l'absence de connaissances complémentaires, est souvent résolu par un choix arbitraire ou au hasard. Or ce type de choix n'est pas adapté dans le contexte scientifique de l'appréhension des données TMA, puisqu'il limite la reproductibilité. Proposer une résolution qui ne repose plus sur le hasard est l'objet des connaissances expérimentales de type protocole.

Ces connaissances doivent alors permettre la représentation de règles d'associa-

tion entre une sous-tâche élémentaire et un composant spécifique, choisi parmi ceux permettant de résoudre ce sous-problème particulier. Dans le cadre du prototype, le système envisagé est pour l'instant très simple, puisqu'il constitue en la définition d'un composant par défaut pour chaque sous-tâche élémentaire.

Ainsi, par exemple, la sous-tâche de gestion de données manquantes est réalisée par défaut par exclusion des individus ayant des données manquantes. La flexibilité est apportée par la possibilité d'écraser ce choix par défaut, au sein de la requête ou au niveau utilisateur. Ainsi, certaines études peuvent être réalisées spécifiquement en recourant à des inférences pour déterminer des valeurs de données manquantes.

A l'avenir, des mécanismes plus complexes peuvent être envisagés. Par exemple, une pondération de l'utilisation de chaque composant, en fonction du type d'étude qui est traité, peut être mise en place.

Ainsi, on peut imaginer, au cours d'une séance d'utilisation, le raffinement progressif de l'étude d'un problème particulier. Dans un premier temps, la gestion des données manquantes par inférence permet d'inclure plus d'individus et ainsi d'avoir une vision plus globale du problème. Ensuite, l'exclusion des données manquantes permet de construire une vue sur des données fiables, qui peut servir de base à une publication ou à d'autres analyses, par exemple statistiques.

Un tel comportement peut être mis en place par le biais d'un système à base de règles, en liant par exemple le composant de gestion de données manquantes à utiliser au nombre de reformulations de la requête. Un système à partir de cas peut lui aussi s'avérer pertinent.

4.4 Gestion des utilisateurs

4.4.1 Introduction

Les notions de tâche et connaissances du domaine ont été abordées dans la section précédente en détails, afin de proposer une représentation adaptée au contexte fonctionnel et technique du prototype. Mais celles-ci sont sans objet sans une description des fonctionnalités manipulant ces entités.

Or, en préalable à toute construction d'un document de synthèse, intervient une problématique liée aux utilisateurs. En effet, afin de permettre une personnalisation de l'interaction avec le système d'assistance à la synthèse, ce dernier doit inclure un système de gestion des utilisateurs. Celui-ci se décompose en deux fonctionnalités complémentaires.

Premièrement, un système d'authentification doit permettre de définir quel utilisateur est en train d'utiliser l'application. De plus, dans le cadre du système envisagé, il doit permettre d'associer chaque utilisateur à son archétype, c'est-à-dire à un groupe d'usagers ayant la même vue partielle sur la taxonomie du domaine d'étude et un même jeu de connaissances expérimentales correspondant à leurs comportements idiosynchratiques.

En pratique, dans le cadre de ma thèse, cette notion d'archétype utilisateur et son incidence sur la session avec le système, en particulier au niveau de l'accès aux concepts du domaine d'étude, n'a pas été suffisamment analysée pour permettre une implémentation fiable. Le système d'authentification du prototype se limite donc à l'identification de l'usager. Afin de réaliser un tel système, il faut disposer d'une liste de couples Nom d'utilisateur/Mot de passe, d'une interface de saisie, correspondant à la page d'accueil du système, ainsi que d'un processus de mise en correspondance des informations saisies avec la liste des utilisateurs.

Deuxièmement, au sein de l'application, un système de gestion des préférences doit permettre à l'utilisateur de définir son profil particulier. Celui-ci requiert un fichier de préférences individuelles, qui est mis à jour après saisie de nouvelles préférences dans un formulaire spécifique. Ce formulaire doit permettre la définition de valeurs par défaut pour des éléments issus des connaissances du domaine d'étude ou la définition de contraintes expérimentales.

La suite de cette section va consister en une présentation succincte de l'architecture logicielle mise en place pour supporter cette gestion des utilisateurs, celle-ci n'étant pas au cœur des problématiques posées par le processus de synthèse.

4.4.2 Architecture logicielle

Au sein du prototype développé, les deux fonctionnalités d'authentification et gestion des préférences sont supportées par des entités logicielles basées sur l'architecture présentée Fig. 4.5. Étant donné qu'ils s'agit d'opérations interactives, le framework Orbeon Forms intervient en position centrale dans les deux cas.

En ce qui concerne l'authentification, le formulaire de connection à l'application, construit au sein du framework Orbeon à partir d'un fichier XForms, permet à l'usager de saisir un nom d'utilisateur et un mot de passe. Le processus d'authentification en tant que tel est mené de façon interne au framework, qui gère aussi la persistance des connections, et se base sur un fichier listant les utilisateurs autorisés. L'utilisateur connecté accède alors à une page d'accueil spécifique permettant d'accéder au processus de saisie de requête et présentant la liste des anciennes requêtes.

Parmi les fonctionnalités disponibles une fois l'utilisateur authentifié se trouve le

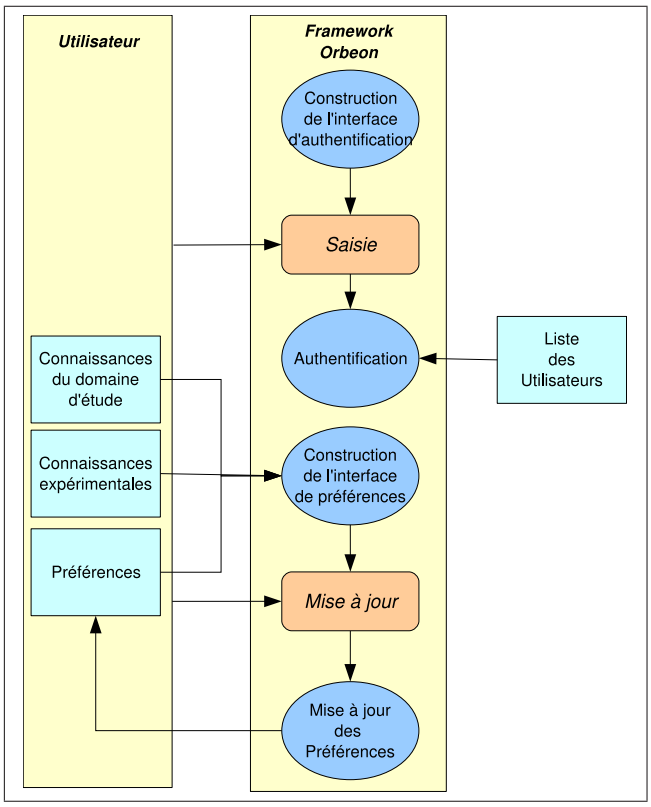


Fig. 4.5: Gestion des utilisateurs - Basée principalement sur des formulaires construits au sein du framework Orbeon Forms, la gestion des utilisateurs inclut une fonctionnalité d'authentification (en haut) et une fonctionnalité de gestion des préférences (en bas).

système de gestion des préférences. Le formulaire de saisie est lui aussi construit au sein du framework Orbeon à partir d'un fichier XForms spécifique. Ce formulaire, à partir du fichier de préférences existant pour l'utilisateur, construit une liste des préférences déjà définies, en permettant leur modification, et propose l'ajout de nouvelles préférences.

Dans le cadre du prototype, ces préférences permettent la définition de valeurs spécifiques à l'utilisateur pour des connaissances du domaine d'étude. Par exemple, l'utilisateur peut spécifier que sa taille préférée pour les classes d'âges est de 10 ans.

Les préférences utilisateur incluent aussi des contraintes expérimentales. En ce qui concerne les contraintes de type «matériel», il s'agit de valeurs par défaut, comme pour les connaissances du domaine d'étude. Pour les contraintes de type «méthodes», il s'agit d'associations d'une sous-tâche élémentaire avec un composant particulier.

4.5 Saisie de requête

4.5.1 Introduction

Comme les systèmes de Recherche d'Information classiques, le prototype, basé sur un modèle de Recherche d'Information orienté tâche, comprend une entrée dans le système de traitement par une requête, ici proposée comme une requête structurée. La forme de cette requête a été précédemment spécifiée comme dépendant de la tâche de synthèse considérée, conduisant à la définition d'un modèle de requête par modèle de tâche. Ce modèle doit être spécialisé par l'utilisateur en fonction de l'étude particulière qu'il veut réaliser.

Ces observations sur la formulation de requête nécessitent de considérer un système interactif, permettant à l'utilisateur de choisir la tâche prototypique qui l'intéresse puis de spécialiser le modèle de requête correspondant, à partir des connaissances du domaine. Cette requête spécialisée doit alors être stockée pour un traitement futur.

Ce processus, mis en place au sein de l'architecture présentée dans le prochain paragraphe, pose un problème cognitif de formulation.

4.5.2 Architecture logicielle

La construction de la requête, dont l'architecture logicielle est présentée Fig. 4.6, est réalisée au sein du framework Orbeon Forms. Celui-ci permet la construction d'interfaces, à partir d'un fichier XForms, dont une partie est basée sur un modèle de requête spécifique de la tâche sélectionnée par l'utilisateur.

Le modèle de requête est un fichier XML qui liste les composantes de la requête qui sont particulières à la tâche considérée et pour lesquelles des saisies doivent être réalisées. En définitive, il liste les rôles inclus dans la requête. A chacun de ces rôles, il associe la partie de taxonomie de connaissances du domaine qui peut servir de source pour la saisie. Ainsi, les éléments de spécialisation qui peuvent être choisis sont issus des connaissances du domaine d'étude pour les rôles de la partie Besoins de la requête et des connaissances expérimentales pour la partie Contraintes Expérimentales. Un fichier de modèle de requête, décrivant les éléments de formulaire nécessaires pour une étude de type «Comparaison», est présenté en Annexe C.

Une fois cette interface construite, l'utilisateur réalise des saisies, qui constituent, d'un point de vue purement logiciel, la formulation en tant que telle. Ce processus, présenté plus en détails en Annexe D, implique le parcours et le remplissage d'un

ensemble de formulaires, permettant le choix du type de requête et du domaine applicatif, puis la description de Généralités, des Contraintes Expérimentales et des Besoins et la validation des saisies réalisées.

Le framework construit ensuite un fichier de requête à partir des informations saisies, et le stocke pour traitement.

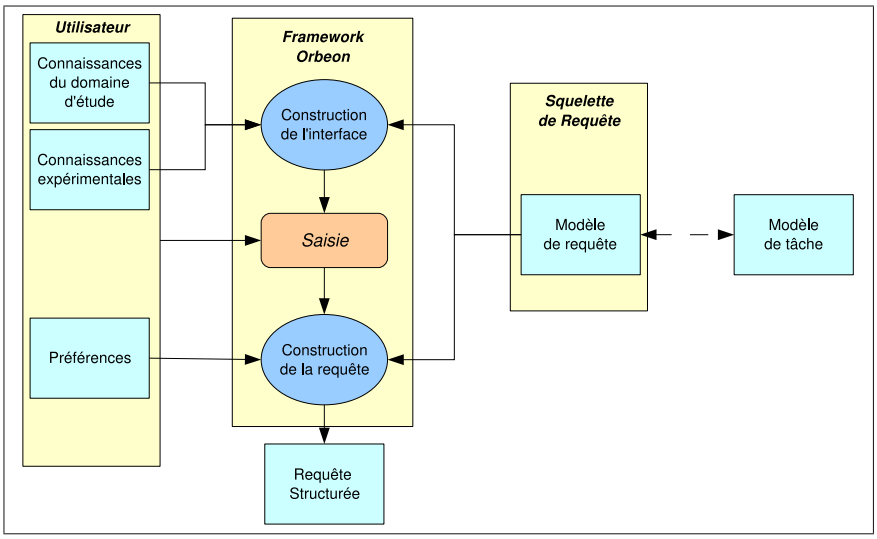


Fig. 4.6: Saisie de requête - Un modèle de requête, spécifique de la tâche choisie par l'utilisateur, est utilisé au sein du framework Orbeon pour construire un formulaire de saisie de requête que l'utilisateur complète avec des informations issues de son archétype. Les données saisies et des informations utilisateur sont alors enregistrées dans un fichier de requête.

Ce processus de saisie de requête pose principalement un problème cognitif de formulation, analysé plus en détails dans le prochain paragraphe.

4.5.3 Problème de formulation

Ainsi que présenté Paragraphe 3.4.3.1, la requête structurée est décomposée en trois parties (Généralités, Besoins, Contraintes Expérimentales), qui chacune regroupe un ensemble de rôles spécifiques du modèle de tâche considéré. La formulation de la requête implique tout d'abord une association informelle entre d'une part des éléments exprimés dans une description de l'étude à réaliser en langue naturelle et d'autre part les rôles du modèle de requête.

Centrale au processus de formulation, cette décomposition du problème biologique en rôles vis à vis d'une tâche prototypique de synthèse est un processus cognitif complexe. C'est en effet à ce niveau qu'apparaît le fossé sémantique et cognitif entre une représentation intellectuelle du besoin et son expression au sein d'une représentation formelle.

Ainsi, pour l'exemple de requête considéré, c'est-à-dire la «comparaison du pourcentage de cellules marquées entre les différentes molécules étudiées, entre tissus tumoraux et adjacents à la tumeur, en fonction de la localisation intracellulaire du marquage, chez les patients atteints d'un cancer du côlon», cette formulation informelle pourrait être représentée par le Tab. 4.2.

Tab. 4.2: Formulation informelle de requête - Le Tab. 3.1 est ici spécialisé de manière informelle pour la requête d'exemple : «comparaison du pourcentage de cellules marquées entre les différentes molécules étudiées, entre tissus tumoraux et adjacents à la tumeur, en fonction de la localisation intracellulaire du marquage, chez les patients atteints d'un cancer du côlon».

Élément du modèle	*Description*
Généralités :	
- Tâche	Comparaison
- Titre	Exemple d'illustration d'une tâche de type comparaison
- Description	Comparaison du pourcentage de cellules marquées entre les différentes molécules étudiées, entre tissus tumoraux et adjacents à la tumeur, en fonction de la localisation intracellulaire du marquage, chez les patients atteints d'un cancer du côlon
- Domaine	TMA
Besoins :	
- But	Pourcentage de cellules marquées
- Critères d'inclusion	Patients souffrant d'un cancer du côlon
- Critères de groupement	molécules étudiées - localisation du tissu par rapport à la tumeur - localisation intracellulaire du marquage
- Critères de tri	Année de naissance
Contraintes expérimentales :	
- Langue	Français
- Couleurs	Noir et Blanc
- Géométrie	-
- ...	
- Application des critères de sélection	Stricte
- Gestion des données manquantes	Exclusion
- Représentativité des groupes	Maximiser la variabilité
- ...	

Ainsi, la description de l'étude en langue naturelle est majoritairement représentée par la partie Besoins.

L'objectif de l'étude, représenté par le rôle «But», est ainsi le pourcentage de cellules marquées.

La population sur laquelle porte l'étude est définie par des «Critères d'inclusion», ici les patients atteints d'un cancer du côlon.

Les «Critères de groupement» permettent de définir selon quels axes seront constitués les groupes à comparer. Ici, il s'agit tout d'abord des molécules étudiées,

conduisant à la construction de quatre groupes, un par molécule. Ensuite, le second critère est la localisation du tissu par rapport à la tumeur. Au sein de chaque groupe de molécule, sera constitué un groupe pour le tissu tumoral et un groupe pour le tissu adjacent à la tumeur. Le dernier critère est la localisation intracellulaire du marquage, ce qui induit la constitution d'un groupe pour chaque compartiment cellulaire (membrane, cytoplasme, noyau) au sein de chacun des groupes qui ont été préalablement construits.

Les «Critères de tri» guident l'ordonnancement des individus au sein des groupes de niveau le plus fin, qui sera réalisé ici en fonction de l'année de naissance.

En ce qui concerne les contraintes expérimentales, les deux types de contraintes apparaissent. Par exemple, les couleurs à employer relèvent des connaissances de type «matériel». Par contre, les trois dernières qui sont présentées indiquent des contraintes de type «méthodes». Ainsi, la première induit une sélection des individus en fonction de critères de sélection par correspondance exacte, et non approximative. La seconde requiert l'exclusion des individus présentant des données manquantes et non l'inférence des valeurs manquantes. La troisième impose de maximiser la variabilité des individus au sein d'un groupe, plutôt que de conserver un groupe homogène, au cas où des individus doivent être exclus par manque de place au sein de la grille documentaire.

Cette formulation n'a pas d'existence réelle au sein du système, mais constitue un préalable intellectuel à la formulation en tant que telle, réalisée par le biais des interfaces. L'objectif poursuivi par le processus de saisie de requête est la construction d'un fichier de requête structuré et formalisé, interprétable par un système informatique, au format XML. Ainsi, la formulation informelle du Tab. 4.2 doit mener à un document tel que celui présenté en Annexe E.

4.6 Instanciation de la tâche

4.6.1 Introduction

Une fois la requête structurée saisie par l'utilisateur, elle est stockée et ajoutée à la liste des requêtes à traiter. Elle est alors repérée par un processus de contrôle du traitement de synthèse qui tourne en permanence. Celui-ci lance le traitement de la requête et change son statut à «En cours d'exécution». Il contrôle aussi les accès concurrents aux ressources par le système de traitement, par exemple pour s'assurer que les identifiants de requête sont bien uniques quand plusieurs sont exécutées en même temps.

La première phase du processus est l'instanciation de la tâche. D'après l'archi-

tecture logique du prototype, cette instanciation implique tout à la fois une opérationnalisation du modèle de tâche, par association de chaque sous-tâche élémentaire avec un composant particulier de la librairie de composants, et sa spécialisation avec des informations issues de la requête, des préférences de l'utilisateur et des connaissances du domaine.

L'architecture logicielle sous-jacente et les éléments qu'elle implique sont présentés dans la suite de cette section.

4.6.2 Architecture logicielle

L'instanciation de la tâche est réalisée par le gestionnaire de tâche, comme présenté Fig. 4.7. Celui-ci parcourt l'ensemble du fichier de modèle de tâche.

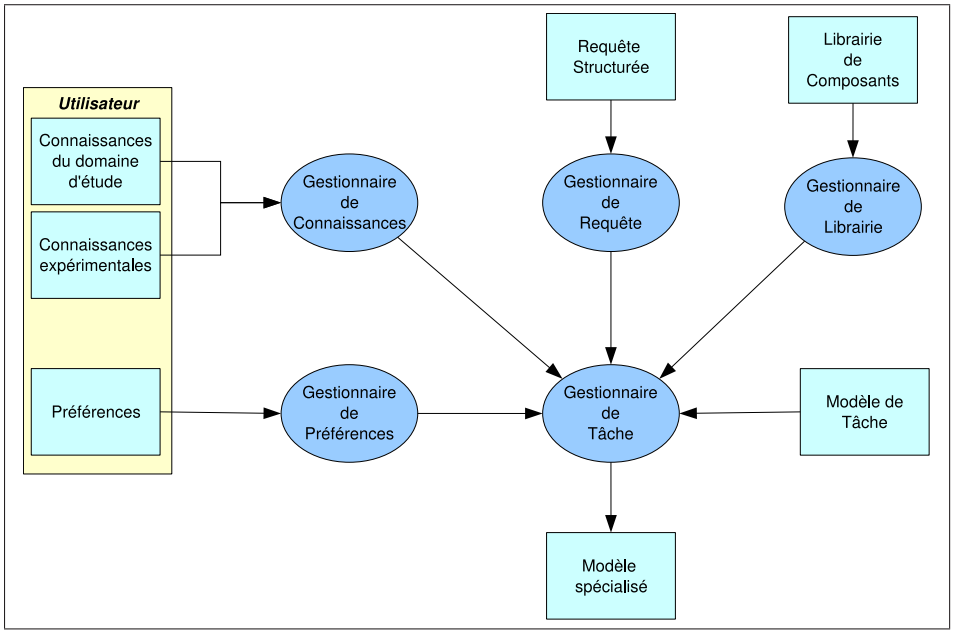


Fig. 4.7: Opérationnalisation du modèle de tâche - Ce processus est réalisé par le gestionnaire de modèle qui parcourt le modèle de tâche et le spécialise selon les indications de sources pour la spécialisation que le modèle contient. Selon les instructions du modèle, le gestionnaire accède à la requête, aux préférences utilisateur et aux connaissances du domaine par l'intermédiaire de gestionnaires spécifiques, récupère les informations de spécialisation et complète les paramètres de chacune des sous-tâches du modèle. Il a aussi recours à un gestionnaire de librairie pour accéder à la librairie de composants et déterminer quel composant utiliser pour chaque sous-tâche élémentaire.

Pour chaque sous-tâche élémentaire du modèle, le gestionnaire de tâche commence par définir quel composant utiliser, en se basant sur des connaissances expérimentales de type «méthodes». Dans le contexte du prototype, comme exposé Paragraphe 4.3.2.3, ces connaissances définissent un composant par défaut pour chaque sous-tâche élémentaire. Ce défaut peut être écrasé par des instructions spécifiques de la requête ou des préférences de l'utilisateur.

En pratique, le gestionnaire de tâche use d'un arbre de décision très simple pour choisir le composant à utiliser. Il commence par vérifier si un composant spécifique est défini dans la requête et si c'est le cas, prévoit d'utiliser celui-là. Si aucune indication n'est donnée dans la requête, il vérifie de la même façon les préférences utilisateur, et si rien n'est indiqué, il recourt au composant par défaut. Ainsi qu'il a été présenté Paragraphe 4.3.2.3, des représentations de connaissances expérimentales plus complexe pourraient permettre à l'avenir une gestion plus fine du choix du composant à utiliser pour chaque sous-tâche élémentaire, rendant ainsi le processus d'opérationnalisation plus adapté aux besoins des usagers.

Le gestionnaire de tâche lit ensuite les instructions de spécialisation, qui définissent les paramètres du composant à utiliser. Chaque paramètre inclut l'indication d'une source de spécialisation primaire et des sources optionnelles. Ces sources peuvent être un élément de la requête, des préférences utilisateur, des connaissances du domaine d'étude ou expérimental, ou le résultat d'une autre sous-tâche, stockée dans une partie du tableau noir. Le gestionnaire teste les sources dans l'ordre une à une en accédant à la source par l'intermédiaire d'un gestionnaire spécifique. Dès qu'une source définie a été trouvée, il utilise ces informations pour spécialiser le paramètre.

Un exemple de spécialisation de paramètre est proposé Fig. 4.8. La sous-tâche qui consiste en le calcul de la taille de la grille du document de synthèse a trois paramètres dont la largeur de la grille. Cette largeur peut être définie dans la requête, comme contrainte expérimentale. Elle peut aussi faire partie des préférences de l'utilisateur. Enfin, une valeur par défaut est définie dans les connaissances du domaine.

Pour spécialiser ce paramètre, le gestionnaire de modèle interroge tout d'abord le gestionnaire de requête pour savoir si la taille de grille est spécifiée dans la requête. Si oui, il utilise cette valeur pour spécialiser la largeur de grille. Sinon, il fait appel au gestionnaire de préférences pour déterminer si l'utilisateur a cette valeur dans son fichier de préférences. Si oui, il utilise la valeur issue des préférences pour spécialiser le paramètre, sinon, il demande la valeur par défaut au gestionnaire de connaissances et l'utilise pour spécialiser la largeur de grille.

Cette valeur de spécialisation est stockée au sein du tableau noir à un emplacement particulier, et la référence de cet emplacement est indiquée au sein du fichier de modèle spécialisé.

Une fois le paramètre spécialisé, le gestionnaire de modèle passe au paramètre suivant. Il a recours à la librairie de composants par l'intermédiaire d'un gestionnaire de librairie pour vérifier que tous les paramètres décrits pour le composant correspondant à la sous-tâche élémentaire sont bien définis. Une fois toutes les sous-tâches spécialisées, le modèle spécialisé, correspondant à la tâche instanciée est stocké dans un fichier XML et il peut être exécuté.

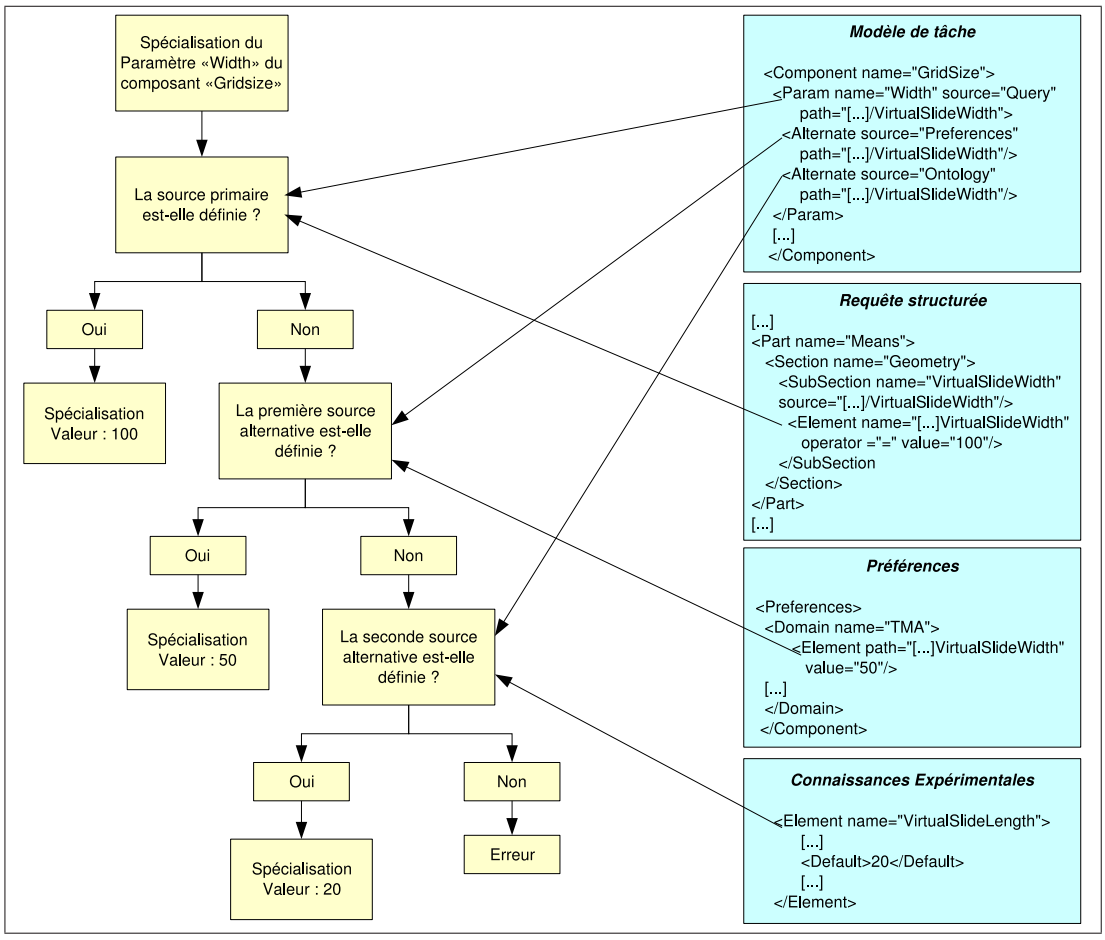


Fig. 4.8: Exemple de spécialisation de paramètre - Cette spécialisation est dirigée par le modèle de tâche selon un arbre de décision très simple, tel qu'illustré ici pour le paramètre «Width» du composant «GridSize» qui détermine la taille de la grille du document de synthèse.

4.6.3 Modèle spécialisé

L'instanciation du modèle de tâche qui vient d'être présentée, par le processus qui a été décrit en début de section, conduit à l'écriture d'un fichier XML de modèle spécialisé ou instance de tâche. Ce fichier reprend la même structure que le fichier de modèle de tâche et indique, au niveau sous-tâche élémentaire, le composant à utiliser, ainsi que les informations de spécialisation. Ces informations de spécialisation sont indiquées sous la forme de la référence à l'emplacement du tableau noir où l'information est localisée. Quelques exemples pour l'étude d'illustration considérée dans ce chapitre sont présentés Tab. 4.3.

Ainsi, par exemple, le composant «CriteriaApplication», qui permet la sélection stricte d'individus selon un jeu de critères reçoit comme paramètre la valeur de critère de sélection définie dans la requête, soit une localisation de la tumeur dans le côlon (TumorLocation = colon). Le composant «Language», lui, est spécialisé à

Tab. 4.3: Modèle de tâche spécialisé - Ce tableau répertorie quelques exemples de composants inclus dans le modèle de tâche et présente, pour l'exemple d'étude considéré, les informations de spécialisation correspondantes, telles que stockées dans le tableau noir.

Composant	*Description*	*Spécialisation*
Sélection :		
- CriteriaApplication	Application stricte d'un critère de sélection	TumorLocation = colon
- GroupingCriteria	Définition de la liste des groupes	Marker / LocationTowardsLesion / IntracellularLocation
- GroupsBuilding	Attribution des individus sélectionnés aux groupes	CleanListItems / GroupsList
- ...		
Organisation :		
- GridSize	Définition de la taille de la grille	LengthWidthRatio / GridLength / GridWidth / Count
- GroupsPositions	Attribution de groupes à des zones de la grille	GridWidth / GridLength / Groups
- ItemsPositions	Placement des individus au sein des zones de la grille	PercentMarkedCells / Items / Groups
- ...		
Présentation :		
- Language	Définition de la langue de l'interface	fr
- Colour	Définition du code couleur à utiliser dans la grille	White / Black
- ...		

partir de préférences utilisateur (fr pour le français).

Enfin, le composant «Gridsize», qui calcule la taille de la grille, a des sources variées pour sa spécialisation : la largeur de grille (GridWidth) est issue de la requête, le rapport entre largeur et longueur (LengthWidthRatio) est défini dans les préférences utilisateur, la longueur de grille (GridLength) est une valeur par défaut des connaissances expérimentales et le nombre total d'individus sélectionnés (count) est inscrit dans le tableau noir par un des composants de sélection.

Par contre, le composant «GroupsBuilding», qui attribue les éléments sélectionnés aux groupes, ne subit pas de spécialisation : ses paramètres sont toujours la liste des items sélectionnés après gestion des données manquantes (CleanListItems) et la liste des groupes (GroupsList) construite par le composant «GroupingCriteria».

Une vue partielle du fichier XML correspondant à ce modèle spécialisé est présentée et commentée en Annexe F.

4.7 Exécution de l'instance de tâche

4.7.1 Introduction

Le modèle de tâche ayant été spécialisé en fonction de la requête structurée, des préférences de l'utilisateur et de contraintes expérimentales, celui-ci sert de guide pour la construction du document de synthèse.

Cette construction est réalisée par le processus de synthèse en tant que tel, qui consiste en l'exécution dirigée d'un ensemble de composants décrits dans l'instance de tâche, selon l'architecture introduite dans le prochain paragraphe.

Ce processus requiert un accès tout à la fois aux diverses parties de l'instance de tâche et à la librairie de composants et implique l'utilisation d'un tableau noir. Ces diverses entités seront décrites par la suite.

4.7.2 Architecture logicielle

Le processus d'exécution d'une instance de tâche est contrôlé par un moteur d'exécution, comme présenté Fig. 4.9.

La première problématique à résoudre pour exécuter le modèle est de déterminer l'ordre d'exécution des composants, opération menée par l'ordonnanceur. L'ordonnanceur accède aux informations d'entrées et sorties (paramètres et cibles) des composants sous forme d'en-têtes de composants, qui sont extraits du modèle spécialisé par l'intermédiaire du gestionnaire d'instance de tâche. L'ordonnanceur construit alors un arbre de dépendances entre composants et, en fonction du contenu courant du tableau noir, indique quels composants peuvent être exécutés à cet instant au moteur d'exécution.

Le moteur d'exécution fait alors appel à l'initialiseur pour instancier et paramétrer les classes JAVA correspondant aux composants exécutables. Les informations concernant les classes à utiliser pour chaque composant sont inscrites dans la librairie de composants, que l'initialiseur interroge par l'intermédiaire d'un gestionnaire de librairie.

Le moteur d'exécution lance alors le traitement des composants initialisés. Celui-ci a lieu en parallèle. Chaque composant écrit les résultats de son traitement au sein du tableau noir, pour les rendre disponibles aux autres composants. Ce tableau noir peut être déchargé au sein d'un fichier XML spécifique, à des fins d'analyse du processus.

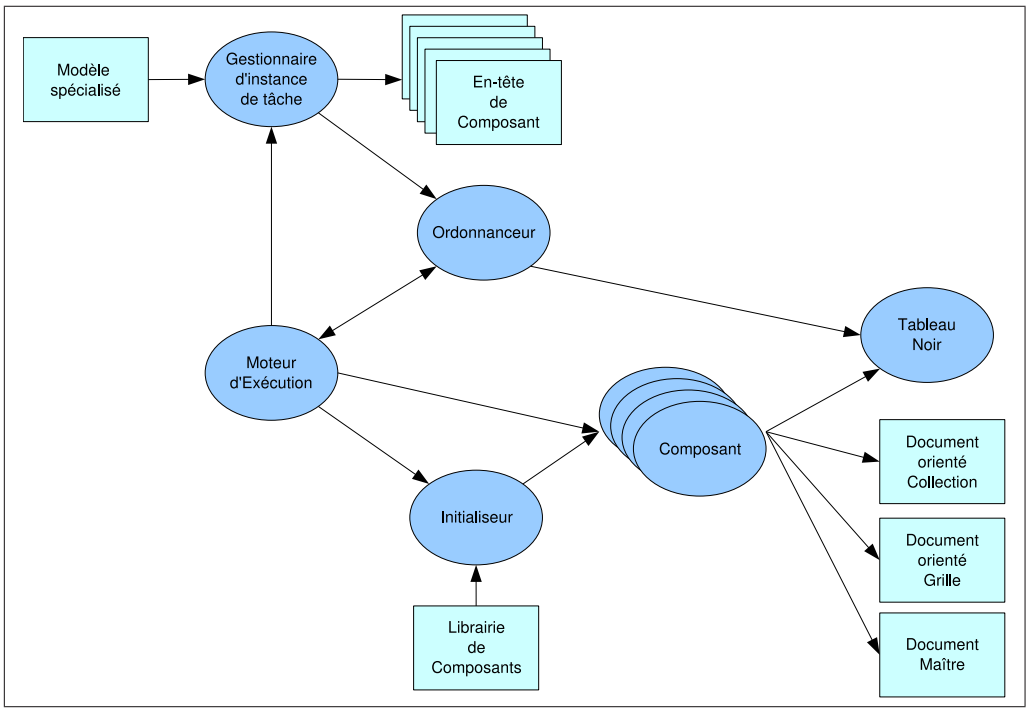


Fig. 4.9: Exécution du modèle opérationnalisé - Celle-ci est réalisée par un moteur d'exécution. Celui-ci fait appel à un ordonnanceur, qui détermine l'ordre d'exécution des composants, et un initialiseur qui prépare l'exécution des composants.

Au cours du processus, certains composants, dits composants de finalisation, écrivent des fichiers de résultat, comme un Document orienté Collection en fin de sélection. Au final, est réalisé un document XML appelé Document Maître qui inclut toutes les informations nécessaires pour présenter un document de synthèse à l'utilisateur.

4.7.3 Entités impliquées

4.7.3.1 Introduction

L'architecture décrite ci-dessus fait intervenir un certain nombre d'entités telles que le modèle de tâche opérationnalisé et sa décomposition en en-têtes de composants, qui guide le processus, la librairie de composants, qui indique les composants logiciels à utiliser pour chaque sous-tâche, le tableau noir, qui sert de support d'échange d'informations entre composants, les documents construits. De plus, un certain nombre de composants accèdent à un corpus documentaire pour réaliser leurs traitements.

Le modèle de tâche opérationnalisé, qui guide le processus, a été présenté en

détails dans la section précédente, et sa décomposition en en-têtes de composants est purement symbolique. Il ne sera donc pas détaillé plus avant. Le document maître, résultat de l'ensemble du traitement, sera quant à lui présenté plus en détails dans la prochaine section.

L'objectif poursuivi ici est donc de présenter la forme de la bibliothèque de composants, du tableau noir et du corpus documentaire.

4.7.3.2 Bibliothèque de composants

Tout comme les connaissances du domaine, la bibliothèque de composants consiste en une hiérarchie de répertoires dans le système de fichiers. La structure de cette hiérarchie recouvre l'organisation des divers modèles de tâche existants. Les feuilles de la hiérarchie sont représentées par des fichiers XML de description d'un composant. Ce fichier de description d'un composant est majoritairement défini à des fins documentaires, afin de spécifier son rôle et ses entrées et sorties. Au niveau fonctionnel, c'est surtout la définition de la classe JAVA qui est essentielle, puisqu'elle permet la construction d'un objet correspondant à ce composant et son exécution.

Le Tab. 4.4 liste les différentes parties d'un fichier de description d'un composant, dont un exemple commenté est proposé en Annexe G.

Tab. 4.4: Description d'un composant - Ce tableau répertorie les composantes principales de la description d'un composant : des généralités et en particulier la classe JAVA correspondante, des cibles et des paramètres, dont la nature éventuellement composée est décrite par un nombre d'occurrences minimum et maximum.

Élément	*Description*
Généralités :	
- Nom	Nom du composant tel qu'il apparaît dans l'instance de tâche
- Titre	Titre explicitant le nom du composant
- Description	Texte décrivant le fonctionnement du composant
- Classe JAVA	Nom de la classe à instancier et dont la méthode execute() doit être lancée pour exécuter le composant
Cibles (pour chacune) :	
- Nom	Nom de la cible (ou sortie) telle qu'elle est utilisée dans le tableau noir
- Titre	Explicitation du contenu de la cible
Paramètres (pour chacun) :	
- Nom	Nom du paramètre tel qu'il est défini dans le tableau noir
- Occurrence Minimum	Nombre minimum d'éléments composant le paramètre
- Occurrence Maximum	Nombre maximum d'éléments composant le paramètre

C'est donc ce fichier qui sert de base pour déterminer quelle classe instancier pour

exécuter un composant particulier. Cette exécution implique en général l'utilisation du tableau noir, présenté par la suite.

4.7.3.3 Tableau Noir

Afin d'éviter la mise en place d'un système de communication par messages entre composants, dont le contrôle peut devenir complexe, l'approche choisie dans le cadre du prototype pour permettre l'échange d'informations entre composants est une approche de type tableau noir. Le tableau noir est envisagé comme une structure d'échange particulière où les composants lisent et écrivent des données à des emplacements définis dans le modèle de tâche, comme présenté de manière schématique pour quelques composants au sein de la Fig 4.10.

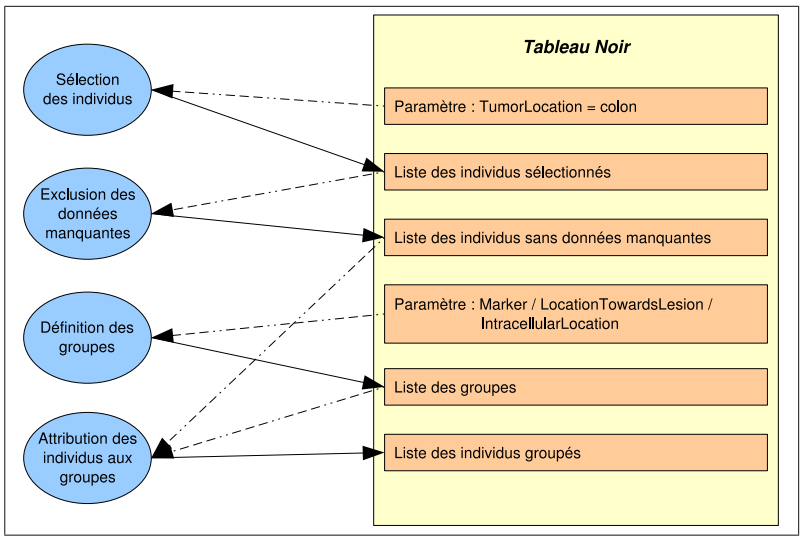


Fig. 4.10: Fonctionnement du tableau noir, vue partielle - Chaque composant lit (flèches en pointillés) et écrit (flèches pleines) des informations à des emplacements du tableau noir qui sont définis dans l'instance de tâche, afin d'échanger les résultats de leurs traitements.

Dans un contexte de développement d'un prototype, il peut s'avérer pertinent d'être en mesure d'accéder au contenu de ce tableau noir, à des fins de détection manuelle d'erreurs ou d'évaluation manuelle de la qualité des résultats produits par les divers composants. Un mécanisme de déchargement de l'ensemble du contenu du tableau noir, sous forme d'un fichier XML spécifique, a donc été mis en place. Ce fichier est présenté plus en détails en Annexe H.

4.7.3.4 Corpus documentaire

Dans le cadre de la plateforme TMA-Explorer, à laquelle s'intègre le système d'assistance à la synthèse, est disponible une base de données PostGreSQL contenant

les données correspondant aux dossiers cliniques des patients et des données histologiques. Dans le cadre du prototype, c'est cette base de données qui est utilisée comme corpus documentaire, mais on pourrait imaginer facilement le recours à une base XML type EXIST[8].

L'avantage de l'utilisation d'une base de données est liée au fait que la plupart des champs consistent en données numériques ou mots prédéfinis. Au niveau représentation de connaissances, ces champs correspondent donc à des concepts quantitatifs ou qualitatifs. Ceci facilite les traitement liés à la sélection des individus, puisqu'ils se limitent à une correspondance exacte entre requête et champs de la base. Ceci est particulièrement intéressant dans le cadre d'un prototype dont l'objectif est de tester les concepts d'un modèle, mais devra à l'avenir être étendu vers des champs textuels, faisant alors intervenir une correspondance plus similaire aux paradigmes de Recherche d'Information.

Par contre, le stockage des données sous forme d'une base de données induit l'éclatement des informations entre de nombreuses tables (une centaine pour le projet TMA-Explorer), ce qui n'est pas cohérent avec la notion de corpus documentaire sous forme de documents structurés. Afin de pallier ce problème, deux vues (au sens base de données de table dynamique virtuelle du terme), correspondant aux dossiers cliniques et aux données histologiques, ont été construites, fournissant de façon virtuelle deux types de documents structurés.

La vue patient donne accès aux informations du dossier clinique, telles l'état-civil, le diagnostic, la séquence thérapeutique subie, le suivi médical. La vue histologie contient des mesures de marquage pour divers marqueurs et tissus et des liens vers les images correspondantes. Chacune de ces vues peut être assimilée à une collection de documents structurés de la même manière, où une ligne dans la base de données correspond à un document particulier, ainsi que présenté Fig. 4.11.

4.8 Affichage du document de synthèse

4.8.1 Introduction

Le document maître construit par le processus de synthèse contient toutes les informations nécessaires pour à la fois proposer une visualisation des informations selon l'étude exprimée dans la requête et permettre une réutilisation des données dans un autre cadre. Il doit aussi donner lieu à des possibilités de reformulation de requête.

[8] http ://exist.sourceforge.net/

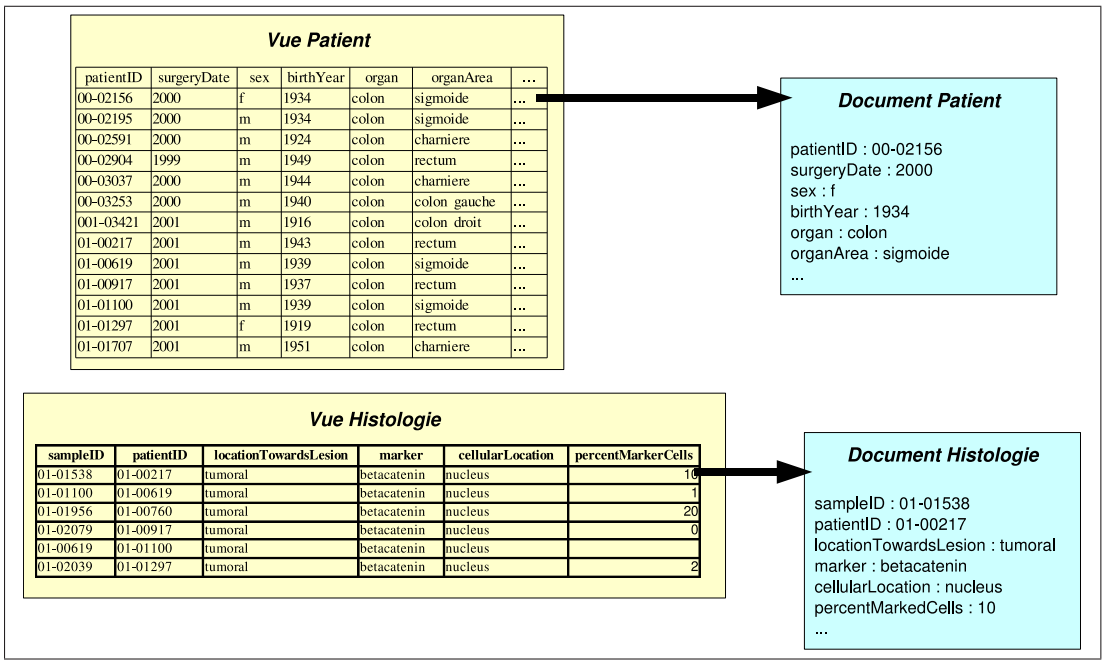


Fig. 4.11: Vues et documents - Au sein de chaque vue de la base de données, une ligne est considérée comme un document virtuel. Ainsi, un document patient correspond à une ligne de la vue patient, et un document histologie à une ligne de la vue histologie.

Mais, en tant que fichier XML, il est difficilement utilisable par l'usager, sans une transformation préalable. Cette section propose une exploration de ce processus d'affichage, à des fins de navigation au sein du document de synthèse et de reformulation. Ceci passe par une description de ce document maître et des méthodes utilisées pour l'afficher.

4.8.2 Architecture logicielle

Le résultat final du processus d'exécution d'une instance de tâche est un document maître au format XML. Deux problématiques associées interviennent alors pour son traitement : l'affichage et la navigation au sein de ce document, permettant de proposer à l'utilisateur un document de synthèse, et un support à la reformulation de requête, reformulation qui est en général motivée par l'observation du résultat du traitement.

En ce qui concerne la première problématique, il s'agit, à partir d'un fichier XML, de construire tout à la fois une vue sur la grille documentaire, et des vues sur les informations associées (vues patient et histologie) accessibles depuis la grille. Ce mécanisme est réalisé au sein du framework Orbeon Forms, par l'intermédiaire de règles XSLT, qui sont appliquées sur le document maître et permettent sa transformation en fichiers HTML. Spécifiques du domaine applicatif, ces règles constituent

un modèle de présentation.

La reformulation de requête repose sur des formulaires similaires à ceux proposés pour la saisie de requête. La différence est que les formulaires sont préremplis avec les informations saisies lors de la description initiale de la requête. Ce préremplissage est réalisé par extraction des informations depuis le fichier de requête et application là aussi de règles XSLT. Lancer l'exécution d'une requête reformulée conduit en pratique à l'écriture d'un nouveau fichier de requête, permettant ainsi de garder une trace du processus de reformulation.

Au final, ces deux opérations sont réalisées selon l'architecture présentée Fig. 4.12 et reposent principalement sur le document maître et des règles XSLT, qui vont être présentés plus en détails par la suite.

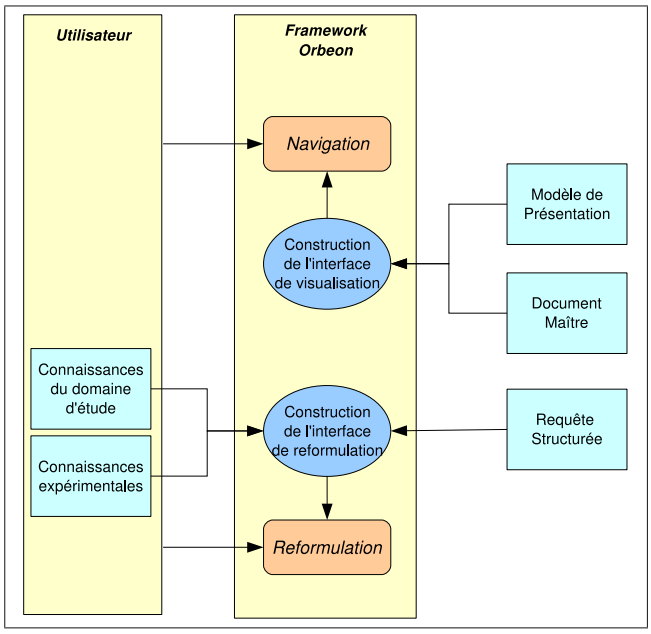


Fig. 4.12: Présentation du document de synthèse - Les deux opérations, permettant la navigation au sein d'un document de synthèse construit à partir d'un document maître et la reformulation de requête sont réalisées au sein du framework Orbeon Forms grâce à des règles XSLT.

4.8.3 Document maître

Le document maître, en tant que support à la construction d'un document de synthèse, doit permettre tout à la fois :

- ⋆ la description du contenu de la grille documentaire et des vues patient et histologie pour chaque item présent au sein du document de synthèse. Ceci correspond donc à des informations issues d'un processus de type sélection. Cette description contribue à constituer le fond du document,

- ⋆ la description de l'organisation des groupes éventuels et des individus au sein de la grille documentaire et au sein de chaque vue. Ceci correspond donc à des informations issues d'un processus de type organisation. Cette description contribue, au niveau conceptuel, au fond du document, mais, au niveau affichage, à sa forme,
- ⋆ la description de spécificités d'affichage, issues d'un processus de type présentation. Cette description contribue à constituer la forme du document,

Au niveau pratique, cette séparation entre fond et forme est peu adaptée à la génération d'un document de synthèse au format HTML, et le document maître est considéré comme constitué en deux sections différentes. La première inclut un en-tête rappelant la requête correspondante ainsi que toutes les informations nécessaires à la construction de la grille documentaire (fond et forme). La seconde décrit les données correspondant à chacune des vues. La structure d'un document maître peut alors être représentée schématiquement comme sur la Fig. 4.13.

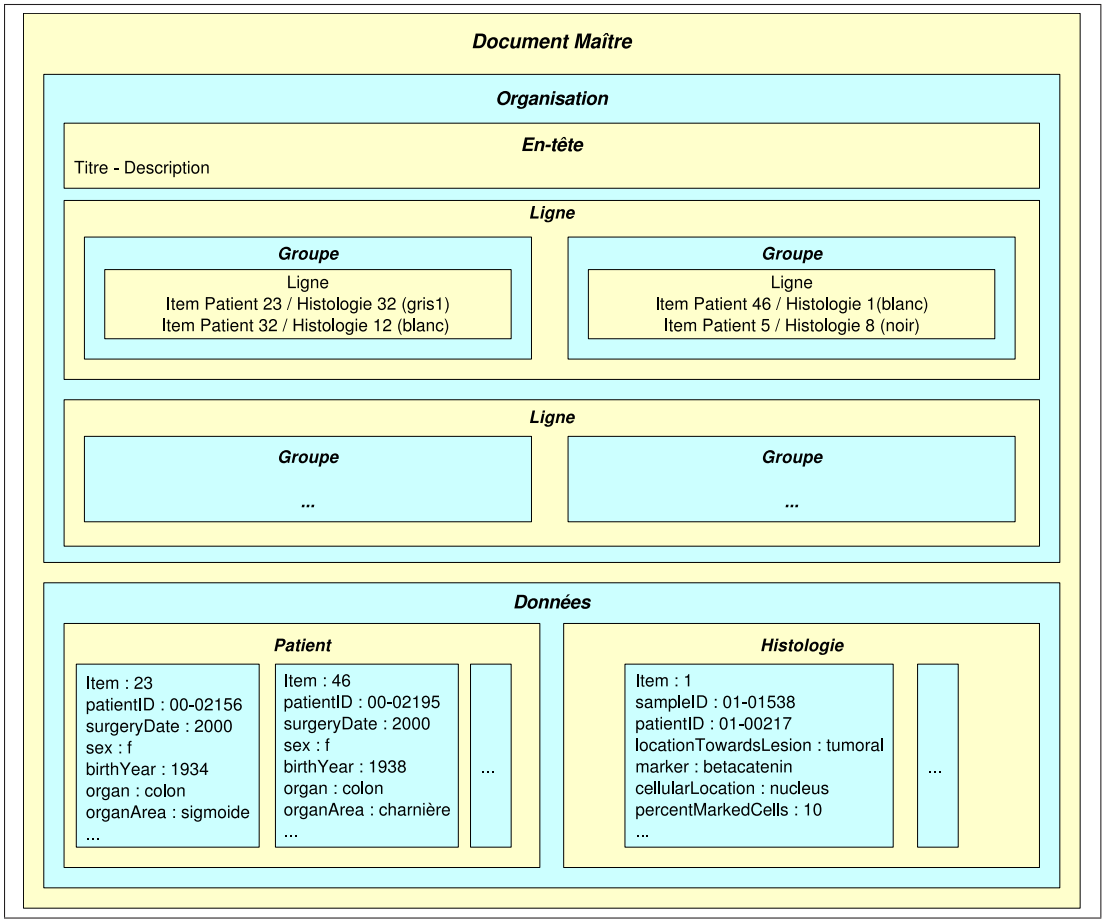


Fig. 4.13: Structure d'un document maître - Cette structure est décomposée en deux sections. La partie Organisation donne les informations nécessaires à la construction d'un en-tête et de la grille documentaire, aussi bien au niveau de l'organisation des groupes et items qu'au niveau couleur des cases. La partie Données indique pour chaque vue les données disponibles pour chaque item.

Au sein de ce document, la partie la plus complexe est la partie Organisation. Celle-ci, hors de l'en-tête, consiste en un imbriquement de Lignes, chaque Ligne correspondant à une zone rectangulaire horizontale. Les lignes peuvent contenir des groupes, découpages en rectangles verticaux des lignes, ou des items, correspondant à des cases de la grille. Au niveau item, en plus des identifiants correspondant à l'item pour chaque vue, une couleur pour le fond de la case, calculée en fonction de la valeur de la variable définie comme but de l'étude, est spécifiée.

Un exemple de fichier XML de document maître est présenté en Annexe I.

4.8.4 Document de synthèse

Après application de règles XSLT du modèle de présentation, le document maître est présenté à l'utilisateur sous forme d'un document de synthèse, dont un exemple est présenté Fig 4.14.

Au sein de cette figure, la grille documentaire, qui constitue le cœur du document, est présentée en vue partielle en haut. Cette grille représente chaque individu par une case dont le niveau de gris correspond au pourcentage de cellules marquées et dont la valeur correspond à un identifiant de l'individu au sein de la grille, ici le patient.

Ces cases sont groupées en zones dont les titres décrivent le contenu, c'est-à-dire le critère qui a permis la constitution du groupe, ici les divers marqueurs, la localisation du tissu par rapport à la lésion et les compartiments intracellulaires. Au sein de chaque zone, les cases représentant les individus sont ordonnées de gauche à droite et de haut en bas selon les valeurs de la variable définie comme critère de tri, c'est-à-dire l'année de naissance.

En cliquant sur le numéro de l'individu dans une case, on accède à la fiche patient correspondante, présentée en bas, à gauche. Cette fiche est décomposée en trois sections. Le haut rappelle l'identifiant du patient au sein de la grille et son numéro de dossier. La partie centrale liste les données histologiques associées à cet individu. Le bas de la fiche présente les données du dossier clinique, classées selon les quatre grandes catégories de la taxonomie du domaine d'étude correspondant au patient : état-civil, diagnostic, thérapeutique, données cliniques.

Chaque item de la liste des données histologiques associées au patient est un lien. Cliquer sur le lien permet d'accéder à une fiche histologique, en bas à droite, présentant les données histologiques en question plus en détails, dont des informations générales sur l'échantillon et le patient correspondant, puis les données en tant que telles, comme le marqueur, la localisation de la zone concernée et diverses mesures de marquage.

4.8.5 Du modèle de présentation au document de synthèse

Le modèle de présentation doit guider la transformation du document maître, au format XML, en un document de synthèse affichable dans un navigateur Web au sein duquel la navigation entre la grille documentaire et les différentes vues est possible, c'est-à-dire un document HTML conservant le fond mais à la forme différente, qu'il s'agisse de l'organisation des éléments ou de leur aspect visuel. Il a donc été conçu comme un fichier de règles XSLT.

En effet, XSLT est un langage standard du W3C[9] qui permet la transformation de documents XML en autres documents XML, et a été conçu comme faisant partie de XSL, le langage de feuilles de style pour XML. Or, un document HTML peut être considéré comme un document XML au jeu de balises spécifique. XSLT est donc un langage adapté pour réaliser la transformation du document maître en document de synthèse.

En pratique, le modèle de présentation consiste en deux jeux de règles distincts. Pour la construction de la grille documentaire, il spécifie pour chaque balise et chaque attribut du document maître un remplacement conditionnel par des balises HTML conduisant à la forme et au fond de la grille. Ainsi, par exemple, les Lignes deviennent des lignes d'un tableau, les Groupes des colonnes et les Items deviennent des cases de la grille. Pour la construction des vues, il définit pour chaque item de chaque vue comment celle-ci doit être présentée au sein d'une fiche spécifique.

La Fig. 4.15 propose une vision schématique de ce processus pour la construction de la grille documentaire.

4.9 Un prototype opérationnel à évaluer

Le modèle de synthèse proposé au Chapitre 3 a été abordé ici d'un point de vue opérationnel, permettant la mise en place d'un prototype. Ce prototype a été conçu selon une architecture logique reflétant les fonctionnalités nécessaires à la réalisation d'un document de synthèse telle qu'une gestion des utilisateurs, une saisie de requête structurée, une instanciation de tâche guidée par la requête, une exécution d'instance de tâche et une présentation du document final.

Ancrée dans un cadre de développement spécifique, cette architecture a induit des choix technologiques particuliers. Le contexte du projet TMA-Explorer a conduit au choix d'une architecture Web trois-tiers, d'un développement en JAVA et du stockage du corpus documentaire au sein d'une base de données. Les besoins

[9] http ://www.w3.org/TR/xslt

de réutilisation de données des usagers potentiels ont suggéré le recours massif au format XML pour représenter les entités impliquées dans le processus de synthèse. Un choix personnel de simplicité de conception et développement a requis la mise en place d'un framework composants spécifique, où les composants communiquent par l'intermédiaire d'un tableau noir, pour implémenter des méthodes de résolution pour chaque sous-tâche élémentaire d'un modèle de tâche.

Une fois mis en place ce cadre de développement, les entités principales impliquées dans le processus de synthèse, c'est-à-dire les notions de tâche et de connaissances applicatives, ont été présentées d'un point de vue implantation. Les diverses fonctionnalités du prototype ont aussi été présentées, conjointement aux choix de développement associés.

En pratique, un système d'assistance à la synthèse a donc été mis en place. Par rapport aux bases conceptuelles qui ont été décrites au Chapitre 3, le développement du prototype a en général conduit à une simplification des problématiques sous-jacentes identifiées au chapitre précédent et à une limitation du champ couvert par le prototype.

Cette simplification a tout d'abord concerné le niveau de raffinement pris en compte dans les diverses entités et processus. Par exemple, au niveau gestion des utilisateurs, la notion d'archétype a été laissée de côté. De plus, la prise en compte de connaissances expérimentales de type «méthodes», qui influencent la résolution d'une sous-tâche élémentaire, est restée naïve. De même, le choix d'une source de spécialisation pour une sous-tâche est réalisé selon des heuristiques très simples, par un arbre de décision figé a priori.

Cette simplification est aussi intervenue selon une dimension résolution de problème. Certaines sous-tâches élémentaires sont fonctionnellement très complexes, sans être facilement décomposables plus avant en problèmes unitaires de petite taille. La résolution de tels problèmes a en général été proposée par des méthodes parfois simplistes, conduisant à une résolution qui est loin d'être optimale.

Par exemple, au sein des tâches prototypiques de type comparaison, une sous-tâche élémentaire calcule l'attribution de zones de la grille à chaque groupe et sous-groupe définis précédemment. Il s'agit d'un problème très complexe, similaire au problème de «bin packing», classique en Recherche Opérationnelle, mais avec des formes de boîtes dynamiques, à surface plus ou moins constante. Il a été résolu de manière rudimentaire, conduisant dans certains cas à un usage non optimal de l'espace disponible.

De plus, le champ couvert par le prototype a aussi été limité. Cette limitation est intervenue selon deux axes : le corpus documentaire et les tâches prototypiques.

En ce qui concerne le corpus documentaire, seuls des éléments associés à des

concepts quantitatifs et qualitatifs des connaissances du domaine ont été pris en compte. Ceci a permis d'éviter de démultiplier les traitements en fonction de la nature des éléments manipulés. De plus, les problématiques complexes relevant de la Recherche d'Information en tant que telle ont pu être laissées de côté. La taille des documents a aussi été limitée : seuls les éléments jugés par les utilisateurs comme vraiment essentiels ont été pris en compte.

Au niveau des tâches prototypiques, seules deux tâches ont fait l'objet de la définition d'un modèle de tâche et du développement des composants correspondants, afin de réduire le nombre de composants à développer. Les tâches de comparaison sont représentées par la tâche prototypique de comparaison d'une variable entre plusieurs groupes, où les éléments de la comparaison sont définis par l'usager, qui a servi d'exemple dans ce chapitre et le précédent. Les tâches d'évolution sont représentées par une évolution monovariée.

Au final, malgré ces simplifications et limites, un prototype fonctionnel est disponible. Se pose alors la question de l'adéquation du logiciel développé au modèle de synthèse de données envisagé et surtout au besoin d'appréhension de données qui a été identifié. Ceci implique de s'intéresser à une validation expérimentale du prototype, objet du prochain chapitre.

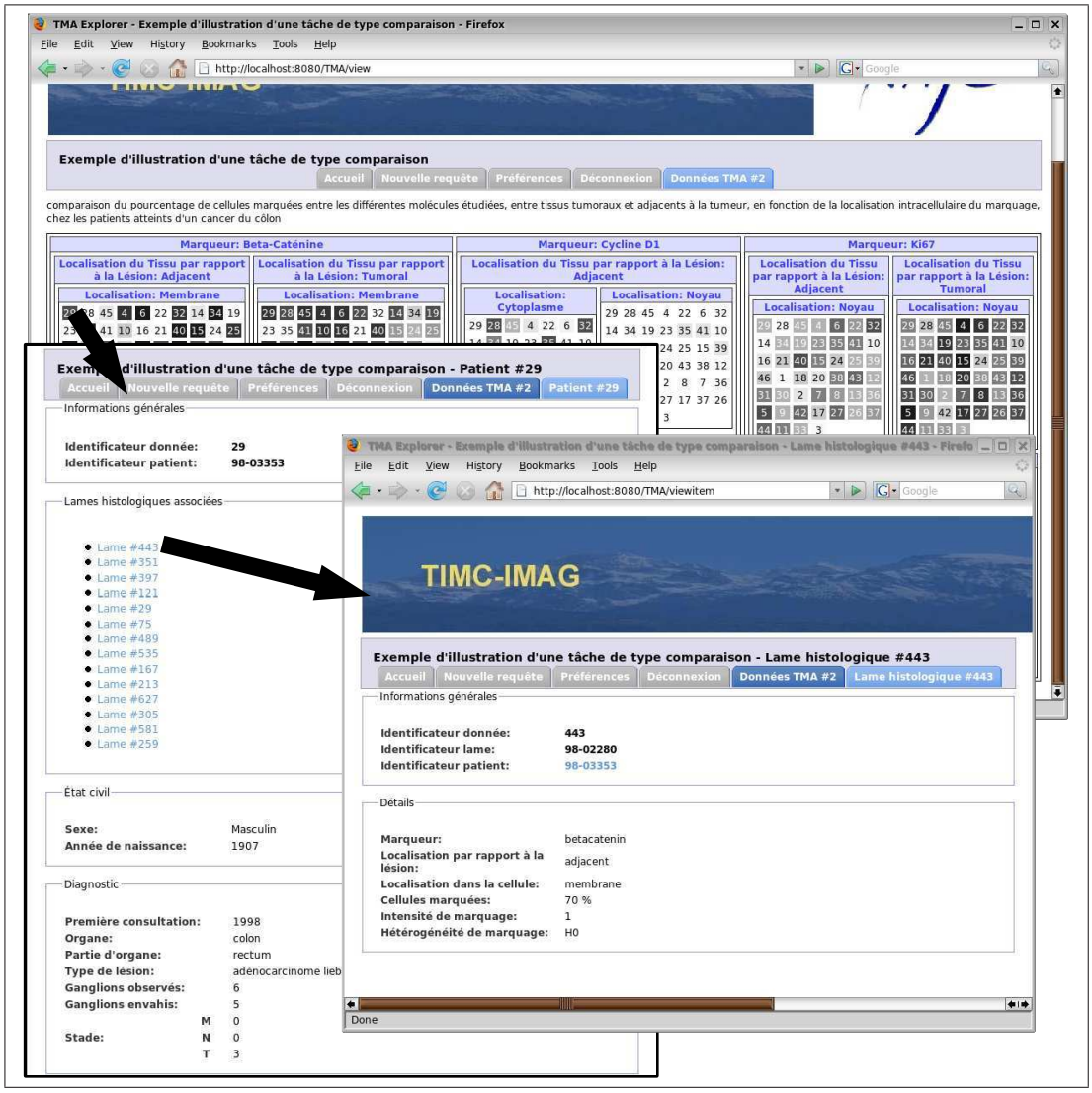


Fig. 4.14: Aperçu d'un document de synthèse - La grille documentaire donne accès à une fiche patient qui présente le dossier clinique et une liste de données histologiques associées. Chaque donnée histologique peut être détaillée sous forme d'une fiche particulière.

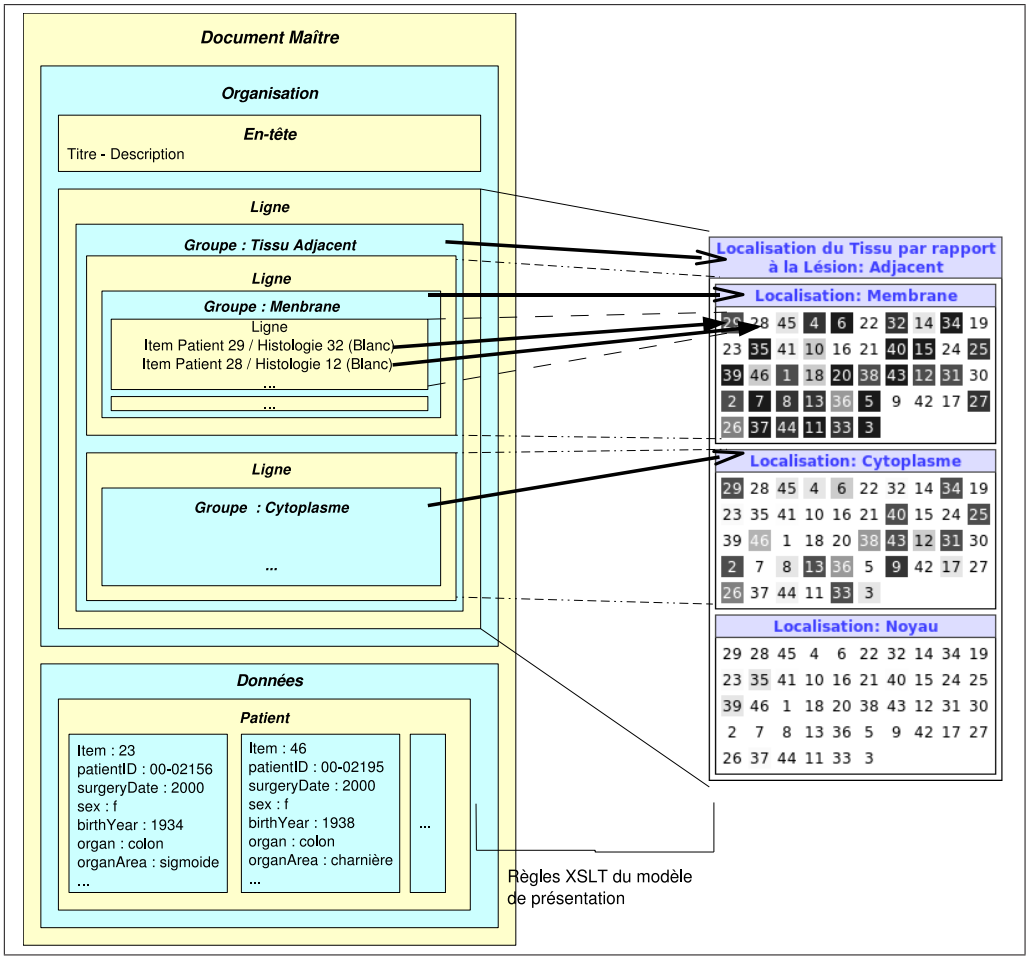


Fig. 4.15: Principe de la transformation du document maître - Ce schéma montre comment l'application des règles XSLT du modèle de présentation transforment le document maître en document de synthèse. Ainsi, les traits fins de divers types montrent comment chaque Ligne du document maître devient une ligne (ou zone rectangulaire horizontale) du document de synthèse. Les flèches à pointe fine correspondent à la transformation des groupes en zones rectangulaires verticales et les flèches à grosse pointe à la transformation des Items en cases.

CHAPITRE

5 Validation expérimentale du concept de synthèse d'information

Les chapitres précédents ont permis, dans le contexte applicatif des Tissue MicroArrays, la description d'un modèle de synthèse, basé sur un paradigme de Recherche d'Information orientée tâche, ainsi que la présentation de l'opérationnalisation de ce modèle au sein d'un prototype de système. Afin de déterminer l'intérêt des propositions qui ont été faites, aussi bien au niveau conceptuel que logiciel, le prototype réalisé doit être évalué au cours d'expérimentations. Cette validation expérimentale est basée sur une méthodologie inspirée tout à la fois de l'évaluation logicielle en général, et, en particulier, de l'évaluation des systèmes de Recherche d'Information ou des sites Web, auxquels le prototype peut se rattacher, du fait du modèle sous-jacent et de sa nature multimédia. L'évaluation passe par des études utilisateur et des études de cas, augmentées de comparaisons avec d'autre systèmes qui peuvent apporter des fonctionnalités similaires.

5.1 Introduction

L'exploration du domaine applicatif des Tissue MicroArrays a permis de dégager un problème d'appréhension des données générées par la technologie. Cette appréhension peut être envisagée en tant qu'étape préalable à une fouille de données ou afin de se replacer dans une démarche expérimentale classique et ne trouve pas de réponse dans les outils informatiques associés à la technique.

La solution envisagée dans ma thèse est la mise en place d'un système de synthèse, basé sur un paradigme de Recherche d'Information orientée tâche, dont les bases conceptuelles ont été précédemment introduites. Un prototype de système, opérationnalisant cette proposition, a été présenté au chapitre précédent. Il s'agit alors de mettre en œuvre une validation expérimentale de ce prototype, afin d'évaluer les réalisations actuelles.

Cette évaluation peut alors être envisagée dans une double perspective. Tout d'abord, le système proposé, en tant que système informatique, relève de l'évaluation logicielle en général. Ensuite, le modèle de synthèse décrit au Chapitre 3 inclut des problématiques qualité spécifiques, inspirées des préoccupations des chercheurs en Recherche d'Information, dont l'évaluation est elle aussi pertinente. Ces deux axes d'évaluation doivent donc être rapidement explorés, afin de définir un cadre expérimental pour la validation du prototype.

En premier lieu, l'évaluation est en effet un élément au rôle important dans le développement logiciel, puisqu'elle permet d'estimer la qualité du produit par rapport aux besoins explicites ou implicites de ses utilisateurs potentiels. Cette évaluation pose le problème de la notion de qualité logicielle. Bien que celle-ci existe depuis des dizaines d'années, il n'y a pas de consensus sur une ou des bonnes mesures à cause de la multitude d'interprétations de la notion de qualité, de la multitude de sens des termes servant à décrire ses aspects. De plus, plusieurs courants de pensée ont conduit à des métriques diverses : Interfaces Hommes/Machine, paradigme de satisfaction du consommateur et autres. Ainsi, les diverses dimensions de cette qualité logicielle incluent des notions telles que l'efficacité, la performance, la qualité de l'information, l'utilisabilité, l'acceptation de la technologie ou la satisfaction de l'utilisateur.

En second lieu, le Chapitre 3 a permis la définition d'un ensemble de problématiques qualité spécifiques au problème de synthèse de données TMA : adéquation à la tâche, dédiée à la correspondance entre la perception de son problème par l'usager et sa représentation au sein d'une requête structurée, pertinence situationnelle, consistant en un jugement système de la correspondance entre la requête et le résultat produit et pertinence interprétationnelle, permettant une évaluation qualité du document de synthèse par l'utilisateur. Ces diverses notions doivent être prises en compte dans le cadre d'une validation expérimentale du système.

Se pose alors la question du choix des métriques à utiliser dans le cadre d'une validation du prototype développé. Dans la suite du chapitre, va tout d'abord être présentée la méthodologie sous-jacente à la définition et à l'évaluation de ces métriques, basée sur un état de l'art du domaine, puis l'évaluation en tant que telle, par une étude utilisateur et des études de cas, avant de présenter un bilan de cette validation.

5.2 Méthodologie

5.2.1 Introduction

Les évaluations qualité spécifiques à la Recherche d'Information orientée tâche ont été présentées au sein du Chapitre 3. Il s'agit alors de s'intéresser à la qualité logicielle en général, et d'estimer en quelle mesure les notions d'adéquation à la tâche, pertinence situationnelle et pertinence interprétationnelle s'intègrent au sein des problématiques de qualité logicielle qui peuvent s'avérer intéressantes dans le cadre du prototype développé.

Ce problème de qualité logicielle n'est pas récent et a été rapidement identifié comme un des facteurs principaux du succès ou de l'échec d'un produit. L'évaluation de cette qualité est donc une thématique qui fait l'objet de nombreux travaux en ingénierie logicielle. Ainsi, [Tian, 2004] propose une revue des divers catégories de modèles et mesures et indique des pistes pour choisir une approche d'évaluation.

De manière schématique, l'évaluation est variable selon la phase du développement et poursuit en général deux buts distincts : évaluation formative en cours de développement pour améliorer le système, évaluation summative à la fin du développement pour voir si le système correspond à ce qui est attendu. Dans le contexte présent, c'est-à-dire celui d'un prototype, il est possible de se placer dans ces deux cadres d'évaluation, de manière conjointe. En effet, la notion de prototype sous-tend un logiciel partiel, incomplet, que l'on veut pouvoir améliorer et auquel on veut ajouter des fonctionnalités, ce qui suggère une évaluation formative. D'un autre côté, le prototype a été construit dans un objectif particulier de réalisation et il est intéressant d'estimer si cet objectif a été atteint, par une évaluation summative.

De plus, on peut distinguer d'un part des évaluations quantitatives, basées sur un jeu de mesures numériques objectives (taux d'erreurs, temps de traitement, taux d'acceptation de la technologie etc.) ou subjectives (valeurs attribuées par un panel d'utilisateurs à des mesures de type utilité ou satisfaction) et d'autre part des mesures qualitatives (en général subjectives comme des commentaires d'usagers recueillis lors d'interviews).

Dans ce contexte, afin d'estimer quelles évaluations et quelles métriques seraient pertinentes dans le cadre du prototype, une revue rapide de l'état de l'art du domaine va être menée. Cette revue va être réalisée en ce qui concerne l'évaluation logicielle en général, et l'évaluation des sites Web en particulier, ces deux types d'entités étant pertinents pour une application Web. Ensuite seront présentées succinctement les évaluations proposées.

5.2.2 État de l'art

5.2.2.1 Introduction

Afin de choisir une méthode et des mesures qualité à appliquer au prototype qui a été développé, une revue rapide de quelques pratiques courantes va être menée ici.

L'objectif n'est bien sûr pas de mener une analyse exhaustive des notions de qualité et d'évaluation dans un contexte d'ingénierie logicielle, qui se trouve bien en dehors du champ couvert par ma thèse. Il s'agit plutôt de fournir quelques pointeurs et exemples qui permettent de guider le choix d'une méthodologie et la définition de mesures qui seront appliquées par la suite pour valider le modèle de synthèse par l'intermédiaire d'une évaluation du prototype.

Le système ayant été développé comme une application Web, l'évaluation logicielle en général, et l'évaluation des sites Web en particulier, sont l'une et l'autre d'intérêt, et vont être rapidement évoquées.

5.2.2.2 Évaluation logicielle

En ce qui concerne l'évaluation logicielle, l'ancienneté et la criticité de cette problématique ont conduit à la définition d'un ensemble de normes ISO[1]. On peut noter par exemple la norme ISO/IEC 14598 qui est consacrée à la qualité logicielle, mais plutôt d'un point de vue processus de développement, qui n'est pas l'objet d'intérêt ici. Par contre, la norme ISO/IEC 9126 se focalise sur la qualité du logiciel en tant que produit et propose un modèle dans sa première partie, et des métriques dans les parties 2, 3 et 4. La partie 4 est d'un intérêt particulier dans l'évaluation du système d'assistance à la synthèse, étant donné qu'elle se focalise sur la qualité en utilisation. Enfin, la norme ISO 9241, consacrée à l'ergonomie logicielle est elle aussi pertinente, étant donné que le système proposé est interactif.

Ces normes proposent en général des modèles qualité et éventuellement définissent

[1] http ://www.iso.org

des métriques mais ne proposent pas de guide pratique pour leur application à des projets concrets. Par contre, elles servent de base à un certain nombre de modèles utilisés en milieu industriel ou académique.

Ainsi, [Lee and Lee, 2005] présentent un modèle d'évaluation qualité basé sur les ISO et une méthode de développement pour les systèmes à composants du ministère de la défense sud-coréen pour définir un ensemble de métriques pratiques classées en catégories de type fonctionnalité, fiabilité, utilisabilité, efficacité, maintenabilité et portabilité. Dans le même esprit, [Gediga et al., 1999] proposent un questionnaire basé sur la norme ISO 9241-10 pour l'évaluation utilisateur de logiciels, nommé IsoMetrics, qui conduit à une évaluation en termes d'adéquation à la tâche, intuitivité, contrôlabilité, conformité aux attentes des utilisateurs, tolérance à l'erreur, adéquation à l'individualisation et adéquation à l'apprentissage. On peut aussi citer [Côté et al., 2005], qui introduisent une application réelle de la norme ISO 9126 et d'un modèle qualité développé par un industriel, MITRE Corporation. En particulier ils s'intéressent à la problématique de correspondance entre composantes des deux modèles, en présentant la traduction de l'un à l'autre.

Certains modèles ont en effet été définis indépendemment des normes ISO, sans pour autant s'en éloigner beaucoup. Ainsi, [Wong, 2003] se base sur une étude quantitative de ce qui est jugé important en terme de qualité logicielle pour des décideurs afin de mettre en place des mesures au sein d'un framework d'évaluation de logiciels. Il obtient un jeu de mesures qu'il classe selon des catégories différentes (économique, fonctionnel, institutionnel, opérationnel, technique, usabilité) mais qui sur le fond sont similaires à celles définies dans les normes ISO.

5.2.2.3 Évaluation de sites Web

Dans le contexte du Web, les métriques proposées pour les logiciels en général ne sont pas forcément adaptées, et d'autres points de vue s'avèrent d'intérêt. Ainsi, [Reix, 2003], après avoir replacé l'évaluation des sites entre problématiques qualité d'interface Homme/Machine et théorie de la satisfaction du consommateur propose d'aller vers une perspective interactionniste, sans forcément proposer de métriques.

Dans une autre direction, [Cohen and Casanova, 2001] introduisent une grille d'analyse des sites Web basée sur des considérations cognitives, par interprétation des phénomènes visuels, conduisant à une catégorisation entre indicateurs perceptifs, indicateurs graphiques et indicateurs d'orientation, très proches de la forme des pages. Ces contingences basées sur la forme sont aussi utilisées par [Ivory et al., 2001], qui montrent que des mesures page par page basées sur des éléments de forme permettent de prévoir une évaluation qualité par un expert pour un site Web, permettant ainsi la définition d'un profil pour de bonnes pages.

Mais l'évaluation des sites Web dépasse en général les simples éléments de forme et se base aussi sur des questionnaires incluant d'autres considérations. Ainsi, [Olsina et al., 2001] introduisent la méthodologie QEM, qui catégorise les problématiques en utilisabilité, fonctionnalité, fiabilité du site, efficacité. [Mich et al., 2003] proposent le modèle 2QCV3Q, qui permet une évaluation de site Web du point de vue auteur et utilisateur selon un ensemble de directions : Quis (identification et caractérisation), Quid (couverture et exactitude), Cur (fonctionnalités, contrôle), Ubi (accessibilité, interactivité), Quando (actualité, maintenance), Quomodo (accessibilité, navigabilité, compréhensibilité), Quibus Auxiliis (ressources, technologies de l'information et de la communication).

Enfin [Tullis and Steton, 2004] présentent une comparaison de questionnaires classiques pour des études utilisateur sur des sites Web, fournissant ainsi un panel d'exemples qui peuvent s'avérer d'intérêt.

5.2.2.4 Des tendances générales pour l'évaluation

Les quelques exemples de méthodes et métriques d'évaluation qualité présentés ici permettent de dégager quelques tendances.

En ce qui concerne les logiciels en général, on peut noter deux grandes catégories de métriques : celles liées à l'artefact logiciel en tant que tel, proposant une évaluation de son développement, de sa maintenabilité et de son évolutivité, et celles liées à l'usage qui en est fait, apportant un point de vue utilisateur.

Au niveau des sites Web, ces métriques se déclinent entre fond (qualité et actualité des informations présentées), forme (rendu des pages) et usage (navigation, recherche, utilisabilité, etc.).

L'étape suivant est donc de déterminer quelles catégories de métriques sont pertinentes pour évaluer le prototype développé au cours de ma thèse.

5.2.3 Méthodes choisies

Dans le contexte d'un prototype de système d'assistance à la synthèse dédié au domaine applicatif des Tissue MicroArrays, il s'agit alors d'estimer quelles métriques courantes en évaluation logicielle seraient pertinentes, tout en prenant en compte les contingences qualité qui ont été évoquées au sein du modèle de synthèse. Cette notion de pertinence des métriques est liée à l'objectif de l'évaluation. Il s'agit donc de garder à l'esprit que le logiciel testé est un prototype qui se veut une opérationnalisation d'un modèle, où c'est le modèle qu'il faut valider. Ce point de vue permet

d'écarter certains types des métriques ou d'en restreindre d'autres, pour proposer un jeu d'évaluations pertinent dans l'objectif recherché.

En ce qui concerne l'évaluation logicielle en général, la majorité des métriques liées à l'objet logiciel sont peu pertinentes. En effet, la qualité du processus de développement, les besoins en maintenance et évolutivité, ne sont pas des éléments critiques dans le cadre d'un prototype visant à valider des propositions conceptuelles. Par contre, des mesures de performance peuvent s'avérer intéressantes : en effet, s'ils sont confrontés à un système à la réactivité faible et produisant beaucoup d'erreurs, les utilisateurs potentiels peuvent juger qu'il est inutile d'aller plus loin dans le développement.

D'un point de vue évaluation de sites Web, la qualité du fond, souvent assimilée à la qualité de l'information, n'est pas vraiment pertinente, car les documents de synthèse ne constituent pas de «vraies» pages et la qualité des résultats dépend plutôt de la qualité du corpus documentaire, problème situé en amont du système de synthèse. La forme, quand elle concerne des aspects de couleurs ou fontes, n'est pas non plus d'un intérêt majeur.

Enfin, en ce qui concerne les mesures évoquées dans le cadre du modèle de synthèse, la pertinence situationnelle, en tant que mesure objective, est difficile à définir et ne sera pas abordée plus avant que les pistes de définition évoquées Paragraphe 3.4.4.2.

Par contre, les points de vue usage de l'évaluation logicielle et de l'évaluation de sites Web semblent tout à fait en adéquation avec la problématique de validation du modèle. De plus, les dimensions adéquation, complétude, expressivité, extensibilité et navigabilité de l'adéquation à la tâche du modèle de synthèse, et les dimensions intuitivité, informativité, utilité, suggestivité et navigabilité de la pertinence interprétationnelle peuvent chacune être considérée comme une dimension des axes d'évaluation courants en ce qui concerne l'évaluation de l'usage.

Cette évaluation de l'usage peut être appréhendée selon plusieurs axes.

Dans un premier temps, l'évaluation de l'usage implique des considérations concernant les fonctionnalités proposées, d'un point de vue système. Il s'agit alors de considérations de type diagnostic : le système propose-t-il des fonctionnalités aux résultats correspondant à ce qui était prévu ? Une telle évaluation peut être réalisée par l'intermédiaire d'études de cas, qui consistent, dans le contexte du prototype de système d'assistance à la synthèse, en une analyse détaillée de quelques exemples d'études menées avec le système.

Dans un second temps, l'évaluation de l'usage induit de s'intéresser au point de vue des usagers potentiels du système. Une telle problématique suggère de mener une étude utilisateurs, basée sur un questionnaire et des interviews réalisés auprès d'un

panel représentatif d'utilisateurs après une session avec le prototype. Le questionnaire, organisant des questions selon les divers axes d'évaluation identifiés précédemment, permet de collecter des données quantitatives, tandis que les interviews fournissent des informations qualitatives. L'ensemble permet d'avoir un point de vue subjectif sur le système.

Au cours des tests utilisateur, des mesures de performances peuvent être réalisées, qui donnent un aperçu tout à la fois sur les éventuels problèmes du système et les difficultés rencontrées par les usagers.

La suite de ce chapitre va présenter ces études de cas et étude utilisateurs.

5.3 Études de cas

5.3.1 Objectifs

L'objectif des études de cas est de permettre d'estimer si le prototype proposé permet d'atteindre des résultats correspondant à ce qui était prévu. Dans le cadre du problème de synthèse, il s'agit donc de déterminer si le système permet d'appréhender un gros volume de données afin de le replacer dans une démarche expérimentale classique ou de préparer une fouille de données.

Dans l'objectif d'évaluer l'appréhension des données, plusieurs axes d'analyse doivent être pris en compte :

- ⋆ adéquation à la tâche : cet axe de l'évaluation, introduit dans le cadre du modèle de synthèse au Paragraphe 3.4.4.1, vise à estimer dans quelle mesure les problèmes de synthèse envisagés peuvent être exprimés sous forme d'une requête structurée,
- ⋆ pertinence interprétationnelle : cet axe de l'évaluation, présenté lui aussi en tant que partie du modèle de synthèse au Paragraphe 3.4.4.3, peut être envisagé selon plusieurs dimensions :
  - ⋆ utilité : il s'agit de déterminer si le document de synthèse proposé permet effectivement d'appréhender le corpus documentaire selon l'axe particulier décrit dans la requête,
  - ⋆ suggestivité : il faut évaluer si l'analyse du document de synthèse induit de nouvelles requêtes ou suggère d'autres études à réaliser avec d'autres outils,
  - ⋆ informativité : il faut aussi estimer si les résultats proposés apportent effectivement quelque chose de nouveau par rapport à l'existant.

Dans la suite de cette section, après une description des moyens utilisés pour l'étude, deux cas seront présentés, correspondant aux deux catégories de tâches

prototypiques disponibles au sein du prototype.

5.3.2 Moyens utilisés

Les études de cas qui seront présentées par la suite ont été réalisées dans le domaine applicatif des TMA, à partir de données disponibles au sein de la plateforme TMA-Explorer. Ce jeu de données va être présenté ici, conjointement aux problèmes biologiques qui serviront de support aux études de cas.

Dans le cadre de ses travaux sur les mécanismes d'oncogenèse, l'équipe du projet TMA-Explorer a sélectionné 162 patients souffrant d'un cancer du côlon et 119 patients atteints d'un cancer du sein, parmi ceux suivi au Centre Régional de Lutte Contre le Cancer de Montpellier. Les dossiers cliniques de la plupart de ces patients sont accessibles au sein de la base de données du projet. Des blocs de biopsie construits lors de l'ablation de leur tumeur dans des tissus tumoraux et des tissus supposés sains sont aussi disponibles pour des études histologiques.

Un jeu de biomarqueurs tumoraux a été révélé sur des lames histologiques complètes et des lames TMA construites à partir des biopsies des patients. Ces molécules étudiées ont été présentées en détails Paragraphe 1.3.2.3 et le Tab. 5.1 rappelle leurs caractéristiques principales.

Tab. 5.1: Résumé des faits biologiques connus à propos des molécules étudiées.

Molécule	*Compartiment cellulaire*	*Rôle*
β-caténine	Membrane, Cytoplasme, Noyau	Cellules normales : produite près de la membrane et détruite - Cellules tumorales : migre à travers le cytoplasme jusqu'au noyau où elle initie a production des cyclines
Cycline D1	Noyau, Cytoplasme	Produite dans le cytoplasme puis migre vers le noyau - Noyau : initie la division cellulaire et ainsi contribue à la prolifération cellulaire
Ki67	Noyau	Marque les cellules en division et est ainsi le témoin de la prolifération cellulaire
Bcl2	Cytoplasme	Empêche la mort cellulaire

Le pourcentage de cellules marquées ainsi que l'intensité de marquage et l'hétérogénéité du marquage ont été évalués par le pathologiste de l'équipe, le Dr. Joëlle Simony-Lafontaine du CRLCC de Montpellier, sur environ la moitié des lames.

Une fois ce contexte posé, il s'agit de définir des exemples de problèmes biologiques à analyser au sein de l'étude de cas. Étant donné que le prototype, pour l'instant, ne propose que deux tâches prototypiques, j'ai choisi de proposer un exemple pour chacune d'elles.

En ce qui concerne la comparaison, l'exemple proposé dans l'étude de cas est celui qui a servi de base aux deux chapitres précédents : «comparaison du pourcentage de cellules marquées entre les différentes molécules étudiées, entre tissus tumoraux et adjacents à la tumeur, en fonction de la localisation intracellulaire du marquage, chez les patients atteints d'un cancer du côlon». L'intérêt de cette étude ayant été explicité Paragraphe 3.2.3.5, ce point ne sera pas évoqué plus avant.

Les données histologiques et les éléments de diagnostic sont bien représentés au sein de l'étude de comparaison. Aussi, en ce qui concerne l'évolution, j'ai choisi de plutôt me focaliser sur des informations du dossier clinique, et en particulier des éléments de pronostic, mais toujours chez des patients atteints d'un cancer du côlon.

Or, pour le cancer du côlon, le stade pTNM de la tumeur est l'élément central du pronostic. Cette classification inclut trois composantes, T, N et M, qui chacune estiment une dimension différente du taux d'envahissement tumoral, selon les critères présentés dans le Tab. 5.2. Ainsi, plus les valeurs de T, N et M sont élevées, plus le pronostic est mauvais pour le patient.

Tab. 5.2: Explicitation de la classification pTNM des tumeurs du côlon - Les composantes T, N et M de la classification correspondent à un envahissement évalué selon des axes différents, à différents degrés.

Composante	*Valeur*	*Description*
T - Envahissement tissulaire progressif, de la lumière de l'intestin vers la périphérie		
	0	pas de tumeur dans l'échantillon
	Tis	intra-épithéliale ou chorion
	1	sous-muqueuse
	2	musculeuse
	3	sous-séreuse
	4	tumeur envahissant la séreuse ou un organe de voisinage
N - Envahissement des ganglions lymphatiques à la périphérie du côlon, estimé en n observant au moins 12 ganglions		
	0	Pas de métastases ganglionnaires
	1	1 à 3 ganglions envahis
	2	plus de 4 ganglions envahis
M - Envahissement à distance, par des métastases		
	0	Pas de métastases
	1	présence de métastases

La question qui peut alors se poser est celle de l'indépendance entre les diverses dimensions de la classification. On peut par exemple se demander s'il y a une relation entre le taux d'envahissement ganglionnaire et l'envahissement tissulaire.

Aussi, le second exemple d'étude sera l'«évolution du nombre de ganglions envahis en fonction du nombre de ganglions observés avec une visualisation de la composante T du stade chez les patients atteints d'un cancer du côlon».

La suite de la section va alors proposer, pour chacun des deux exemples, une exploration des problématiques d'évaluation évoquées : adéquation à la tâche, utilité des résultats, suggestivité des résultats et informativité des résultats. Cette dernière dimension de l'évaluation va être menée par comparaison avec un autre outil, différent pour chacune des tâches prototypiques considérées.

5.3.3 Exemple de comparaison

5.3.3.1 Adéquation à la tâche

Il s'agit ici de déterminer de manière qualitative si l'exemple d'étude envisagé peut s'exprimer sous une forme de requête structurée telle qu'envisagée au sein du prototype. Ceci revient à déterminer quels groupes de mots de la description de l'étude en langue naturelle doivent être utilisés pour spécialiser chacun des rôles de la requête structurée.

Les rôles de type «Généralités» sont facilement définissables à partir de la description de l'étude.

Ensuite, il s'agit d'une requête de type comparaison, et les rôles à spécialiser en fonction de l'étude à mener sont donc principalement les rôles de type «Besoins», c'est-à-dire le but, des critères d'inclusions, des critères de groupement et des critères de tri.

Cette décomposition en rôles peut être réalisée de la manière suivante :

- ⋆ But : il s'agit de l'objectif de l'étude, c'est-à-dire la variable dont la valeur va servir à la coloration des cases de la grille du document de synthèse. Ici, il s'agit du pourcentage de cellules marquées,
- ⋆ Critères d'inclusion : ces critères servent de base à la définition de la population à étudier, par sélection d'individus pertinents. Ici, il s'agit de patients atteints d'un cancer du côlon, ce qui suppose une localisation de la tumeur dans le côlon,
- ⋆ Critères de groupement : ces éléments permettent la constitution des groupes à comparer. Ici, le premier critère de groupement concerne les molécules étudiées, conduisant à la constitution d'un groupe pour chacune de ces quatre molécules. Le second critère est la localisation du tissu par rapport à la tumeur, induisant la construction au sein de chacun des groupes précédents, de deux groupes : un pour le tissu tumoral et l'autre pour le tissu adjacent à la tumeur. Le dernier critère est la localisation intracellulaire du marquage, conduisant à la division des groupes précédents en trois sous-groupes : membrane, cytoplasme et noyau,
- ⋆ Critères de tri : aucun critère n'est défini dans la description du problème,

aussi l'année de naissance a été choisie arbitrairement, pour la variété de valeurs qu'elle induit.

Enfin, les rôles correspondant aux «Contraintes expérimentales» ont majoritairement été laissés de côté, mis à part la largeur de grille, qui donne le meilleur rendu dans cet exemple avec une valeur de 50.

La décomposition de la description de l'étude en une formulation informelle peut alors être représentée par le Tab. 5.3.

Tab. 5.3: Formulation informelle d'un exemple de comparaison - Les différents rôles de la requête sont ici spécialisés en fonction de l'exemple d'étude : «comparaison du pourcentage de cellules marquées entre les différentes molécules étudiées, entre tissus tumoraux et adjacents à la tumeur, en fonction de la localisation intracellulaire du marquage, chez les patients atteints d'un cancer du côlon».

Élément du modèle	*Description*
Généralités :	
- Tâche	Comparaison
- Titre	Exemple d'illustration d'une tâche de type comparaison
- Description	Comparaison du pourcentage de cellules marquées entre les différentes molécules étudiées, entre tissus tumoraux et adjacents à la tumeur, en fonction de la localisation intracellulaire du marquage, chez les patients atteints d'un cancer du côlon
- Domaine	TMA
Besoins :	
- But	Pourcentage de cellules marquées
- Critères d'inclusion	Localisation de la tumeur dans le côlon
- Critères de groupement	Molécules étudiées - Localisation du tissu par rapport à la tumeur - Localisation intracellulaire du marquage
- Critères de tri	Année de naissance
Contraintes expérimentales :	
- Géométrie	Largeur de grille : 50

Dans le cadre de cet exemple précis, la formulation de la requête, à partir de la description d'un problème biologique de comparaison, apparaît donc comme matériellement possible.

5.3.3.2 Utilité des résultats

Une fois que la requête structurée a été définie, elle sert de base à la construction d'un document de synthèse, selon le processus qui a été décrit plus en détails au sein du Chapitre 4. La question qui se pose alors est de déterminer si l'affichage

graphique des résultats peut apporter des informations intéressantes au biologiste, dans une perspective d'appréhension des données.

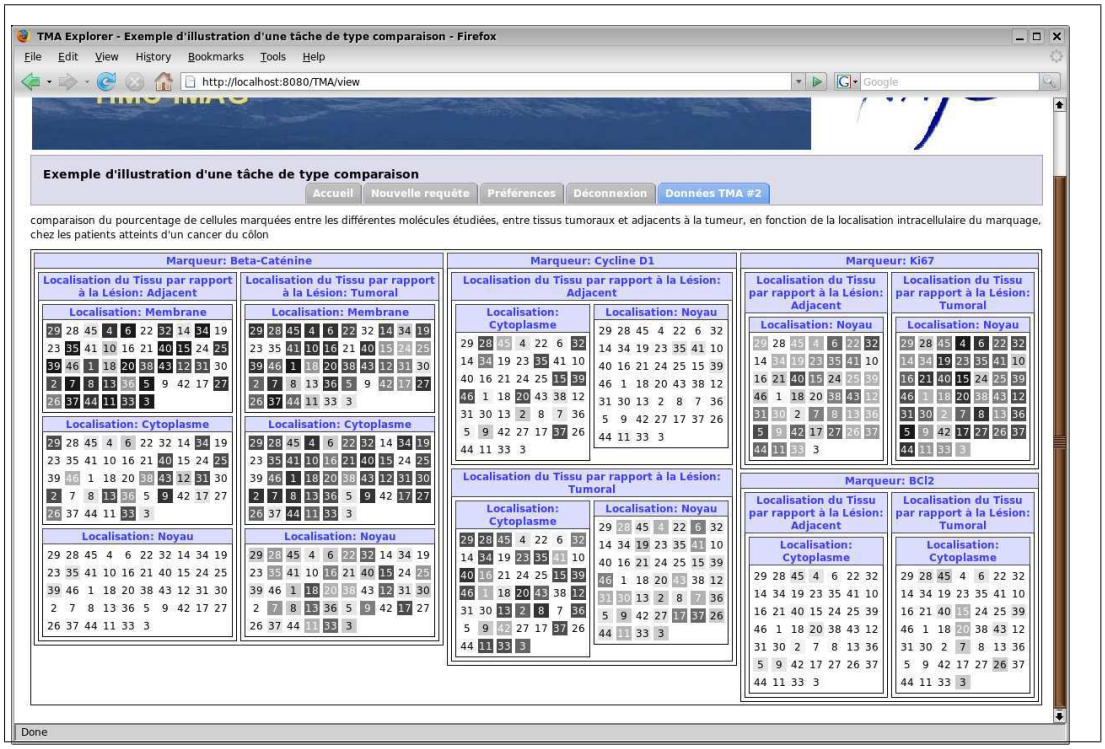


Fig. 5.1: Grille d'un document de synthèse pour un exemple de comparaison - Cette grille représente chaque individu par une case dont le niveau de gris correspond au pourcentage de cellules marquées et dont la valeur est un identifiant de l'individu au sein de la grille, ici le patient. Ces cases sont groupées en zones dont les titres décrivent le contenu, c'est-à-dire le critère qui a permis la constitution du groupe, ici les divers marqueurs, la localisation du tissu par rapport à la lésion et les compartiments intracellulaires. Au sein de chaque zone, les cases représentant les individus sont ordonnées de gauche à droite et de haut en bas selon les valeurs de la variable définie comme critère de tri, c'est-à-dire l'année de naissance.

Des copies d'écran d'un tel document de synthèse ont été présentées Fig. 4.14. Le cœur de ce document est constitué par la grille documentaire, qui est le résultat du processus de synthèse en tant que tel. Cette grille documentaire est présentée pour l'étude considérée Fig. 5.1.

Cette grille documentaire représente chaque individu par une case dont le niveau de gris correspond au pourcentage de cellules marquées (but de la comparaison), avec une douzaine de niveaux, de 0 (blanc) à 100% (noir), et dont le numéro est un identifiant interne du patient. Ces cases sont groupées au sein de zones qui sont identifiées par un titre décrivant leur contenu. Chaque case joue le rôle de lien vers une vue patient, qui elle-même contient des liens vers les informations histologiques correspondantes.

Une observation d'ensemble rapide laisse apparaître des différences majeures entre marqueurs : certains sont présents pour toutes les localisations intracellulaires,

alors que d'autres sont limités au noyau des cellules par exemple. Il apparaît aussi des différences de marquage entre cellules normales et tumorales pour tous les marqueurs, ce qui suggère de s'intéresser de manière indépendante à chacun des groupes de plus haut niveau, le niveau marqueur.

La Fig. 5.2 consiste en un focus sur le groupe du marqueur β-caténine.

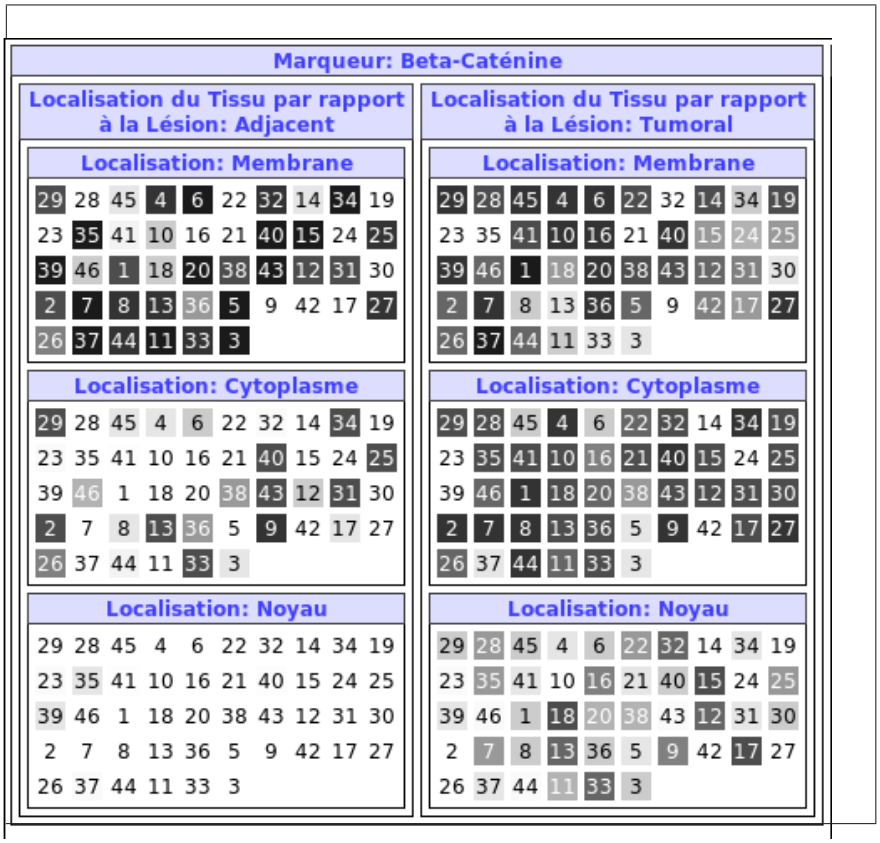


Fig. 5.2: Groupe du marqueur β-caténine pour l'exemple de comparaison - Cette figure consiste en une vue partielle de la grille du document de synthèse, centrée sur le marqueur β-caténine.

Au sein de cette zone, on peut noter que pour les cellules situées dans du tissu adjacent à la tumeur (à gauche) la localisation de la β-caténine est majoritairement dans la membrane, alors que pour les cellules tumorales (à droite), le marquage est présent dans tous les compartiments cellulaires. Or, les connaissances biologiques indiquent que dans les cellules normales, la β-caténine est exprimée dans la membrane, puis dégradée dans le cytoplasme. Dans les cellules tumorales, des mutations ou/et d'autres événements antérieurs dans le réseau moléculaire de la β-caténine empêchent cette dégradation et la β-caténine migre alors à travers le cytoplasme jusqu'au noyau où elle augmente la prolifération cellulaire en activant la transcription de gènes (tels que ceux qui codent les molécules de la famille des cyclines). En conséquence, l'affichage est cohérent avec les connaissances biologiques.

La Fig. 5.3 est elle centrée sur le groupe de la Cycline D1.

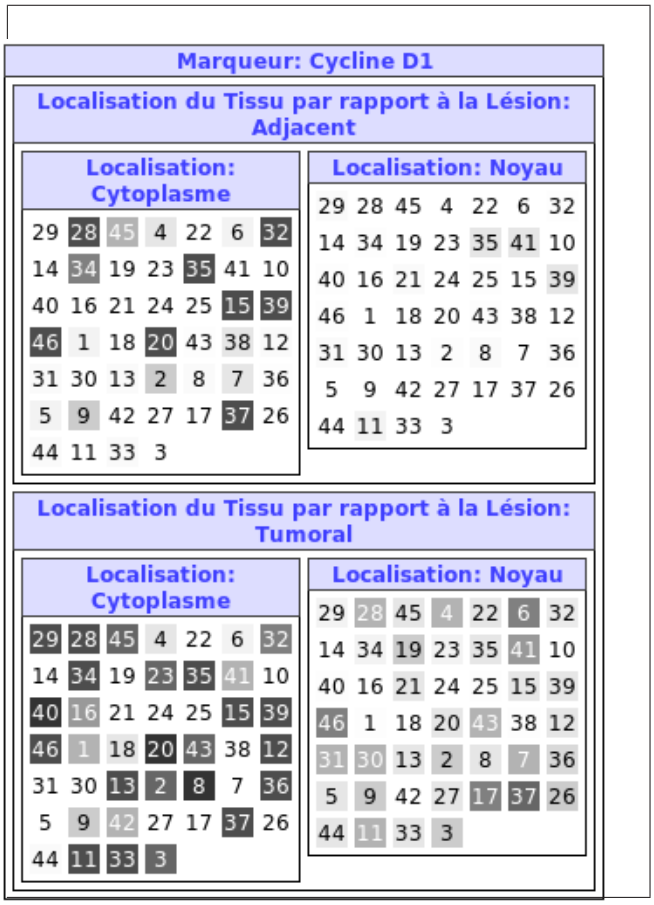


Fig. 5.3: Groupe du marqueur Cycline D1 pour l'exemple de comparaison - Cette figure consiste en une vue partielle de la grille du document de synthèse, centrée sur la Cycline D1.

Au sein du groupe de la Cycline D1, on peut remarquer que pour les cellules adjacentes à la tumeur (en bas), la localisation intracellulaire de la Cycline D1 est principalement dans le cytoplasme, alors que pour les cellules tumorales, le marquage est aussi présent dans le noyau. D'un point de vue biologique, il est connu que la Cycline D1 contribue à l'initiation de la mitose. De plus hauts pourcentages de cellules marquées en Cycline D1 impliquent donc un taux de mitoses plus élevé et un niveau de prolifération plus important. Ceci est en adéquation avec les tissus tumoraux où la division cellulaire est rapide et anarchique. Ici encore, l'affichage est en correspondance avec les connaissances du domaine d'étude.

La Fig. 5.4 est focalisée sur les groupes du marqueur Bcl2 et du marqueur Ki67.

Au sein du groupe correspondant au Ki67 (en haut), la coloration implique un plus grand nombre de noyaux marquées dans le tissu tumoral (à droite) que dans le tissu adjacent à la tumeur (à gauche). On peut s'attendre à ce marquage plus important dans le tissu tumoral, puisque les tumeurs sont connues pour leur taux de croissance cellulaire plus important.

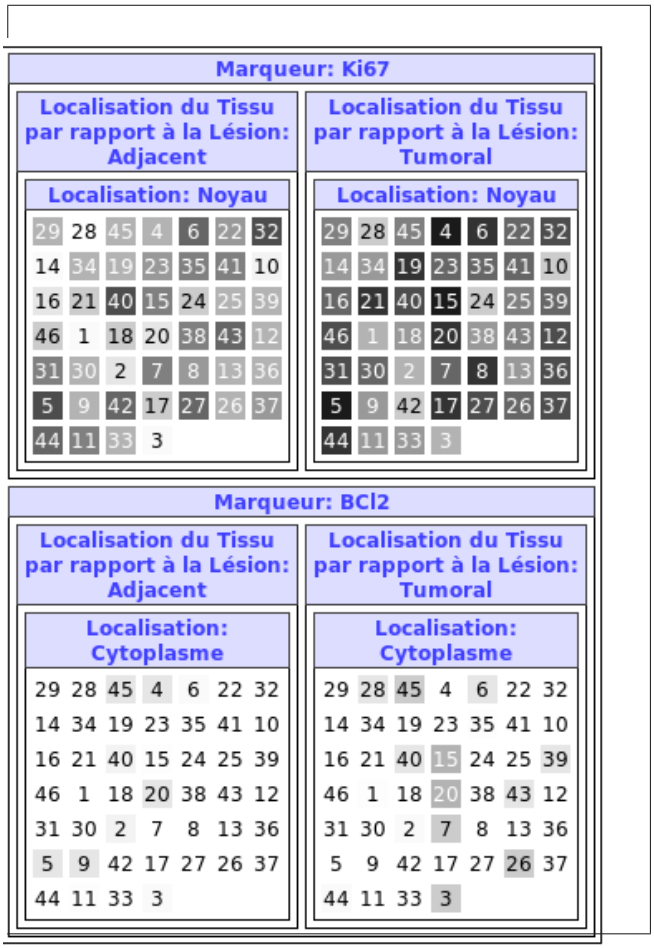


Fig. 5.4: Groupes des marqueurs Ki67 et Bcl2 pour l'exemple de comparaison - Cette figure consiste en une vue partielle de la grille du document de synthèse, centrée sur les marqueurs Ki67 et Bcl2.

Pour le Bcl2 (en bas), le marquage est faible mais on peut quand même noter que la répartition du marquage est plus étendue dans le tissu tumoral (à droite) que dans le tissu adjacent à la tumeur (à gauche). De plus, il est connu que cette molécule inhibe l'apoptose. Une grille documentaire présentant un marquage accru en Bcl2 dans le tissu tumoral, qui impliquerait un taux réduit de mort cellulaire, correspond aussi aux connaissances biologiques.

La grille documentaire du document de synthèse offre ainsi un aperçu à la fois simple et en phase avec des informations biologiques connues. C'est un résultat important qui suggère que de nouvelles hypothèses pourraient être explorées par ce biais, en tant que préalable à une analyse statistique par exemple.

Le couplage de cette visualisation avec des outils de navigation associés à l'espace documentaire construit offre de nouvelles perspectives et pourrait aussi aider à suggérer de nouvelles explorations de données.

5.3.3.3 Suggestivité des résultats

5.3.3.3.1 Introduction

La notion de suggestivité recouvre le fait que l'observation de la grille induise chez l'usager de nouvelles idées d'études à réaliser. Cette suggestivité peut être envisagée selon deux perspectives.

D'un part, l'usager peut être poussé à utiliser les résultats de la synthèse pour réaliser des analyses avec d'autres outils, par exemple pour valider statistiquement les conclusions biologiques dont l'observation de la grille lui a donné l'intuition.

D'autre part, les conclusions qu'il a tirées peuvent l'amener à formuler de nouvelles hypothèses, qu'il va tester en reformulant sa requête afin de refléter la nouvelle étude qu'il veut réaliser pour étudier sa nouvelle hypothèse.

Ces deux directions de la suggestivité vont être explorées plus avant ici.

5.3.3.3.2 Recours à d'autres outils

Grâce au document de synthèse qui a été construit, il a été possible de tirer des conclusions biologiques à partir du jeu de données à explorer, sans avoir à recourir à aucun outil de fouille de données, qui en général nécessitent une certaine expertise pour être utilisés correctement. Mais même si ces conclusions sont en accord avec des connaissances biologiques préalables, il est en général nécessaire de les confirmer en réalisant une analyse statistique sur les données.

Un test de Wilcoxon pour les données couplées a donc été utilisé pour comparer les groupes entre tissus tumoraux et adjacents à la tumeur. Les résultats de ce test sont présentés Tab. 5.4.

Ce test statistique montre des différences significatives entre les tissus tumoraux et adjacents à la tumeur pour les mêmes groupes que ceux qui ont été identifiés au sein du document de synthèse. Le paradigme de synthèse peut donc être être utilisé comme une méthode exploratoire préliminaire pour pré-valider des hypothèses avant de recourir à des outils statistiques gourmands en temps qui demandent une expertise importante.

Tab. 5.4: Résultat du test de Wilcoxon pour les données couplées - Ce test compare échantillon par échantillon les différents groupes entre tissus tumoraux et tissus adjacents à la tumeur. Il fait l'hypothèse que les individus dans les deux groupes sont identiques en ce qui concerne le pourcentage de cellules marquées et le but est de valider ou rejeter l'hypothèse. z est une mesure de la similarité entre les deux groupes pour la variable testée et la mesure $Prob|z|$ évalue la probabilité que l'hypothèse d'identité soit vraie.

Molécule	*Localisation intracellulaire*	*z*	*Prob|z|*	*Conclusion*
β-caténine	Noyau	5.320	0	différence significative
	Cytoplasme	4.484	0	différence significative
	Membrane	-1.026	0.3049	pas de différence significative
Cycline D1	Noyau	5.501	0	différence significative
	Cytoplasme	2.894	0.0038	différence significative
Ki67	Noyau	5.357	0	différence significative
Bcl2	Cytoplasme	1.408	0.1592	différence significative

5.3.3.3.3 Reformulation de requête

Conjointement, la vue sur les données qui a été présentée apporte beaucoup d'informations mais elle reste trop générale pour des investigations spécialisées parce que le champ d'étude choisi (l'oncogenèse dans les cancers du côlon) est trop étendu. Par contre, cette vue suggère des focalisations pour des études futures.

Par exemple, le stade de la tumeur a été défini comme un outil de pronostic. En particulier, la composante N du stade, qui correspond à l'envahissement ganglionnaire, est un signe de l'agressivité de la tumeur. On peut alors se demander si cette agressivité peut être corrélée avec le niveau de marquage pour chacune des molécules étudiées. Pour cette étude, on va se focaliser sur le marqueur β-caténine.

Le problème biologique devient alors : «comparaison du pourcentage de cellules marquées entre tissus tumoraux et adjacents à la tumeur, en fonction de la localisation intracellulaire du marquage et en fonction de la composante N du stade, chez les patients atteints d'un cancer du côlon, pour le marqueur β-caténine».

La requête correspondante peut être formulée comme relevant d'une tâche de «comparaison». Son but reste l'expression d'un marquage, par la mesure du pourcentage de cellules marquées. La population de l'étude (c'est-à-dire les critères d'inclusion) consiste toujours en les patients atteints d'un cancer du côlon, mais s'y ajoute un marqueur spécifique, la β-caténine. Ensuite, aux groupes à comparer (les critères de groupement) déjà définis (marqueur, localisation par rapport à la lésion et localisation intracellulaire du marquage), s'ajoute la composante N du stade.

La grille documentaire correspondante est présentée Fig. 5.5.

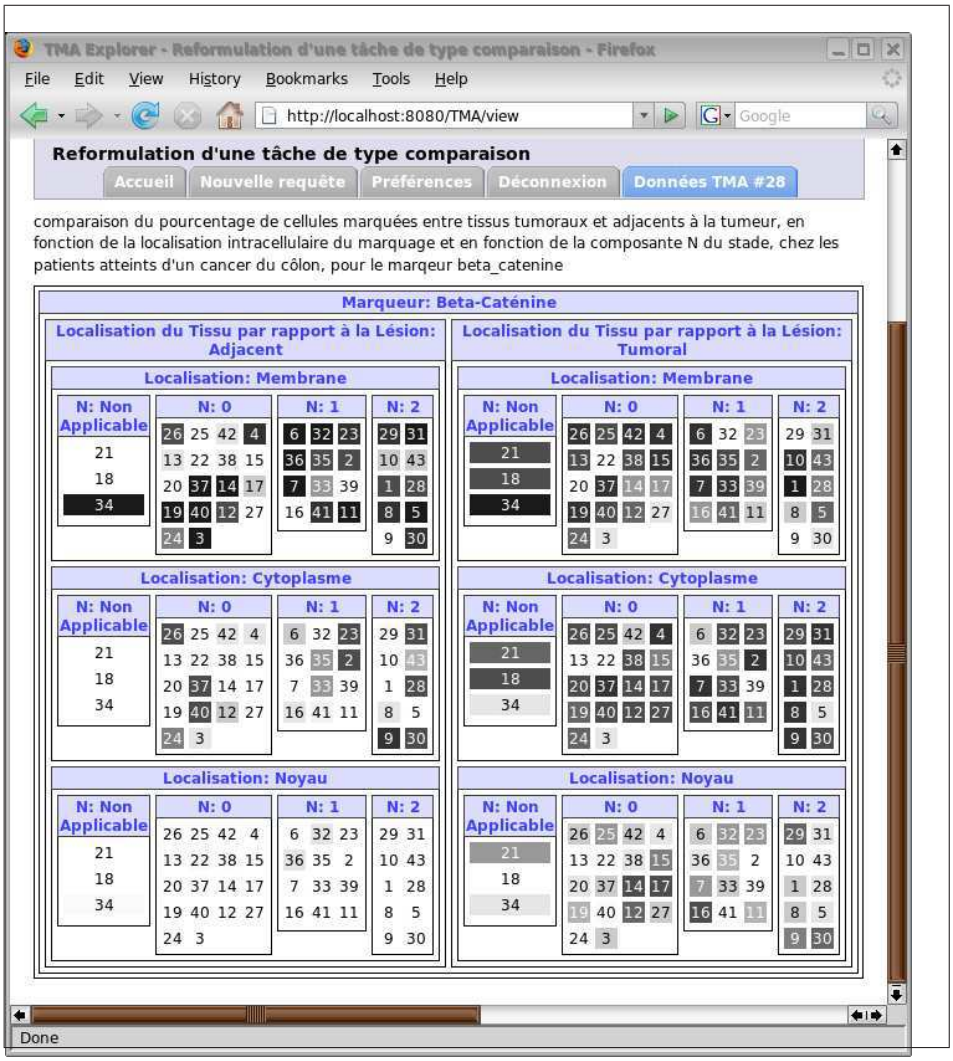


Fig. 5.5: Grille documentaire pour la comparaison reformulée - Cette reformulation conduit à la disparition des groupes de niveau marqueur mais fait apparaître de nouveaux groupes de niveau composante N du stade.

Cet affichage laisse toujours apparaître les différences qui ont été notées entre tissu tumoral et tissu adjacent à la tumeur. Mais la composante N du stade ne semble pas être liée avec un marquage plus important, et la répartition des individus à pourcentages de cellules marquées en β-caténine variables semble plus ou moins homogène entre les différentes valeurs de N.

Biologiquement, il semble donc qu'il n'y ait pas de relation entre un pourcentage de cellules marquées en β-caténine élevé et un envahissement ganglionnaire important. On peut donc en déduire que ce sont d'autres molécules que la β-caténine qui sont impliquées dans des phénomènes d'infiltration de la tumeur. Ceci suggère donc de s'intéresser à d'autres molécules au cours d'études futures.

5.3.3.4 Informativité des résultats

5.3.3.4.1 Introduction

Les précédents paragraphes ont permis de montrer de façon qualitative l'adéquation à certaines tâches de comparaison du modèle de synthèse envisagé, l'utilité du document de synthèse pour inférer des connaissances biologiques et comment un tel document peut suggérer des analyses avec d'autres outils ou de nouvelles requêtes.

La question qui se pose alors est l'intérêt d'un tel système dans l'absolu et l'apport des résultats proposés par rapport à un autre outil permettant la construction d'une vue similaire. Afin d'explorer cette problématique, la méthode choisie est une comparaison des résultats du système de synthèse avec ceux d'un autre outil qui permet de construire des représentations proches.

Pour évaluer la validité des études relevant d'une problématique de comparaison, la confrontation est menée avec l'outil Treemap[2], de l'Université du Maryland, dont le principe est présenté dans [Shneiderman, 1992].

Dans la suite de ce paragraphe va être présenté comment une visualisation similaire à celle proposée par le document de synthèse peut être construite avec Treemap. Une comparaison critique entre les deux constructions va alors être menée.

5.3.3.4.2 Construction d'une visualisation avec Treemap

L'outil Treemap permet la construction de visualisations compactes de hiérarchies au sein d'un espace découpé en zones rectangulaires dont la surface et la couleur de fond peuvent être associées à des variables d'intérêt. Il se base sur des fichiers textes particuliers, au format très simple. La première étape de l'utilisation de cet outil est donc la construction d'un fichier d'entrée à partir de la base de données du projet. Pour ce faire, le contenu de la vue histologie a été déchargé dans un fichier texte, en prenant soin d'éliminer tous les individus ayant des données manquantes, Treemap ne supportant pas les champs vides. Cette tâche a pris à peu près une demi-heure.

Ensuite, le fichier texte généré est chargé au sein de l'outil. Le résultat est similaire à la copie d'écran de la Fig. 5.6. On peut noter que le résultat du chargement est une grille «vierge», où aucune hiérarchie ni aucune légende ne sont définies. L'affichage préliminaire n'est donc absolument pas guidé par une étude à réaliser et la construction de la visualisation est réalisée par le biais de manipulations a posteriori.

[2] http ://www.cs.umd.edu/hcil/treemap/

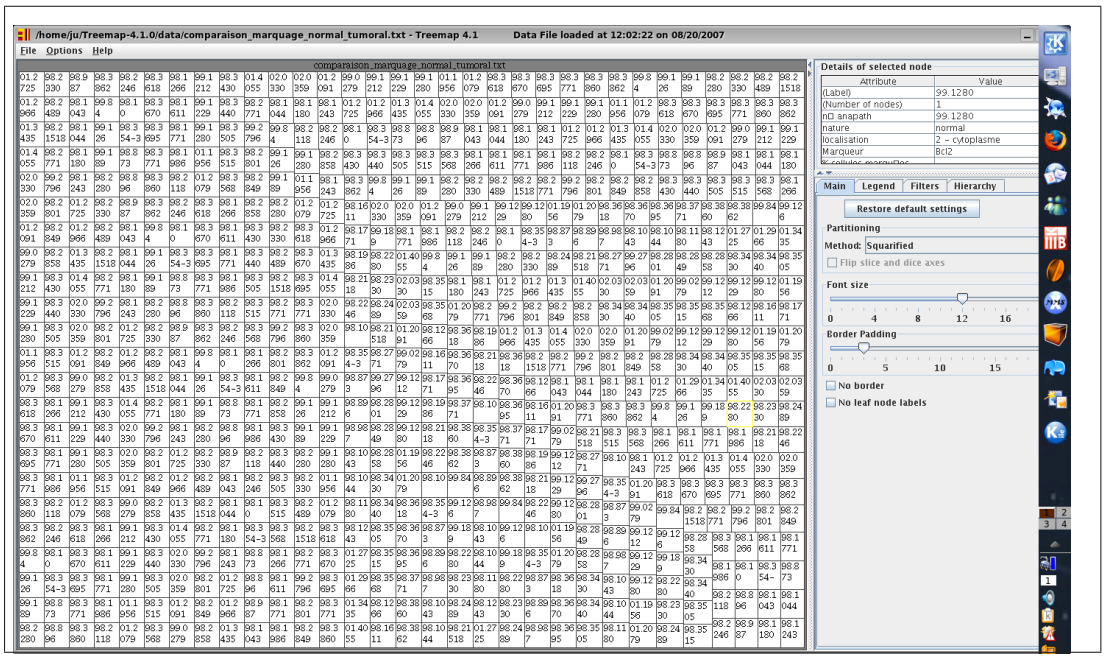


Fig. 5.6: Chargement des données dans Treemap - Après chargement des données, chaque individu est représenté par une case sans aucune organisation et sans code couleur.

La première étape de construction de la vue Treemap est la définition d'une légende. Celle-ci correspond au choix d'une variable à utiliser pour définir la couleur de fond de chaque case, réalisée par le biais d'une liste déroulante contenant la liste des variables du fichier, ainsi qu'au choix du code couleur correspondant, par le biais d'un sélecteur de couleur au sein d'une palette. Dans le contexte de la synthèse, ceci reviendrait à définir un but à l'étude, ainsi que des couleurs au sein des connaissances expérimentales du domaine.

Le résultat de la définition de cette légende correspond à la copie d'écran de la Fig. 5.7. Chaque case est alors colorée en fonction du pourcentage de cellules marquées correspondant, de 0 (blanc) à 100% (noir).

La dernière étape est la constitution d'une hiérarchie. Celle-ci correspond à la définition d'une liste ordonnée de variables qui seront utilisées pour constituer des groupes puis des sous-groupes, etc. Dans le contexte de la synthèse, ceci reviendrait à définir des critères de groupement.

Le résultat de la définition de cette hiérarchie correspond à la copie d'écran de la Fig. 5.8. Les individus sont groupés dans une premier temps en quatre zones correspondant aux divers marqueurs. Chaque zone est alors divisée en deux zones correspondant aux tissus tumoraux et adjacents à la tumeur. Ensuite, des zones sont définies pour chaque localisation intracellulaire.

Ce découpage est réalisé pour utiliser au mieux l'espace disponible, aussi la vi-

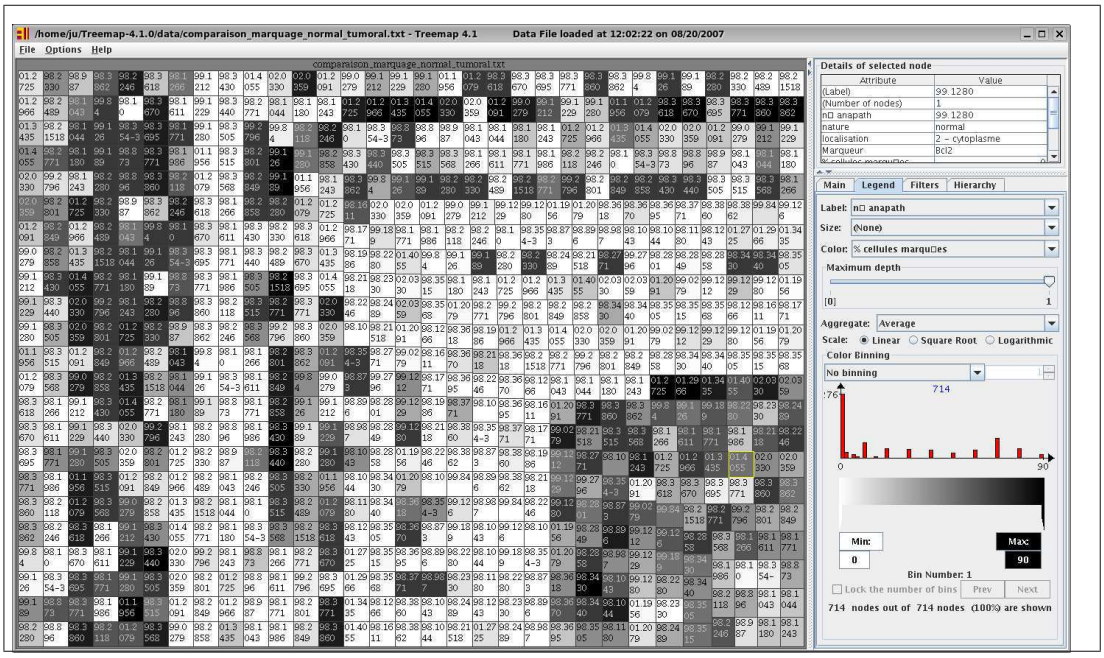


Fig. 5.7: Définition d'une légende dans Treemap - La définition d'une légende permet d'indiquer la variable qui sert de base à la coloration, ainsi que le code couleur à utiliser.

sualisation dépend de la taille de la fenêtre de l'application. Ici, cette taille a été ajustée pour donner le résultat le plus similaire à celui de la grille synthétique de la Fig. 5.1. On peut aussi noter que la taille des cases individuelles n'est pas constante, afin d'occuper tout l'espace disponible.

Treemap ne permet pas de définir un ordonnancement des individus au sein d'une zone de la grille. Il n'y a donc pas d'équivalent aux critères d'ordonnancement de la synthèse.

Par contre, la taille des cases peut être associée à la valeur d'une variable. Cette fonctionnalité n'a pas été utilisée ici.

5.3.3.4.3 Comparaison critique

Les visualisations proposées par Treemap et par construction d'un document de synthèse apparaissent comme similaires. Il est alors légitime de se demander ce que le système d'assistance à la synthèse apporte de plus par rapport à un outil tel que Treemap.

La différence est tout d'abord purement conceptuelle. Treemap est un outil de Visualisation d'Information, focalisé sur la présentation compacte en deux dimensions de hiérarchies. En tant que tel, il n'est adapté que pour comparer des groupes et la

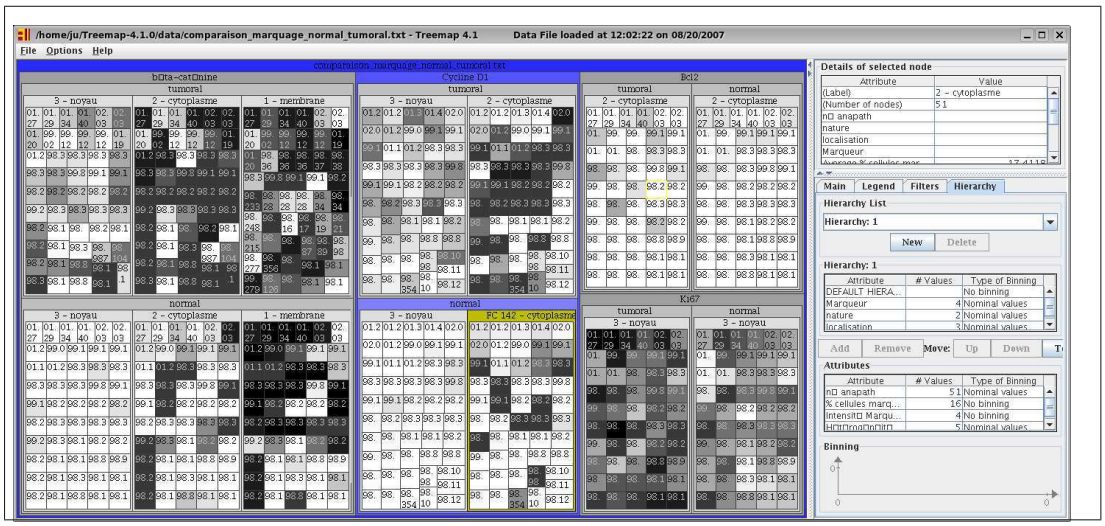


Fig. 5.8: Définition d'une hiérarchie dans Treemap - La définition d'une hiérarchie permet la constitution de groupes et sous-groupes selon des valeurs de variables définies par l'utilisateur.

représentation de grilles correspondant à des problématiques de type évolution ou distribution se seraient pas possibles. Le système d'assistance à la synthèse se veut un système de Recherche d'Information orientée tâche, qui permet l'exploration de multiples tâches et donc types de documents de synthèse.

Ensuite, dans Treemap, le focus sur le jeu de données est fourni par la sélection manuelle réalisée lors de la construction du fichier de données. Un changement de focus induit en général la construction d'un nouveau fichier spécifique. Ceci ne facilite donc pas l'exploration du corpus documentaire : la réalisation de nouvelles études suggérées par l'observation du résultat conduit en général à une tâche longue et fastidieuse de construction d'un fichier adapté. Par contre, dans le cadre de la synthèse, le focus est décrit explicitement dans la requête, par l'ensemble des critères définis comme «Besoins». Changer de focus consiste alors en une reformulation, activité simple et rapide à mettre en œuvre.

Puis, la visualisation au sein de Treemap est construite a posteriori, par le biais d'une définition d'une légende et d'une hiérarchie. L'exploration du jeu de données n'est donc pas dirigée. Par contre, l'entrée dans le système de synthèse par une requête permet de spécifier a priori un objectif pour cette exploration, une étude qui sert de cadre au traitement. L'analyse d'une hypothèse est alors explicite, au lieu d'être implicite pour Treemap.

Par la suite, la navigation dans l'espace documentaire est très contrainte dans Treemap. Seules les informations présentes dans le fichier en entrée peuvent être accédées. Par contre, le document de synthèse est envisagé en contexte, et les informations connexes sont disponibles par l'intermédiaire des vues patient et histologie. Ceci permet une plus grande flexibilité dans l'exploration d'une problématique biologique.

Ces diverses remarques font apparaître une différence fondamentale entre les deux approches : la composante demandant le plus gros effort réflexif de la part de l'utilisateur. Ainsi, pour Treemap, la construction du fichier en entrée est une activité qui implique une planification pesée puisqu'elle conditionne ce qu'il sera possible de visualiser par la suite. Au sein du système de synthèse, l'effort cognitif est déporté vers la formulation de requête, mais les erreurs ne sont pas trop coûteuses en temps, puisqu'elles induisent une simple reformulation.

Il apparaît donc que la construction d'un document de synthèse pour une tâche prototypique de type comparaison permet une exploration dirigée d'un jeu de données plus souple que la visualisation proposée par un outil de type Treemap, grâce à l'entrée dans le système par une requête structurée et aux possibilités de navigation dans les données apportées par les vues associées à la grille documentaire.

5.3.3.5 Un premier cas instructif

L'étude de cas qui vient d'être présentée a permis de montrer de manière qualitative un certain nombre de qualités des documents de synthèse relevant de problématiques de comparaison.

Tout d'abord, le problème biologique d'exemple a pu être exprimé sous forme de requête structurée, permettant l'entrée dans le système de synthèse. Ensuite, le document de synthèse a permis de mettre en lumière un certain nombre de faits biologiques connus, montrant ainsi que de nouvelles connaissances pourraient être suggérées par ce biais et permettant d'estimer cette construction comme utile.

De plus, l'observation de la grille a permis de suggérer une étude statistique et une reformulation de requête, donnant ainsi des indices sur la suggestivité de la construction.

Enfin, une comparaison avec l'outil Treemap a permis la mise en exergue des spécificités et de l'intérêt du processus de synthèse par rapport à un outil de Visualisation d'Information.

5.3.4 Exemple d'évolution

5.3.4.1 Adéquation à la tâche

Après l'analyse qui vient d'être présentée d'un exemple de tâche de comparaison, il s'agit de s'intéresser à une problématique d'évolution et de mener le même type d'exploration. l'objectif est alors de déterminer, comme pour l'exemple de compara-

ison, si l'étude envisagée peut être exprimée sous forme de requête structurée, ce qui revient à associer à chaque rôle de la requête une composante de l'étude. Dans le cas d'une tâche prototypique d'évolution, seuls les «Besoins» sont différents par rapport à un cas de comparaison, et seuls ceux-ci seront analysés en détails. La décomposition en rôle de cette partie «Besoins» devient alors :

- ⋆ But : il s'agit de l'élément à visualiser, soit ici la composante T du stade,
- ⋆ Première dimension : cet élément permet de définir en fonction de quoi l'évolution sera représentée, c'est-à-dire ici le nombre de ganglions observés,
- ⋆ Seconde dimension : ceci correspond à l'élément dont l'évolution est présentée, ici le nombre de ganglions envahis,
- ⋆ Critères d'inclusion : ces critères permettent de limiter la population étudiée ; ici la limite porte sur la localisation de la tumeur des patients dans le côlon.

Cette décomposition conduit alors à une formulation informelle de requête, présentée Tab. 5.5.

Tab. 5.5: Formulation informelle d'un exemple d'évolution - Les différents rôles de la requête sont ici spécialisés en fonction de l'exemple d'étude : «évolution du nombre de ganglions envahis en fonction du nombre de ganglions observés avec une visualisation de la composante T du stade chez les patients atteints d'un cancer du côlon».

Élément du modèle	*Description*
Généralités :	
- Tâche	Évolution
- Titre	Exemple d'illustration d'une tâche de type évolution
- Description	évolution du nombre de ganglions envahis en fonction du nombre de ganglions observés avec une visualisation de la composante T du stade chez les patients atteints d'un cancer du côlon
- Domaine	TMA
Besoins :	
- But	Composante T du stade
- Première dimension	Nombre de ganglions observés
- Seconde dimension	Nombre de ganglions envahis
- Critères d'inclusion	Patients atteints d'un cancer du côlon

Dans le cadre de cet exemple précis, la formulation de la requête, à partir de la description d'un problème biologique d'évolution, apparaît donc elle aussi comme possible, tout comme pour l'exemple de comparaison qui a été étudié précédemment.

5.3.4.2 Utilité des résultats

Maintenant que la requête structurée a été définie, le processus de synthèse peut être appliqué pour construire le document de synthèse correspondant. Comme dans l'exemple de comparaison, il s'agit alors d'évaluer dans quelle mesure le document

présenté à l'usager apporte des informations d'intérêt pour le problème exprimé dans la requête.

Une copie d'écran de la grille documentaire correspondante est présentée Fig. 5.9. Dans le cadre du prototype, la tâche d'évolution envisagée a été développée sous une forme simplifiée, ne permettant de présenter qu'un seul individu, l'individu moyen, pour chaque ensemble d'individus présents au même jeu de valeurs selon les deux dimensions de l'évolution. A l'avenir, l'objectif serait de proposer une présentation de plusieurs individus, dont le nombre serait défini par un nouveau rôle de la requête. La Fig. 5.10 illustre cette possibilité.

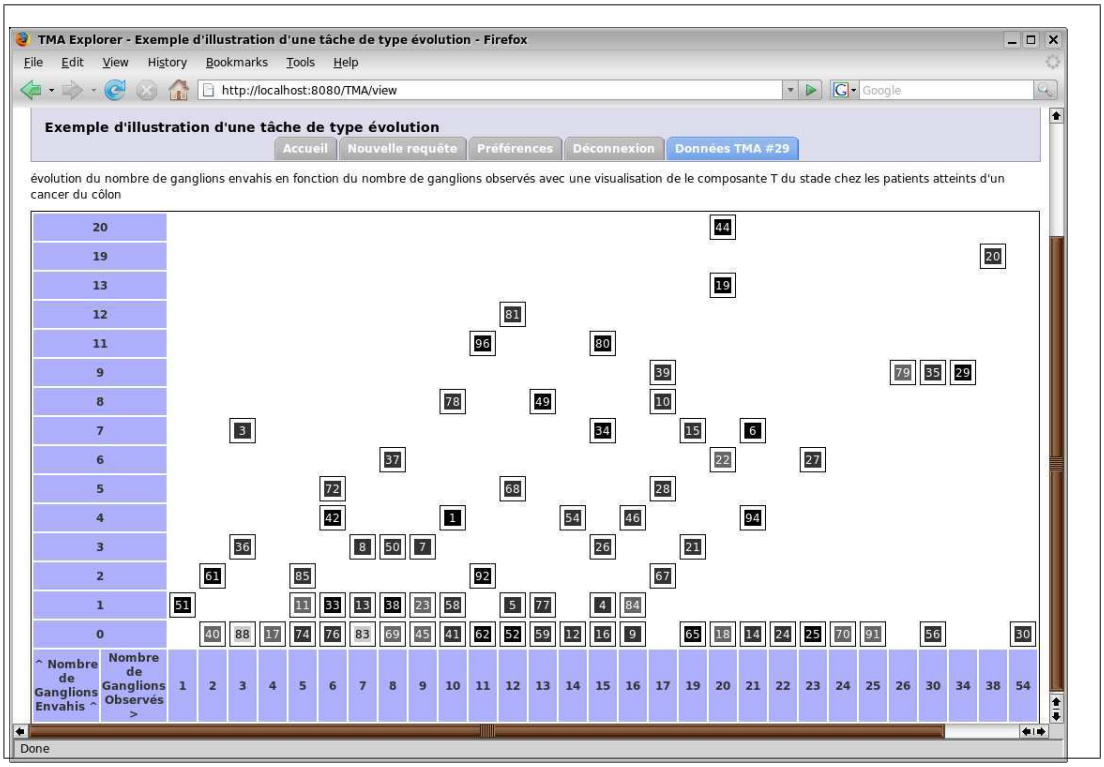


Fig. 5.9: Grille d'un document de synthèse pour un exemple d'évolution - Cette grille organise les éléments pertinents selon les deux axes définis dans la requête, le nombre de ganglions observés et le nombre de ganglions envahis. Pour chaque combinaison de valeurs selon ces deux axes, pour laquelle des individus existent, l'individu moyen du groupe correspondant est choisi et présenté au sein d'une case donc la couleur de fond est liée au but de l'étude (ici la composante T du stade), de gris clair (stade 1) à noir (stade 4). Le numéro présenté dans la case est celui de l'individu moyen et donne accès à sa fiche.

L'analyse des résultats peut alors être envisagée selon les deux axes qui ont été présentés comme objectif de l'appréhension des données : un préalable à une fouille de données et une validation d'hypothèses.

Selon le point de vue exploration préalable à une fouille, il s'agit d'observer la structure du jeu de données, par exemple afin d'identifier des données aberrantes. Ici, on observe le nombre de ganglions envahis en fonction du nombre de ganglions

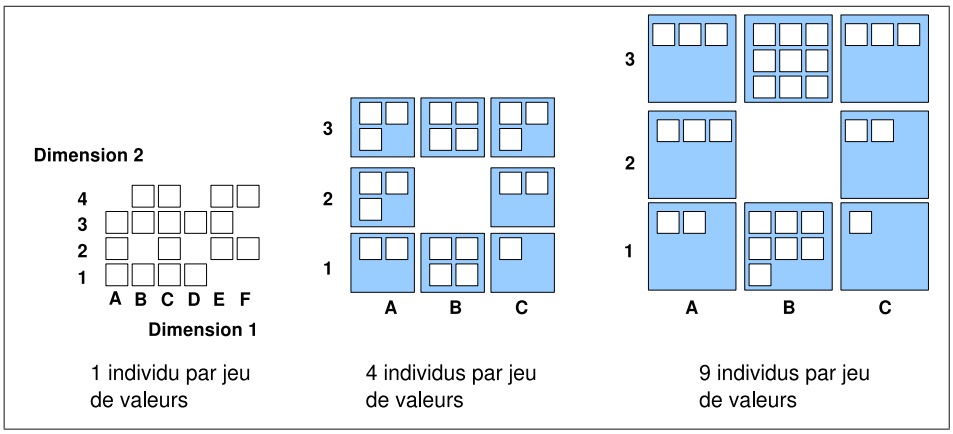


Fig. 5.10: Extension de la notion d'évolution - Actuellement, seul un individu moyen par jeu de valeurs des deux dimensions est présenté. À l'avenir, l'objectif serait d'en présenter plusieurs, où le nombre, ici par exemple 4 ou 9, serait défini par un nouveau rôle de la requête.

observés. Logiquement, il ne devrait pas y avoir plus de ganglions envahis qu'observés et tous les individus devraient être localisés sous la diagonale matérialisée sur la Fig. 5.11. Or ce n'est pas le cas pour le patient 3 : d'après son dossier clinique, 7 ganglions ont été jugés comme envahis parmi les 3 observés. Il s'agit typiquement d'une erreur de saisie.

Or il est important, avant de recourir à des méthodes de fouille de données plus évoluées, d'identifier de telles erreurs. En effet, leur rareté leur donne en général une valeur prédictive très élevée, conduisant à la définition de règles ou de modèles erronés. Le document de synthèse proposé permet donc, dans le cadre de cet exemple, une appréhension des données intéressante en tant que préalable à une fouille de données plus poussée.

D'un point de vue exploration du problème biologique, plusieurs remarques peuvent être formulées.

Tout d'abord, les bonne pratiques en matière de classification de tumeurs, définies par des organismes internationalement reconnus, recommandent l'observation d'au moins 12 ganglions. Or, au sein de la grille documentaire, on peut voir un certain nombre d'individus pour lesquels un nombre plus faible de ganglions ont été observés. Ces individus se trouvent à gauche de la ligne verticale de la Fig. 5.12. Il est alors légitime de se demander pourquoi.

Or, le comptage des ganglions est réalisé sur la pièce opératoire fraîche, juste après l'ablation de la tumeur, soit par l'anatomopathologiste, soit par le technicien qui réalise des biopsies d'archive. Plusieurs circonstances peuvent alors expliquer l'observation de moins de 12 ganglions.

Tout d'abord, il est possible que la personne en charge du comptage n'ait pas

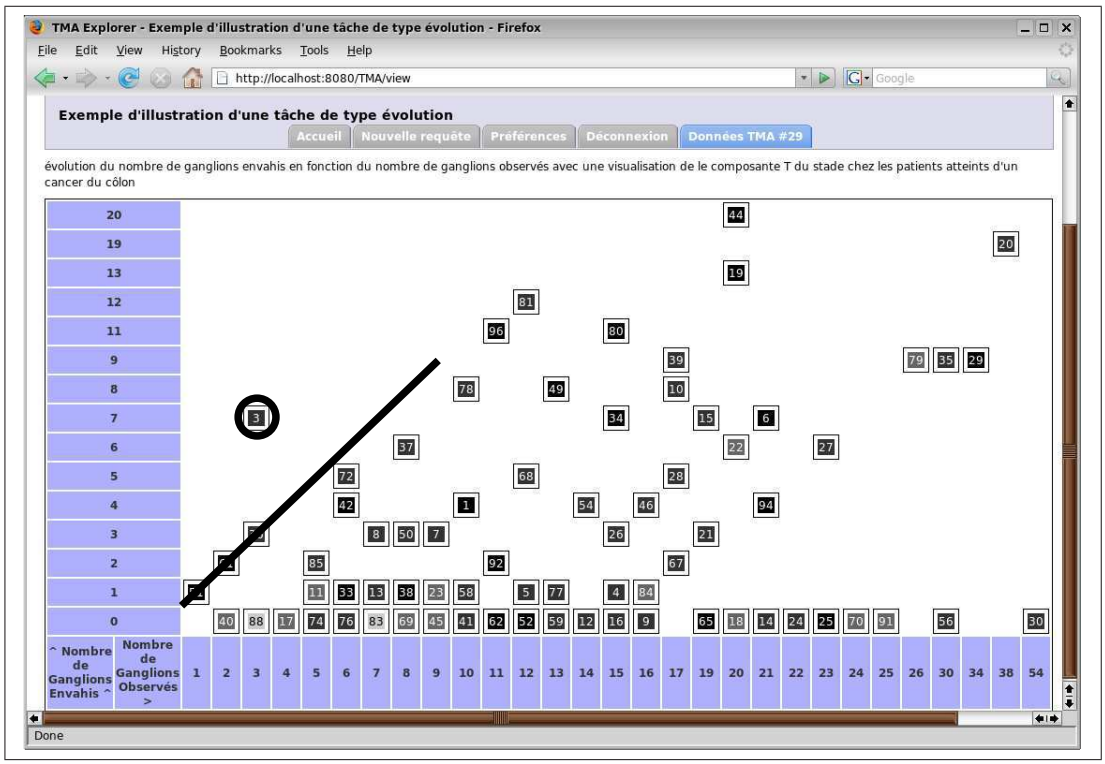


Fig. 5.11: Mise en évidence d'un individu aberrant pour l'exemple d'évolution - Logiquement, il ne peut pas y avoir plus de ganglions envahis que de ganglions observés. Tous les individus devraient être sous la diagonale marquée par le trait. Or le patient 3 est localisé au-dessus du trait. Il comporte des données aberrantes : 7 ganglions envahis sur 3 observés.

trouvé assez de ganglions pour en observer 12. Ensuite, l'intervention chirurgicale peut ne pas avoir assez enlevé de tissu graisseux périphérique pour que 12 ganglions soient inclus. Enfin, il est connu que certains traitements préopératoires, tels qu'une chimiothérapie ou une radiothérapie, provoquent une «fonte» des ganglions, qui restent alors en nombre limité. Il pourrait alors être intéressant d'évaluer quelle théorie est applicable pour les différents patients concernés.

Ensuite, on peut se focaliser sur les patients ayant une nombre de ganglions observés supérieur à 12, selon les recommandations, afin d'observer plus finement le lien entre taux d'envahissement ganglionnaire et envahissement tissulaire. Ceux-ci se trouvent donc à droite du trait vertical de la Fig. 5.12.

Parmi ces individus, on peut distinguer ceux ne présentant aucun envahissement ganglionnaire (sous le trait horizontal de la Fig. 5.12) et ceux où des ganglions sont envahis.

Il ne semble pas y avoir de différence notable au niveau de la composante T du stade entre les deux. Un fort taux d'envahissement ganglionnaire (individus localisés vers le haut de la Fig. 5.11), ne semble pas non plus lié à la composante T du stade. On peut donc supposer que les deux dimensions T et N de la classification pTNM

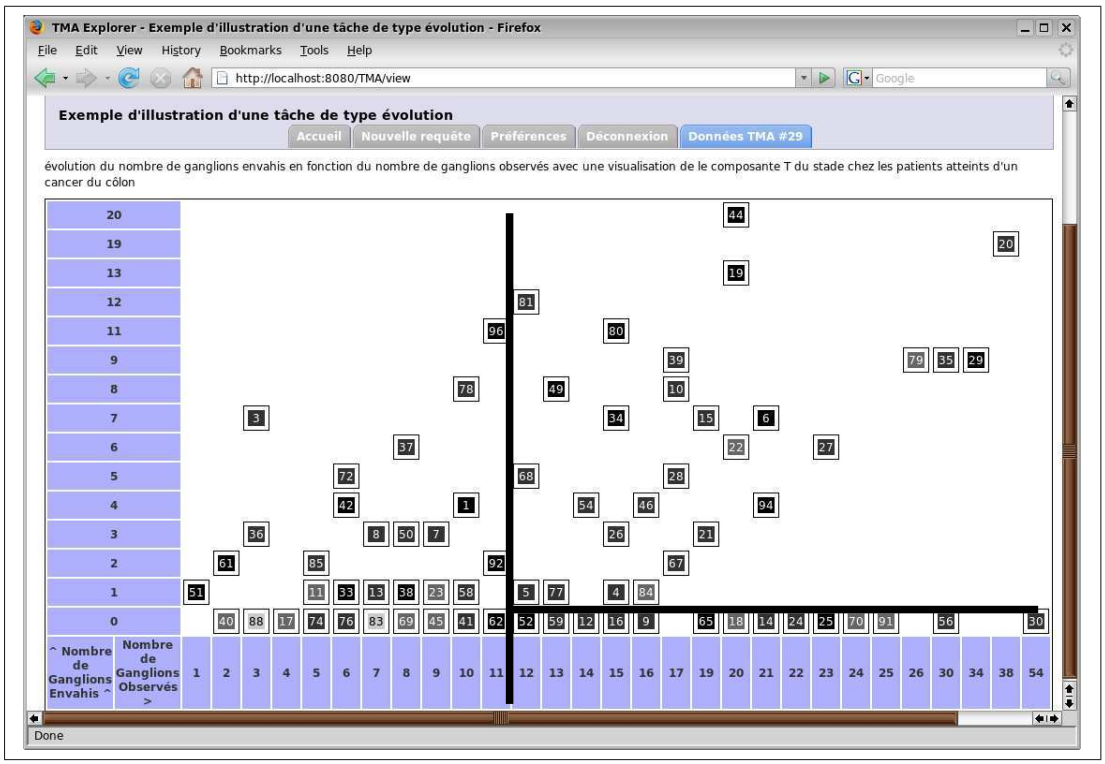


Fig. 5.12: Des pratiques de classification au sein de l'exemple d'évolution - Le trait vertical sépare les individus pour lesquels 12 ganglions au moins ont été observés (à droite). Le trait horizontal sépare les individus qui présentent un envahissement ganglionnaire (au-dessus) de ceux qui n'en présentent pas.

sont indépendantes, ce qui est en fait une information biologique connue.

Ici, comme pour l'exemple de comparaison, l'observation de la grille du document de synthèse permet d'illustrer des faits biologiques connus, ce qui suggère que ce type de tâche prototypique pourrait être utilisé pour évaluer des hypothèses nouvelles.

5.3.4.3 Suggestivité des résultats

5.3.4.3.1 Introduction

Dans le cadre de l'exemple de tâche d'évolution, cette notion de suggestivité va être uniquement envisagée selon la perspective de suggestion de nouvelles études à réaliser en utilisant le système d'assistance à la synthèse. Dans ce contexte, deux thématiques vont être explorées.

Tout d'abord, l'exploration des résultats menée au paragraphe précédent a fait apparaître des interrogations sur les patients dont moins de 12 ganglions ont été observés, contrairement aux recommandations communément admises. Des hypothèses

explicatives ont été émises et il serait pertinent de les explorer.

Ensuite, la grille a permis de suggérer que l'envahissement tissulaire (composante T du stade) n'est pas liée à l'envahissement ganglionnaire. Par contre, la composante N est directement évaluée à partir du nombre de ganglions envahis. Il pourrait alors s'avérer intéressant, dans une perspective de traque des données incohérentes, par exemple pour préparer une analyse statistique incluant les diverses dimensions du stade, de vérifier que les valeurs de N attribuées aux différents patients sont bien valides par rapport au nombre de ganglions envahis.

Ces deux problématiques correspondent à des reformulations de requête qui vont être explicitées dans la suite de ce paragraphe.

5.3.4.3.2 Explication de l'observation de moins de 12 ganglions

La première problématique découlant de l'exemple d'évolution présenté précédemment est celle de l'explication d'une observation de moins de 12 ganglions. L'une des hypothèses explicatives qui ont été émises est l'influence d'un traitement préopératoire de type chimiothérapie ou radiothérapie, qui peut provoquer une «fonte» des ganglions. C'est cette hypothèse qui va être testée ici.

Afin de tester cette hypothèse, il s'agit de proposer une reformulation de la requête. Ici, on s'intéresse à l'effet d'une chimiothérapie ou d'une radiothérapie préopératoire. L'information concernant la présence ou l'absence d'un tel traitement devient alors le but de l'étude, en remplacement de la composante T du stade. Présence d'une chimiothérapie préopératoire et présence d'une radiothérapie préopératoire sont deux concepts indépendants des connaissances du domaine, et en l'état du prototype, ils ne peuvent être étudiés conjointement. Leur étude conduit donc au final à l'expression de deux requêtes qui, en langage naturel, seraient :

- ⋆ Pour la chimiothérapie : «évolution du nombre de ganglions envahis en fonction du nombre de ganglions observés avec une visualisation de la présence ou absence d'une chimiothérapie préopératoire chez les patients atteints d'un cancer du côlon»,
- ⋆ Pour la radiothérapie : «évolution du nombre de ganglions envahis en fonction du nombre de ganglions observés avec une visualisation de la présence ou absence d'une radiothérapie préopératoire chez les patients atteints d'un cancer du côlon».

Les grilles documentaires correspondant à ces deux requêtes sont présentées Fig. 5.13 pour la chimiothérapie et Fig. 5.14 pour la radiothérapie. Ces deux grilles présentent en blanc les individus ayant subi un traitement préopératoire et en noir ceux qui n'en ont pas reçu. Le trait vertical matérialise la limite de 12 ganglions observés.

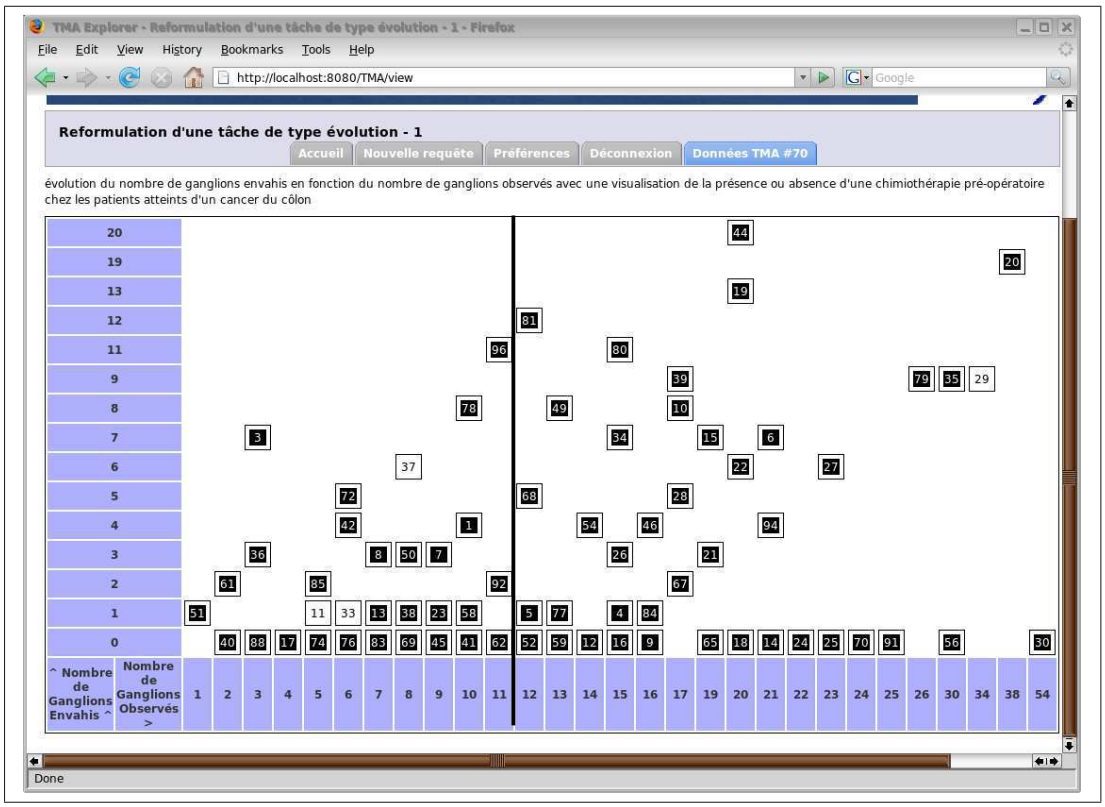


Fig. 5.13: Test de l'hypothèse chimiothérapie expliquant l'observation de moins de 12 ganglions - Le trait vertical sépare les individus pour lesquels moins de 12 ganglions ont été observés (à gauche). La majorité des individus ayant subi une chimiothérapie présentent moins de 12 ganglions observés.

Au total, 31 patients ont moins de 12 ganglions observés. Dans le cas de la chimiothérapie, 3 des 4 patients ayant subi ce traitement présentent moins de 12 ganglions observés. Ce type de traitement peut donc être éventuellement considéré comme intervenant parmi les facteurs explicatifs de l'observation de moins de 12 ganglions. Par contre, dans le cas de la radiothérapie, les patients ayant subi ce traitement sont répartis entre les deux ensembles (plus ou moins de 12 ganglions observés) et son rôle ne peut pas être évalué avec les données disponibles. Le traitement préopératoire peut donc intervenir en tant que facteur explicatif, mais ce n'est pas le seul. Les informations disponibles au sein des dossiers cliniques ne sont malheureusement pas suffisantes pour tester d'autres hypothèses, comme celle d'une pièce opératoire n'incluant pas assez de tissu périphérique pour y trouver 12 ganglions.

Par contre, cette reformulation permet de mettre en exergue une potentielle limite à l'expressivité de la requête. Le test conjoint des deux hypothèses de chimiothérapie et radiothérapie n'est pas possible, car la structure de la requête n'inclut pas de notion correspondant au OR booléen, qui serait applicable ici. Les deux éléments (présence d'une chimiothérapie préopératoire et présence d'une radiothérapie préopératoire) sont en effet de même nature (deux concepts booléens) et leurs valeurs possibles sont identiques (vrai ou faux), ce qui permettrait un éventuellement traite-

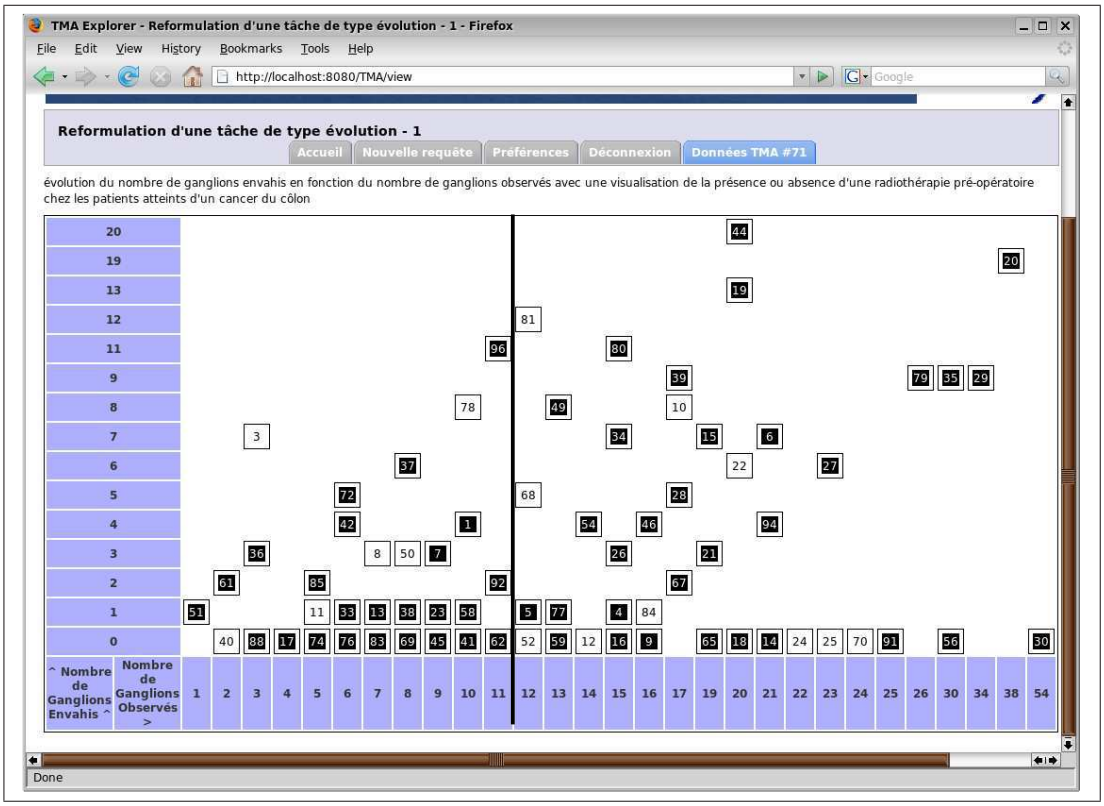


Fig. 5.14: Test de l'hypothèse radiothérapie expliquant l'observation de moins de 12 ganglions - Le trait vertical sépare les individus pour lesquels moins de 12 ganglions ont été observés (à gauche). Les individus ayant subi une chimiothérapie sont répartis équitablement entre les deux parties de la grille.

ment conjoint des deux concepts.

5.3.4.3.3 Vérification de cohérence de la composante N du stade

La seconde problématique découlant de l'exemple d'évolution présenté précédemment est la vérification de cohérence de la composante N du stade. En effet, ainsi que présenté au sein du Tab. 5.2, cet élément de la classification pTNM est directement liées au nombre de ganglions envahis. Ainsi,

- ⋆ Tous les individus ayant 0 ganglion envahi devraient avoir une valeur de N égale à 0,
- ⋆ Tous les individus ayant 1 à 3 ganglions envahis devraient avoir une valeur de N égale à 1,
- ⋆ Tous les individus ayant plus de 4 ganglions envahis devraient avoir une valeur de N égale à 3.

Afin de vérifier que c'est bien le cas, il suffit de modifier le but de la question biologique. Le nouveau problème devient l'«évolution du nombre de ganglions envahis

en fonction du nombre de ganglions observés avec une visualisation de la composante N du stade chez les patients atteints d'un cancer du côlon».

Au niveau pratique, il suffit de reformuler la requête d'exemple décrite Paragraphe 5.3.4.1 en changeant le but de la requête. Au lieu de la composante T du stade, celui-ci devient la composante N.

La grille correspondant du document de synthèse résultat est présentée Fig. 5.15.

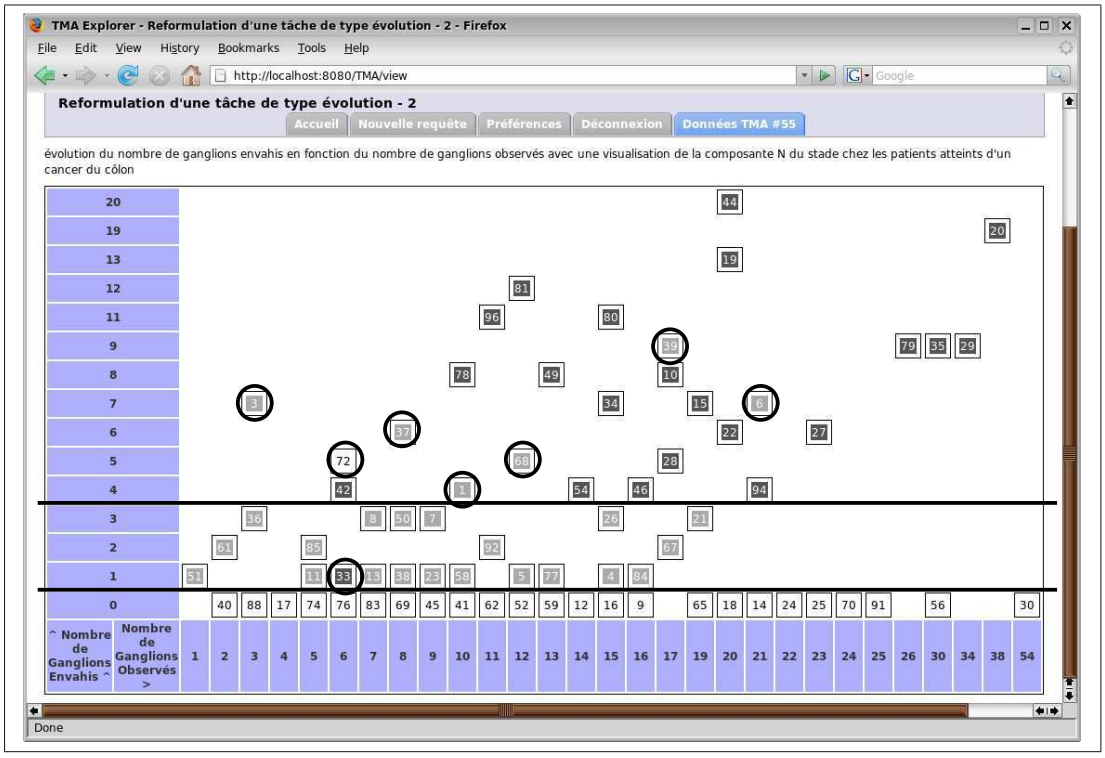


Fig. 5.15: Reformulation pour visualiser la composante N du stade - Le document de synthèse est similaire à celui de la Fig. 5.9. Par contre ici la couleur de fond des cases correspond à la composante N du stade, de 0 (blanc) à 2 (noir). La correspondance avec le nombre de ganglions envahis est matérialisée par les traits horizontaux. Les cases dont la couleur est non conforme sont entourées.

Cette étude fait apparaître huit cas où le choix de la valeur de composante N du stade n'est pas en accord avec le nombre de ganglions envahis. Pour la plupart, il s'agit sans doute d'erreurs. En particulier, on retrouve ici le patient 3, qui a 7 ganglions envahis sur 3 observés, avec une valeur de N de 1, ce qui correspondrait à 1 à 3 ganglions envahis. Ceci semble conforter l'idée de l'erreur de saisie, et en particulier une inversion de valeur entre nombre de ganglions observés et envahis. Pour d'autres, en particulier les patients 4 et 68, ils sont très proches de la limite entre deux catégories (4 et 5 ganglions envahis bien qu'ils soient indiqués de classe N1) et d'autres considérations ont peut-être influencé le choix d'une classe N1.

5.3.4.4 Informativité des résultats

5.3.4.4.1 Introduction

De même que pour l'étude d'un cas de comparaison, les paragraphes précédents ont permis une validation qualitative de l'adéquation de certaines tâches d'évolution avec le modèle de synthèse envisagé, l'utilité du document de synthèse pour inférer des connaissances biologiques et préparer une fouille de données, et comment un tel document peut suggérer de nouvelles études.

Ici aussi, il s'agit alors d'estimer quel est l'intérêt d'un tel système d'assistance à la synthèse dans l'absolu et l'apport des résultats proposés par rapport à d'autres outils. Comme pour l'exemple de comparaison, cette estimation est menée par confrontation du document de synthèse avec les résultats d'un autre outil qui permet de construire des représentations proches.

Pour évaluer l'informativité des études relevant d'une problématique d'évolution, il faut choisir un logiciel permettant de construire des nuages de points à partir d'un tableau de données. Un outil de statistiques ou un tableur permettent de réaliser de tels graphiques. Ici, le choix a porté sur un tableur, celui inclus dans la suite bureautique OpenOffice[3], car c'est le type d'outil le plus simple à utiliser pour un novice.

Dans la suite de ce paragraphe, va être présenté comment une visualisation similaire à celle proposée par le document de synthèse peut être construite avec le tableur. Une comparaison critique entre les deux constructions va alors être menée.

5.3.4.4.2 Construction d'une visualisation avec le tableur

La plupart des tableurs permettent la construction de graphiques à partir de tableaux de données. En particulier, une représentation des données sous forme de nuage de points est possible. La première étape de la construction de la visualisation est donc la préparation de ce tableau de données.

Pour ce faire, il s'agit dans un premier temps d'extraire de la base de données, et en particulier de la vue patient, les informations nécessaires (composante T du stade de la tumeur, nombre de ganglions observés, nombre de ganglions envahis) et de les charger dans le tableur. Ensuite quelques transformations sont nécessaires.

Tout d'abord, il faut éliminer les lignes présentant des données manquantes, qui sont parfois mal gérées par les tableurs. Ensuite, il faut organiser les données de

[3] http ://fr.openoffice.org/

façon à permettre une visualisation différenciée de la composante T du stade. Cette transformation est illustrée par la Fig. 5.16.

T	Ganglions observées	Ganglions envahis
4	10	4
3	10	4
3	3	7
3	15	1
3	12	1
4	21	7
3	9	3
3	7	3
3	16	0
3	17	8

Ganglions observés	Ganglions envahis			
	T1	T2	T3	T4
1			1	
1				1
2		0		
2				2
3	0			
3			7	
3			3	
3			3	
4		0		

Fig. 5.16: Transformation des données pour utiliser le tableur - Les données extraites de la base de données, à gauche, sont manuellement réorganisées, à droite, pour permettre la construction d'un nuage de points similaire au document de synthèse d'une tâche prototypique d'évolution.

L'utilisation des outils de construction de graphiques du tableur permet alors de construire un nuage de points tel que celui présenté Fig. 5.17.

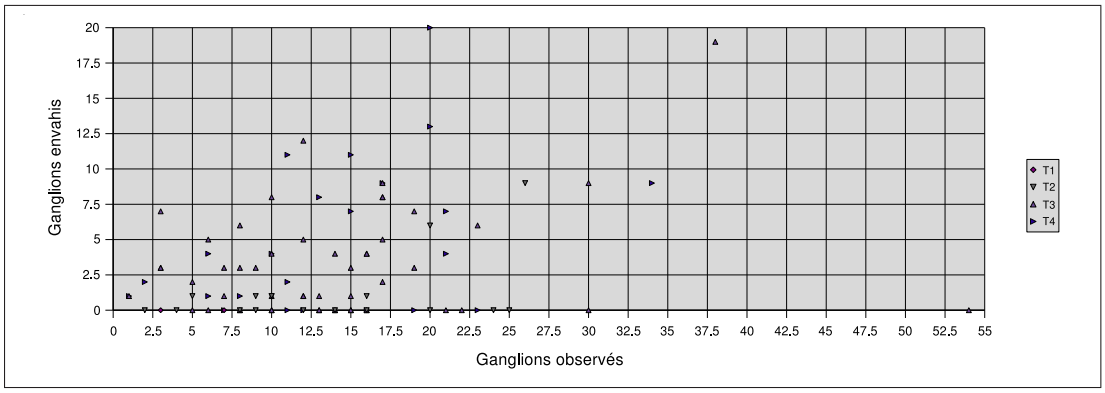


Fig. 5.17: Nuage de points construit avec le tableur - Les individus sont présentés au sein d'un nuage de points où l'abscisse correspond au nombre de ganglions observés et l'ordonnée au nombre de ganglions envahis. Les points ont une forme et une couleur qui dépendent du stade, selon la légende présentée à droite.

5.3.4.4.3 Comparaison critique

Les visualisations proposées par les fonctionnalités graphiques d'un tableur courant et par le prototype apparaissent comme similaires. Comme pour Treemap et les tâches prototypiques de comparaison, il est alors ici aussi légitime de se demander ce que le système d'assistance à la synthèse apporte de plus par rapport à un simple tableur tel que celui inclus dans la suite bureautique OpenOffice.

La différence est là aussi tout d'abord conceptuelle. Un tableur est un outil de gestion de données très simple, dont les fonctionnalités statistiques, qu'elles soient numériques ou graphiques, restent limitées à des analyses descriptives basiques. En

tant que tel, il ne peut permettre la construction de visualisations élaborées telles que celles de Treemap ou de l'exemple de comparaison. Le système d'assistance à la synthèse se veut un système de Recherche d'Information orientée tâche et là encore présente l'avantage de fédérer plusieurs approches pour appréhender les données, par les diverses tâches prototypiques proposées.

Ensuite, comme dans Treemap, le focus sur la collection de données est réalisé par la sélection et l'organisation manuelle réalisées lors de la construction du fichier de données. Un changement de focus induit là encore la construction d'un nouveau fichier spécifique. Ceci pose les mêmes limites que Treemap sur l'exploration du corpus documentaire et en particulier la réalisation de nouvelles études en conséquence de l'observation des résultats d'une étude précédente, qui est au contraire simple et rapide dans le cas de la synthèse.

De plus, la représentation graphique est construite en relation directe avec les données stockées dans la feuille du tableur. L'exploration du jeu de données est donc dirigée non pas par une requête explicite, mais par la sélection et l'organisation manuelle des données au sein de la feuille du tableur. En plus d'être implicite, la mise en place du cadre de l'étude implique une manipulation manuelle du jeu de données, qui peut rapidement devenir complexe et dans tous les cas peut induire des erreurs.

Ensuite, la navigation dans l'espace documentaire est inexistante dans un graphique construit au sein d'un tableur, ce qui limite fortement l'exploration du contexte de l'étude.

Enfin, un tel graphique conserve l'échelle des plages de valeurs représentées. La représentation est donc une image de la distribution des individus, étirée entre valeurs minimum et maximum. Dans le cas du document de synthèse, la grille est une visualisation compacte où les zones vides sont occultées. Si le graphique constitue une représentation fidèle, il présente l'éventuel inconvénient d'un tassement des individus dont les valeurs de variables sont faibles, alors que la grille du document de synthèse donne une place identique à tous, permettant un appréhension des données elles-mêmes plutôt que de l'espace où elles s'inscrivent.

Ces diverses remarques font apparaître la même différence fondamentale entre les deux approches : la ou les composantes demandant le plus gros effort réflexif de la part de l'utilisateur. L'effort cognitif est là aussi déporté de la conception du fichier, pour le tableur, vers la formulation de requête, pour le système de synthèse.

Comme pour les tâches de comparaison, il apparaît donc que la construction d'un document de synthèse pour une tâche de type évolution permet une exploration dirigée d'un jeu de données plus souple qu'un simple nuage de points construit au sein d'un tableur, grâce à l'entrée dans le système par une requête structurée, aux possibilités de navigation dans les données apportées par les vues associées à la grille

documentaire et à une visualisation compacte des données.

5.3.4.5 Un second cas enrichissant

Comme pour l'exemple de comparaison, l'étude de cas qui vient d'être présentée a permis de montrer de manière qualitative certaines qualités des documents de synthèse relevant de problématiques d'évolution.

Tout d'abord, le problème biologique d'exemple a pu être exprimé sous forme de requête structurée, point d'entrée dans le système de synthèse. Ensuite, le document de synthèse a permis de mettre en lumière tout à la fois une donnée incohérente et des faits biologiques connus, montrant ainsi qu'il s'agit d'un bon préalable à une fouille de données et que de nouvelles connaissances pouvaient être inférées. Ceci a permis là aussi de montrer l'utilité du document construit.

De plus, l'observation de la grille a permis de suggérer deux nouvelles études, confortant la suggestivité de la présentation : l'un permettant l'exploration d'une hypothèse explicative aux faits observés, l'autre permettant de révéler des données incohérentes qui seraient à écarter d'une analyse statistique par exemple. Enfin, une comparaison avec un nuage de points construit avec un simple tableur a permis tout à la fois une validation informelle et rapide de la grille documentaire construite et la révélation des spécificités et de l'intérêt du processus de synthèse par rapport à un logiciel proposant des outils simples d'analyse descriptive des données.

5.3.5 Un premier diagnostic qualitatif encourageant

Les deux études de cas qui ont été menées ici ont permis de réaliser un premier diagnostic du système, de manière qualitative.

En effet, dans les deux cas, il a été possible de formuler les problématiques biologiques considérées au sein de requêtes structurées, donnant quelques indices en ce qui concerne l'évaluation de l'adéquation à la tâche.

De plus, l'analyse des résultats proposés, au sein de documents de synthèse, a permis de tirer des conclusions biologiques ou de noter des incohérences, apportant des informations d'intérêt dans le cadre des problématiques étudiées. Les informations présentées ont aussi permis de préparer des fouilles de données, tout à la fois en proposant des sélections de données pertinentes et en mettant en exergue des données incohérentes, comme un individu ayant plus de ganglions envahis qu'observés dans l'exemple d'évolution. Il a ainsi été possible de replacer le jeu de données dans une démarche expérimentale classique, en réalisant des tests statistiques sur les don-

nées sélectionnées dans le cadre d'une étude de comparaison. Ces exemples ont donc permis de montrer l'utilité du système d'assistance à la synthèse.

Enfin, les fonctionnalités du système ont pu être évaluées par comparaison avec des visualisations construites avec d'autres outils. Une comparaison critique a permis de cerner l'intérêt du processus de synthèse par rapport à deux outils permettant la construction de visualisations équivalentes.

Mais ces éléments de diagnostic, s'ils sont intéressants par les informations qu'ils apportent sur les notions d'adéquation à la tâche, utilité, qualité des résultats produits ou intérêt des fonctionnalités proposées, ne sont pas suffisants. En effet, menés par moi-même et donc par un expert dans l'utilisation du système, ils ne sont pas indicatifs des capacités du système en utilisation réelle. Une évaluation centrée sur l'usage, menée par le biais de tests utilisateurs, est donc nécessaire. Celle-ci fait l'objet de la prochaine section.

5.4 Étude utilisateurs

5.4.1 Objectifs

Comme présenté précédemment, il s'agit de s'intéresser ici au point de vue des usagers potentiels. Afin d'accéder à un tel point de vue, il est impossible de mettre en place des métriques objectives et l'évaluation se base sur des réponses à questionnaire et commentaires d'utilisateurs ayant testé le système.

De manière générale, on vise à déterminer l'utilisabilité et l'efficacité du système. Par utilisabilité, on entend tout à la fois des problématiques d'ergonomie, de navigabilité, d'apprentissage de l'utilisation de l'outil. L'efficacité peut être envisagée selon deux axes : une efficacité «utilisateur», qui peut être assimilée à la notion de pertinence utilisateur des systèmes de Recherche d'Information, et donc aux notions d'adéquation à la tâche et pertinence interprétationnelle décrites dans le Paragraphe 3.4.4, et une efficacité «système», qui correspond à des mesures de performances.

Aussi les points à évaluer peuvent être résumés dans la liste suivante :

- ⋆ Utilisabilité de l'outil,
- ⋆ pertinence utilisateur :
  - ⋆ adéquation à la tâche,
  - ⋆ pertinence interprétationelle,
- ⋆ performances du système.

Après une présentation du cadre de l'étude, et en particulier du panel d'utilisateurs qui a été constitué ainsi que du plan d'expérience mis en place, chacune de ces problématiques sera décrite plus en détails et les résultats de l'expérimentation selon chacun de ces axes seront présentés.

5.4.2 Cadre de l'étude

5.4.2.1 Panel d'utilisateurs

Étant donné que le temps disponible pour réaliser les études utilisateurs a été très limité, il s'est avéré impossible de recruter des usagers potentiels hors de l'équipe du projet. La taille du panel d'utilisateurs est donc réduite, ce qui limite les possibilités d'analyse statistique des résultats.

Par contre, le groupe d'usagers qui a testé le système présente une grande variété de profils, ce qui permet d'avoir un aperçu des relations potentielles entre les évaluations et les divers niveaux d'expertise, tant en ce qui concerne les problématiques biologiques liées à la recherche en oncologie en utilisant la technologie des TMA qu'en ce qui se rapporte à l'utilisation de l'outil informatique.

En effet, ce panel inclut tout d'abord un doctorant en informatique, relativement novice en ce qui concerne les problématiques biologiques. Ensuite, il inclut deux étudiants dotés d'un Master 2 en biologie qui suivent une formation complémentaire d'un an en informatique (conception, développement et autres). Puis ont été intégrés deux membres du personnel de laboratoire qui réalisent les blocs et lames TMA : un technicien et un ingénieur.

Enfin, trois personnes sont considérées comme des experts du domaine : un doctorant qui travaille sur les problématiques d'imagerie associées à la technologie des TMA, un maître de conférence en biologie, chef de projet pour TMA-Explorer, dont le thème de recherche est centré sur l'oncologie, et un anatomopathologiste qui termine une thèse sur l'étude des transformations tumorales dans le cancer du côlon en utilisant les TMA.

Le Tab. 5.6 résume les caractéristiques des huit utilisateurs composant le panel.

Ces utilisateurs potentiels se sont prêtés à une séance d'expérimentation avec le système qui est décrite plus précisément dans le prochain paragraphe.

Tab. 5.6: Présentation du panel d'utilisateurs - Ce panel compte un novice vis à vis du domaine d'étude, deux étudiants de Master 2, deux membres du personnel technique du laboratoire et trois experts, doctorants ou chercheur.

Utilisateur	*Compétences informatique*	*Compétences biologiques*
Novices en biologie :		
Novice	Doctorant	Niveau lycée
Étudiants de Master 2 en biologie :		
Étudiant 1	Formation complémentaire d'un an	Niveau Master 2
Étudiant 2	Formation complémentaire d'un an	Niveau Master 2
Personnel de laboratoire :		
Technicien	Novice	Niveau Licence
Ingénieur	Utilisateur standard	Niveau Master 2
Experts du domaine :		
Doctorant imagerie	Doctorant	Niveau Doctorat
Maître de conférence	Utilisateur standard	Chercheur en oncologie
Anatomopathologiste	Utilisateur standard	Expert en cours de doctorat

5.4.2.2 Plan d'expérience

Afin d'avoir des informations comparables entre les sessions de tests des différents utilisateurs, les tests ont été menés selon un plan expérimental précis, détaillé ici.

Tout d'abord, il faut s'assurer d'une homogénéité de déroulement de la séance entre les différents utilisateurs. Pour se faire, un scénario de test a été défini. Ce scénario décrit succinctement les diverses manipulations qui seront testées.

Dans un premier temps, il s'agit de se connecter au système. Ensuite, des problématiques biologiques relevant des deux types de tâches prototypiques disponibles, définies avec les biologistes, sont proposées. Ces problématiques sont celles qui ont fait l'objet de l'étude de cas de la section précédente. A charge de l'usager de décrire ces problématiques par l'intermédiaire de l'interface de saisie de requête et d'interpréter le document de synthèse résultant. Ensuite, l'opportunité est laissée de réaliser quelques études librement définies par l'utilisateur. Enfin, un changement de focus de l'exemple de comparaison, conduisant à une reformulation de requête, est proposé.

Afin de limiter les interactions avec l'expérimentateur, ce scénario est proposé aux utilisateurs sous une forme rédigée imprimée qu'ils ont l'opportunité de lire avant le début de la session de test. Le document correspondant est présenté en Annexe J.

Ensuite, comme il a déjà été indiqué, le point de vue de l'utilisateur sur le système n'est pas accessible par des mesures objectives et est en général obtenu par

l'intermédiaire de questionnaires. Dans le cadre de l'expérience, ce questionnaire est décomposé en trois sections : généralités, couvrant les problématiques d'utilisabilité de l'outil, formulation de requête, consacrée aux problématiques d'adéquation à la tâche, et résultat de la requête, focalisée sur l'estimation de la pertinence interprétationnelle.

Chaque section du questionnaire regroupe un ensemble d'affirmations pour lesquelles l'utilisateur donne un niveau d'approbation, de 1 (pas d'accord) à 5 (tout à fait d'accord). En fin de document, l'opportunité est laissée à l'usager de compléter son appréciation par des commentaires sur des thématiques qui auraient été oubliées au sein du questionnaire. Une image du document est présentée en Annexe K.

Puis doit être ensuite défini le contexte expérimental de déroulement des séances de test. En général, il est admis parmi les bonnes pratiques de réaliser ces tests en l'absence de l'expérimentateur, dans une salle spécialement équipée pour permettre l'observation sans déranger l'utilisateur, dotée d'une caméra, etc. Des systèmes d'enregistrement des traces des actions de l'utilisateur permettent d'accéder aux nombres de clics, lieux des clics, temps passé de manière automatique.

Malheureusement, il n'a pas été possible de réaliser une expérimentation dans ces conditions idéales, du fait des contraintes horaires des différents participants. Les observations ont donc été réalisées par un observateur présent dans la même pièce, qui s'est efforcé de réaliser :

- un chronométrage du temps passé sur chaque écran,
- une prise de note des remarques faites à voix haute par l'usager au cours la séance de test,
- une observation du comportement de l'usager au cours de son interaction avec le système (erreurs, retours en arrière, etc.), permettant une estimation qualitative et non quantitative des difficultés rencontrées.

Ces différents éléments permettent de compléter l'évaluation réalisée par le biais du questionnaire avec des informations qualitatives supplémentaires. De plus, ils permettent une estimation de la notion d'efficacité.

Enfin, ces séances de tests ont servi de cadre à l'évaluation des performances du système, conjointement à des tests réalisés par mes soins. Le temps de traitement, entre le moment où l'exécution de la requête a été demandée et la mise à disposition des résultats a été chronométré pour l'ensemble de études libres réalisées par les usagers, donnant ainsi un aperçu plus étendu des temps de traitement.

Les résultats des différentes évaluations menées dans le cadre de cette étude utilisateurs vont être présentés dans les prochains paragraphes.

5.4.3 Utilisabilité de l'interface

5.4.3.1 Introduction

La première dimension de l'évaluation de l'usage du système menée d'un point de vue utilisateur est celle de l'utilisabilité du système. Cette notion regroupe des considérations d'ergonomie, de facilité d'utilisation, de facilité d'apprentissage, de satisfaction générale. Bien qu'il ne soit pas central dans le cadre d'un prototype, où les considérations d'ergonomie en particulier ne sont pas nécessairement au cœur de la conception du système, cet axe qualité mérite qu'on s'y attache même brièvement. En effet, des considérations d'utilisabilité peuvent influencer les jugements des usagers ou leur causer des difficultés dans d'autres dimensions plus cruciales de l'évaluation.

Cette utilisabilité est estimée selon deux axes.

Dans un premier temps, celle-ci passe par l'analyse des résultats du questionnaire pour la partie Généralités, ainsi que des commentaires des usagers se rapportant à des notions d'ergonomie ou d'interaction avec le système. Ceci permet d'accéder à une évaluation tout à la fois quantitative, par les appréciations posées sur les affirmations du questionnaire, et qualitative, par les remarques écrites ou formulées oralement.

Dans un second temps, elle se base sur les chronométrages des différentes formulations de requête et interprétations des résultats, afin de donner une estimation plus quantifiée de la facilité d'utilisation et surtout d'un effet apprentissage de l'utilisation du système.

Les prochains paragraphes vont être consacrés à ces deux dimensions de l'évaluation.

5.4.3.2 Analyse des résultats du questionnaire

La section Généralités du questionnaire regroupe un ensemble de six affirmations organisées selon diverses dimensions de la notion d'utilisabilité : facilité d'utilisation, convivialité, facilité d'apprentissage, navigabilité, satisfaction générale. Le Tab. 5.7 recense ces dimensions et les affirmations correspondantes, associées à un code d'identification. Seule la facilité d'utilisation est décomposée en plusieurs dimensions : l'appréciation de la facilité d'utilisation en tant que telle et l'aisance resentie par l'utilisateur en cours d'interaction avec le système. Les autres composantes s'expliquent d'elles-mêmes par les affirmations associées.

Tab. 5.7: Dimensions de l'évaluation de l'utilisabilité du système - Ce tableau présente, pour chaque dimension de l'utilisabilité du système, les affirmations de la partie Généralités du questionnaire correspondantes et le code qui permettra de les identifier dans les schémas suivants.

Dimension de l'évaluation	*Affirmations*	*Code*
Facilité	- De façon générale, je suis satisfait de la facilité d'utilisation du système.	F1
	- Je me suis senti à l'aise dans l'utilisation du système.	F2
Convivialité	- L'interface du système est plaisante.	C
Apprentissage	- De façon générale, il a été facile d'apprendre à utiliser le système.	A
Navigabilité	- La navigation au sein des diverses pages est intuitive.	N
Satisfaction	- De façon générale, je suis satisfait de l'utilisation du système.	S

Les résultats de ce questionnaire, correspondant à une valeur entre 1 (pas d'accord) et 5 (tout à fait d'accord) pour chaque affirmation pour chaque utilisateur du panel sont présentés en Annexe L. Ces données quantitatives ont servi de base à la construction des boîtes à moustache, qui permettent une analyse descriptive des données, présentées Fig. 5.18. Les boîtes à moustaches proposent en effet, pour chaque affirmation, la valeur moyenne, les valeurs minimum et maximum et les limites des premier et troisième quartiles.

De manière générale, malgré la faible taille du panel les usagers semblent montrer un intérêt pour le prototype qu'ils ont testé, puisque la moyenne est en général proche de 4. Seuls le maître de conférence et l'utilisateur novice ont exprimé quelques réserves.

Le novice a expliqué son malaise par son manque de recul vis à vis du domaine applicatif qu'il connaît peu, induisant des difficultés d'appréhension des concepts et de leur manipulation. Son expertise informatique lui a aussi permis de noter des défauts d'ergonomie (placement de certains boutons, utilisation d'une même icône pour plusieurs sémantiques différentes, etc.) qui sont passés inaperçus des autres usagers plus novices en la matière.

Le maître de conférence a noté un manque de documentation et aide contextuelle, qui seraient essentiels pour un produit fini et surtout une utilisation en autonomie : ma présence dans la salle a permis aux utilisateurs de poser des questions et de se faire expliquer le fonctionnement de l'outil quand les manipulations à réaliser étaient peu claires ou inhabituelles. Mais de tels manques d'informations à destination de l'usager sont à attendre de la part d'un prototype.

Ainsi, pour la plupart des utilisateurs, «Il suffit de prendre l'habitude de comment ça marche et après c'est facile à utiliser», comme l'a dit l'un deux. Cette notion de temps d'adaptation à l'utilisation d'un nouvel outil, facteur important pour l'adop-

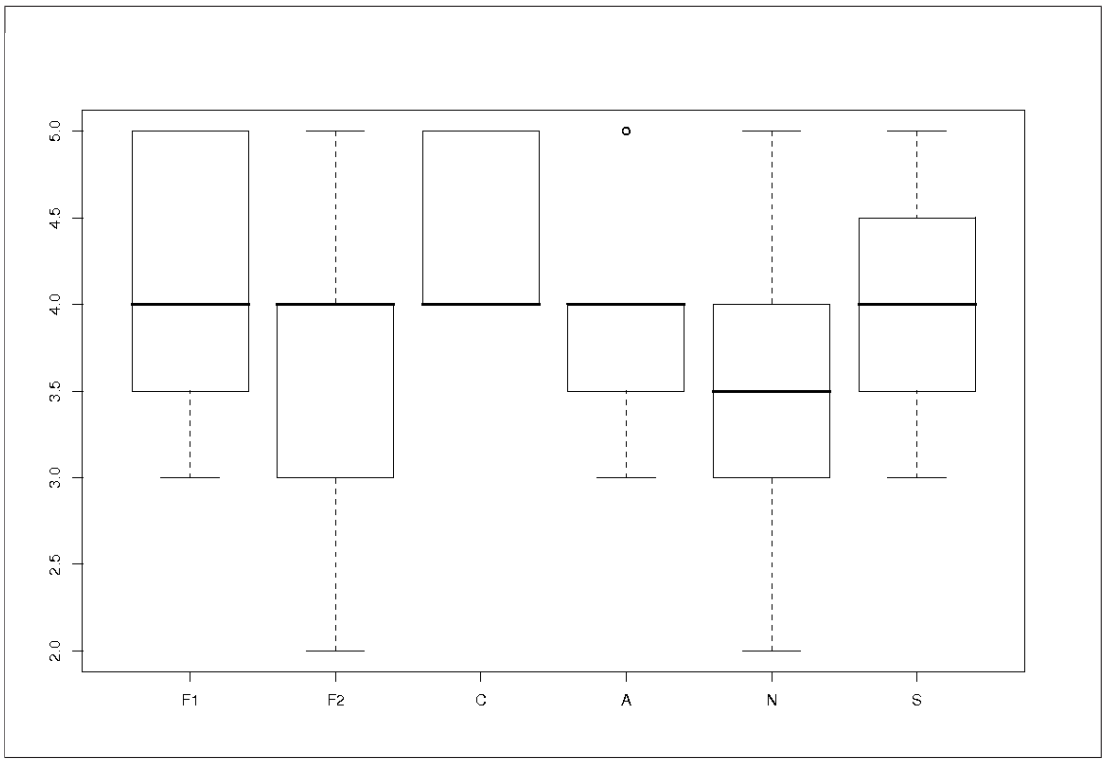


Fig. 5.18: Boîtes à moustaches pour les questions d'utilisabilité - Ces boîtes correspondent aux résultats de la partie Généralités du questionnaire. Les diverses questions sont représentées par leur code tel que défini dans le Tab. 5.7.

tion d'une nouvelle technologie, est une problématique courante dans l'évaluation logicielle, et son analyse va faire l'objet du prochain paragraphe.

5.4.3.3 Effet de l'apprentissage

La notion de facilité d'apprentissage, évaluée par l'une des affirmations du questionnaire, peut aussi être analysée de manière plus objective par des chronométrages du temps passé à formuler les problèmes biologiques sous forme de requête structurée et à interpréter les résultats présentés dans le document de synthèse.

Lors d'une séance de test, chaque utilisateur a en effet mené en général 4 études :

- ⋆ un exemple de comparaison,
- ⋆ un exemple d'évolution,
- ⋆ une étude libre, en général une comparaison,
- ⋆ une reformulation de l'exemple de comparaison.

Pour chacune, il a tout à la fois formulé la requête et interprété les résultats.

Il est alors légitime de se demander si cette impression de facilité croissante dans

l'utilisation du système, liée à un effet d'apprentissage, peut être observée objectivement dans les temps passés à mener ces activités au cours des 4 études. Les Fig. 5.19 et Fig. 5.20 présentent de manière graphique, pour chaque utilisateur, le temps passé à formuler une requête (Fig. 5.19) et à interpréter le document de synthèse (Fig. 5.20), au cours du temps, la notion de temps étant représentée par les études successives.

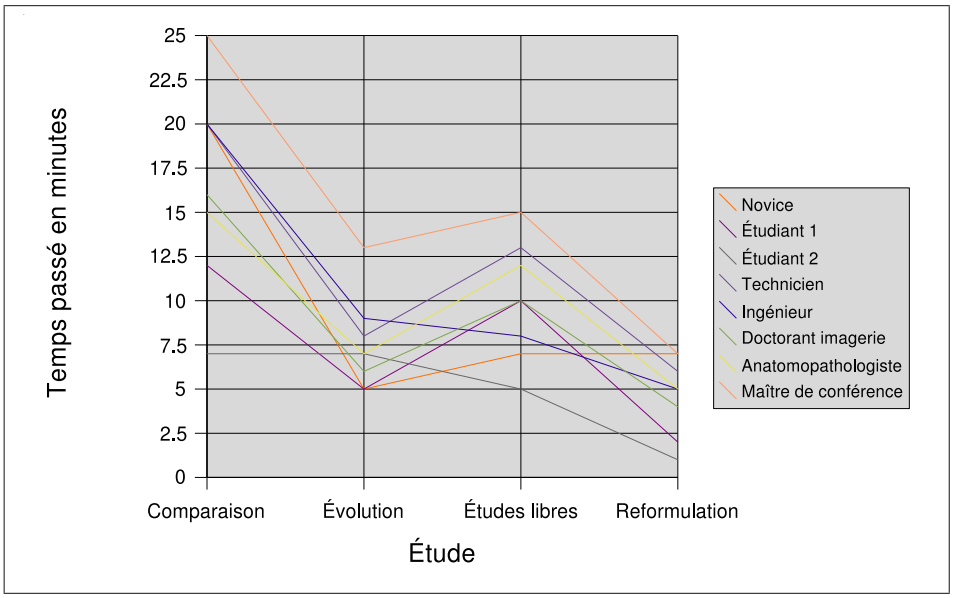


Fig. 5.19: Temps passé pour la formulation - Ce graphique présente le temps passé pour la formulation de chaque requête par chacun des 8 utilisateurs du panel.

Pour la formulation, de manière générale, on peut noter une diminution importante du temps passé entre la comparaison et l'évolution, une augmentation pour l'étude libre, et encore une baisse pour la reformulation. Le pic pour l'étude libre peut être expliqué par la prise en compte, dans le temps chronométré, de la définition du problème biologique à étudier, problème qui est déjà défini pour les autres exemples. Il semble donc y avoir un effet non négligeable de l'apprentissage sur la facilité d'utilisation de l'outil pour la formulation de requête.

Par contre, l'effet de l'expertise informatique et l'expertise du domaine sont difficiles à interpréter. Il semblerait tout de même que les utilisateurs bénéficiant d'une double compétence en biologie et en informatique sont «avantagés» : ils passent moins de temps à la formulation dès les premiers essais, alors que les autres peinent plus au début, soit à cause de leur manque de compréhension des problématiques biologiques, soit à cause de difficultés de prise en main de l'interface liées à un manque d'habitude de l'outil informatique.

Pour l'interprétation, de manière générale, on peut aussi noter une diminution importante du temps passé entre la comparaison et l'évolution, puis ce temps reste plus ou moins constant. Une exception notable est l'utilisateur novice. Son temps d'interprétation reste élevé, quelle que soit l'étude, sauf pour l'étude libre. Ceci

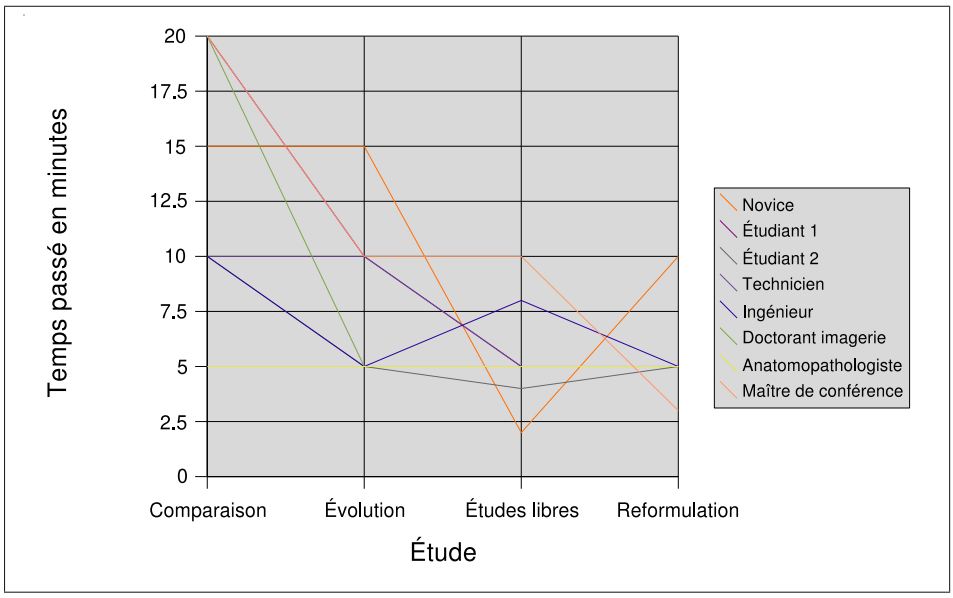


Fig. 5.20: Temps passé pour l'interprétation - Ce graphique présente le temps passé pour l'interprétation de chaque document de synthèse par chacun des 8 utilisateurs du panel.

est intrinsèquement lié à son manque de connaissance du domaine applicatif : il a eu besoin de beaucoup de temps pour comprendre les affichages et en tirer des conclusions dans les exemples où le problème biologique était défini. Par contre, son étude libre était très simple, conduisant à une grille où peu de commentaires étaient possibles et l'effort d'interprétation faible. Par contre, l'ingénieur a construit un problème biologique complexe pour son étude libre, conduisant à une interprétation élaborée, prenant plus de temps.

On peut noter aussi ici un effet de l'expertise sur le domaine d'étude : les étudiants et le novice connaissent peu les problématiques d'oncologie et leur interprétation prend plus de temps au début. Les autres utilisateurs baignent au quotidien dans le contexte des TMA et sont à l'aise avec l'interprétation dès le début.

5.4.3.4 Une utilisabilité satisfaisante

L'étude de cette première composante de l'évaluation utilisateurs, consacrée à la notion d'utilisabilité, a permis de mettre en exergue un certain nombre de points.

Tout d'abord, les utilisateurs semblent appréciateurs de l'outil en ce qui concerne sont utilisabilité. Le système est considéré comme facile d'utilisation, l'interface est jugée plaisante, la navigation intuitive et les usagers se sont exprimés comme satisfaits.

Ensuite, les utilisateurs ont en général exprimé une facilité d'apprentissage de

l'outil, notant qu'au cours de la séance de test, soit en général moins d'une heure, ils avaient appris à s'en servir. Cette facilité d'apprentissage a été évaluée plus objectivement, en analysant le temps passé à la formulation de requête et à l'interprétation des résultats au cours des diverses études menées par chaque usager. Tout en confirmant cet effet de l'apprentissage, cette analyse a aussi permis de montrer un effet de l'expertise tout à la fois informatique et du domaine applicatif, qui est très spécialisé.

Il est donc légitime de considérer que l'utilisabilité du système est satisfaisante. Il s'agit alors de s'interroger sur les fonctionnalités de l'outil, et en particulier sur le jugement porté par l'usager sur les deux composantes principales où il intervient : la formulation de la requête, évaluée par la mesure d'une adéquation à la tâche, et l'analyse des résultats, estimée par une pertinence interprétationnelle. Ces deux composantes de l'évaluation font l'objet des prochains paragraphes.

5.4.4 Adéquation à la tâche

5.4.4.1 Introduction

La notion d'adéquation à la tâche recouvre les problématiques liées à la formulation de requête, à l'expression d'une problématique biologique sous la forme d'une requête structurée relevant d'une tâche de synthèse particulière. L'analyse de cette notion a été menée en détails parmi les bases conceptuelles du système d'assistance à la synthèse, au sein du Paragraphe 3.4.4.1, puisqu'elle fait partie des éléments du modèle de synthèse.

L'objectif ici est de mener une évaluation tout à la fois qualitative et quantitative de cette adéquation à la tâche, selon les dimensions identifiées au Paragraphe 3.4.4.1. En pratique, les affirmations correspondant à ces diverses dimensions font l'objet de la partie Formulation de Requête du questionnaire d'évaluation. Les utilisateurs ont aussi exprimé un certain nombre de commentaires, écrits ou oraux, se rapportant à cette composante de l'évaluation. L'analyse de ces résultats va donc être menée ici en deux temps.

Dans un premier temps, après un rappel des éléments évalués, une analyse quantitative des résultats des questionnaires va être menée. Dans un second temps, une analyse qualitative des remarques des utilisateurs va être présentée.

Tab. 5.8: Dimensions de l'évaluation de l'adéquation à la tâche - Ce tableau présente, pour chaque dimension de l'adéquation à la tâche, les affirmations de la partie Formulation de requête du questionnaire correspondantes et le code qui permettra de les identifier dans les schémas suivants.

Dimension de l'évaluation	*Exemples d'affirmations*	*Code*
Adéquation	- Les tâches proposées, telles que je les comprends, correspondent à de vrais problèmes d'appréhension de données.	A1
	- Les tâches proposées (comparaison, évolution et distribution) sont adaptées aux problématiques courantes du domaine.	A2
	- L'organisation taxonomique du vocabulaire est adéquate.	A3
Complétude	- Tous les types de tâches que je voudrais réaliser sont supportées.	C1
	- Le vocabulaire proposé est complet par rapport au domaine d'étude.	C2
Expressivité	- La transposition d'un problème que je voudrais explorer en une requête ne pose pas de problème.	Exp1
	- La syntaxe de la requête permet de décrire les tâches avec précision.	Exp2
	- La syntaxe de la requête permet d'exprimer toutes les tâches de comparaison ou d'évolution que je voudrais réaliser.	Exp3
Extensibilité	- Il est facile d'explorer les données en modifiant la requête.	Ext1
Navigabilité	- L'organisation des éléments de l'interface de saisie est claire.	N1
	- La saisie de la requête est facile à réaliser.	N2

5.4.4.2 Analyse quantitative

La section Formulation de requête regroupe 11 affirmations organisées en adéquation, complétude, expressivité, extensibilité et navigabilité. Ces diverses dimensions de l'adéquation à la tâche ont été identifiées au Paragraphe 3.4.4.1. Elles sont rappelées au sein du Tab. 5.8, conjointement aux affirmations correspondantes et aux codes qui ont été associés à chacune.

Les résultats de cette partie de questionnaire, correspondant là aussi à une valeur entre 1 (pas d'accord) et 5 (tout à fait d'accord) pour chaque affirmation pour chaque utilisateur du panel sont présentés en Annexe L. Ces données quantitatives ont servi de base à la construction des boîtes à moustaches, présentées Fig. 5.21. Les boîtes à moustaches présentent en effet, pour chaque affirmation, la valeur moyenne, les valeurs minimum et maximum et les limites des premier et troisième quartiles.

Ces résultats, qui restent à prendre avec précaution vu le petit nombre d'individus impliqués, semble indiquer un jugement plutôt positif des usagers en ce qui concerne l'adéquation à la tâche. Toutes les affirmations présentent une appréciation supérieure ou égale à la moyenne.

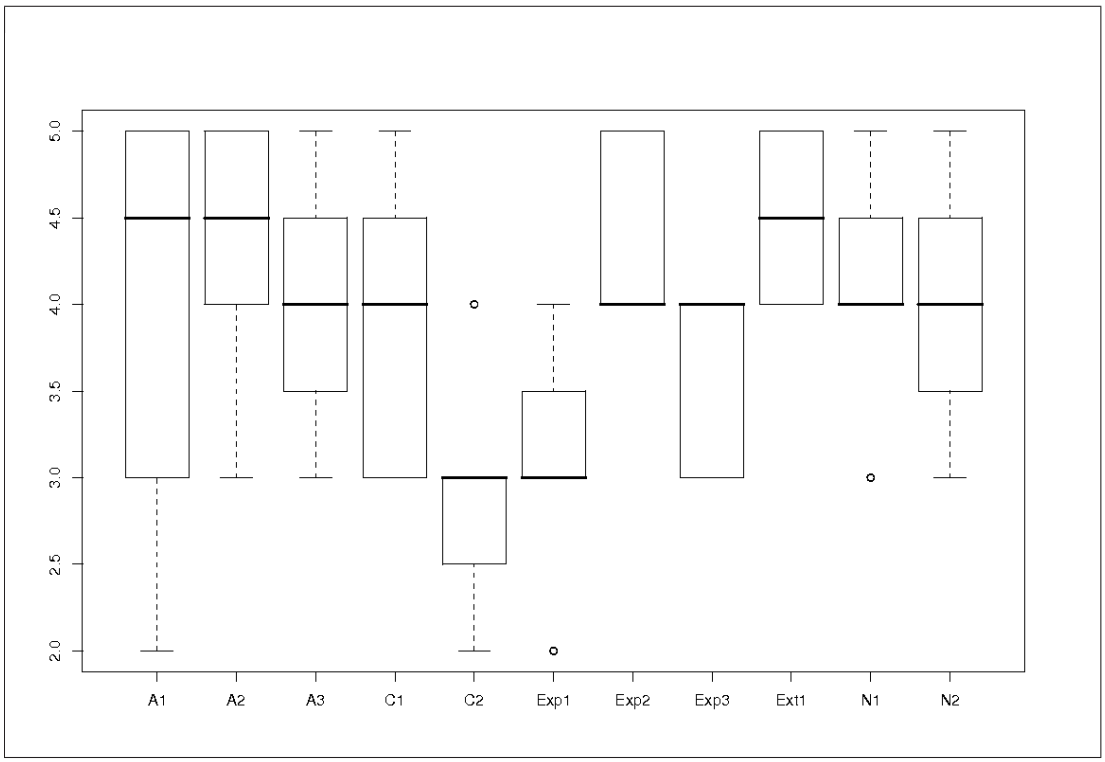


Fig. 5.21: Boîtes à moustaches pour les questions d'adéquation à la tâche - Ces boîtes correspondent aux résultats de la partie Formulation de requête du questionnaire. Les diverses questions sont représentées par leur code tel que défini dans le Tab. 5.8.

De manière plus détaillée, en ce qui concerne l'adéquation des tâches prototypiques en tant que telles, les tâches proposées sont estimées comme en adéquation avec les problématiques courantes du domaine et l'organisation de la taxonomie est aussi estimée comme correcte.

Les résultats concernant la complétude sont plus mitigés, en particulier en ce qui concerne les types de tâches proposés. Ceci est en adéquation avec l'état de développement du prototype : seules deux tâches prototypiques sont proposées, ce qui est peu par rapport à l'ensemble qui a été évoqué au Paragraphe 3.4.2.1, et les utilisateurs ont bien senti ce besoin de types de tâches complémentaires, permettant d'exprimer d'autres problématiques intéressantes pour eux.

En ce qui concerne l'expressivité, la syntaxe de la requête, conduisant à une requête structurée, est appréciée pour la précision de la description du problème biologique étudié. Par contre, les utilisateurs ne sont pas unanimes quant à la possibilité de formuler des requêtes pour tous les problèmes de comparaison et d'évolution, point qui sera analysé plus en détails au cours de l'analyse qualitative. De plus, la transposition d'un problème biologique sous forme de requête est jugée comme plutôt difficile ou difficile à évaluer par beaucoup d'utilisateurs, point qui sera aussi évoqué dans l'analyse qualitative.

En ce qui concerne l'extensibilité, les utilisateurs sont unanimes quant à la facilité d'exploration des données offerte par les possibilités de reformulation. Enfin, au niveau navigabilité, les usagers sont là aussi plutôt satisfaits.

Ces résultats quantitatifs sont associés à des commentaires, écrits ou oraux, qui apportent un éclairage complémentaire sur la perception des problématiques de formulation par les usagers. Leur analyse fait l'objet du prochain paragraphe.

5.4.4.3 Analyse qualitative

5.4.4.3.1 Introduction

Les évaluations quantitatives menées à partir des réponses au questionnaire sont associées à des commentaires, écrits en bas du questionnaire ou oraux, dont certains sont focalisés sur les problématiques de formulation de requête. En tant que tels, ils relèvent de la problématique d'adéquation à la tâche et servent de base à l'évaluation qualitative menée ici.

Les diverses remarques des usagers ont pu être classées en trois catégories, qui seront analysées par la suite : difficultés d'évaluation de la formulation, difficultés dans l'expression d'une requête et limites identifiées à l'expressivité de la requête.

5.4.4.3.2 Difficultés d'évaluation de l'expressivité de la requête

Les usagers ont unanimement exprimé des difficultés à compléter le questionnaire de manière satisfaisante pour eux, en particulier pour les notions d'adéquation, complétude et expressivité. D'ailleurs, un certain nombre d'entre eux n'ont pas répondu à certaines questions concernant ces problématiques, ou ont indiqué avoir choisi une valeur intermédiaire, à cause justement de ces difficultés d'évaluation.

Les usagers sont en effet d'avis que la simple séance très dirigée d'interaction avec le système, qui a servi de base à l'évaluation, est trop limitée pour leur permettre de se faire une idée claire sur certains points. On peut citer entre autres la complétude et l'adéquation de l'organisation taxonomique du vocabulaire, les éventuels besoins en autres types de tâches prototypiques pour étudier d'autres types de problèmes ou les limites à l'expressivité de la requête pour les tâches de comparaison et d'évolution.

Pour les utilisateurs, il faudrait utiliser le prototype pendant plus longtemps, chercher à étudier un plus grand nombre de problèmes biologiques définis par leurs soins, pour être en mesure de donner une évaluation correcte de ces éléments.

De plus, cette difficulté d'évaluation est accrue pour les utilisateurs qui manquent d'expertise sur le domaine applicatif. Ainsi, le novice a peu fourni d'appréciations pour ce type d'affirmation, de son propre aveu par manque de recul par rapport aux problématiques d'oncologie et d'analyse de l'expression de molécules au sein d'échantillons de tissus.

5.4.4.3.3 Difficultés à exprimer une requête

Les utilisateurs ayant participé à la session de tests ont tous indiqué qu'il n'était pas forcément évident d'exprimer les problèmes biologiques d'exemple sous forme de requête. Ces difficultés sont pointées de façon marginale au niveau de l'interface de saisie de requête. L'interface est en effet jugée facile d'utilisation une fois que le principe a été compris. En particulier on peut citer l'enchaînement de listes déroulantes pour parcourir la taxonomie du domaine d'étude et la nécessité d'ajouter l'élément sélectionné pour les composantes de la requête acceptant les sélections multiples. Surtout, c'est la mise en correspondance des différentes parties du problème biologique avec les différentes parties de la requête qui est jugée comme demandant un effort intellectuel. Ici, c'est donc la décomposition de la question biologique en rôles de la requête qui est pointée du doigt.

Or, dans le contexte de cette évaluation, la formulation en langage naturel des problèmes biologiques d'exemple a été orientée en fonction du système d'assistance à la synthèse. Les études à réaliser ont été présentées de façon à indiquer de manière implicite les rôles correspondant à chaque concept évoqué, en utilisant le vocabulaire défini dans la taxonomie du domaine. Ceci constitue donc un biais, qui en théorie facilite l'expression de la tâche à réaliser sous forme d'une requête structurée.

Il est alors légitime de se demander quelles seraient les difficultés de formulation pour des problèmes biologiques imaginés par un utilisateur qui connaît peu le système et la taxonomie du domaine. Malheureusement, en cette matière, les études libres réalisées par les utilisateurs ont été proposées dans le contexte de la session de test, en général sur un modèle proche de ceux rencontrés avec les exemples étudiés précédemment, et le fait que les utilisateurs aient réussi à les exprimer n'est pas forcément significatif.

Ce type de constats permet de mettre le doigt sur le problème commun à tous les systèmes de Recherche d'Information : l'effort intellectuel nécessaire à la formulation d'une requête et le fossé cognitif et sémantique entre besoin perçu et besoin exprimé dans la requête. Il serait alors intéressant d'essayer d'évaluer si l'effort est supérieur pour construire une requête structurée, par comparaison avec une requête standard comme pour un système de Recherche d'Information classique. Ceci permettrait d'évaluer si le coût intellectuel de formulation n'est pas trop élevé par rapport aux bénéfices de la requête structurée. Ce type d'étude n'a malheureusement pas pu être

mené, faute de temps.

5.4.4.3.4 Limites identifiées à l'expressivité

Les études conduites par les utilisateurs ont, en plus de permettre d'identifier une potentielle difficulté d'expression de la requête, ont aussi permis de révéler quelques limites à l'expressivité de la requête.

Tout d'abord, les différents éléments sélectionnés pour spécialiser chaque rôle de la requête sont traités de manière indépendante, sauf s'ils concernent le même concept. Par exemple sélectionner une date de naissance inférieure à 1950 et une date de naissance supérieure à 1920 conduit à un traitement prenant en compte une date de naissance inférieure à 1950 ET supérieure à 1920, de manière automatique. Par contre, dans tous les autres cas, les éléments sont considérés comme indépendants. Or cette hypothèse d'indépendance est fausse, ce qui peut poser un certain nombre de problèmes.

D'une part, ceci laisse la possibilité à l'utilisateur de faire des sélections incohérentes : par exemple, il est possible de limiter la liste des patients à ceux atteints d'un cancer du sein, puis de demander une comparaison dont le critère de groupement est la partie du côlon dans laquelle est localisée la tumeur, ce qui est complètement dénué de sens. Pour empêcher ce type de sélections, il faudrait mettre en place un système de dépendances entre éléments de la taxonomie, conduisant à une neutralisation de certaines branches de la taxonomie en fonction des choix réalisés.

D'autre part, les éléments composés ne peuvent pas être représentés. Cette notion d'éléments composés est intimement liée à la notion de dépendance entre éléments. Ainsi, dans l'exemple précédent, la partie de l'organe est liée à l'organe considéré. Un élément composé correspondrait alors par exemple à la partie côlon droit de l'organe côlon. De même, une mesure est liée à un marqueur. Un élément composé pourrait alors être le pourcentage de cellules marquées en marqueur Ki67. Il est pour l'instant impossible de décrire ce type d'élément par le biais de la requête structurée et leur éventuelle prise en compte dans le processus de synthèse doit faire l'objet d'une analyse plus poussée.

Ensuite, dans le cadre de l'étude de la suggestivité des résultats de l'exemple d'évolution, au Paragraphe 5.3.4.3, j'ai déjà montré une autre limite à l'expressivité de la requête, liée à l'absence de prise en compte d'une notion correspondant au OR des modèles de Recherche d'Information booléens. Dans cet exemple, une impossibilité de tenir compte à la fois de la présence d'une radiothérapie et d'une chimiothérapie préopératoire a été identifiée, montrant le problème au niveau du but de la requête. Mais cette même impossibilité est aussi présente au niveau des critères de sélection. De manière implicite, ces critères sont traités comme s'ils étaient sé-

parés par des AND, sauf quand ils portent sur le même concept. Auquel cas, un OR est utilisé. Mais de manière générale, il est impossible de définir des alternatives, ce qui évidemment limite l'expressivité de la requête.

Enfin, tous les rôles de la partie Besoins de la requête, qui correspondent à la description du problème biologique en tant que tel, sont obligatoires. Or, il est apparu au cours des tests que les usagers n'éprouvent pas forcément le besoin de tous les renseigner. En particulier, les critères de sélection et de tri pourraient être optionnels, permettant ainsi la construction d'un document portant sur tous les individus présents dans la base de données ou d'un document où les individus d'un groupe ne sont pas ordonnés selon un critère particulier.

5.4.4.4 Une certaine adéquation au besoin

L'étude de la seconde composante de l'évaluation utilisateurs, consacrée à la notion d'adéquation à la tâche, a permis de caractériser une certain nombre d'éléments.

Tout d'abord, une analyse quantitative des résultats des questionnaires pour la partie correspondant à cette dimension semble indiquer une satisfaction générale des usagers, malgré quelques difficultés.

Ensuite, une analyse qualitative des commentaires des utilisateurs a révélé quelques points posant particulièrement problème. En particulier, l'évaluation des composantes adéquation, expressivité et complétude a été estimée par plusieurs utilisateurs comme difficile après seulement une session d'environ une heure avec l'outil. De plus, la transposition d'un problème biologique sous forme d'une requête structurée est un problème difficile, qui demande un important effort cognitif pour attribuer des rôles aux différentes composantes du problème. Pour finir, les études libres ont permis de révéler quelques limites de l'expressivité de la requête, par exemple la possibilité laissée à l'usager de construire une requête incohérente, l'absence de prise en compte d'une notion similaire au OR booléen, l'impossibilité de définir des éléments composés ou l'obligation de remplir tous les champs de la partie Besoins décrivant le cœur de l'étude.

5.4.5 Pertinence interprétationnelle

5.4.5.1 Introduction

La notion de pertinence interprétationnelle recouvre des problématiques liées à l'évaluation des résultats de la synthèse, à la construction de connaissances par l'usager à partir d'un document synthétique. L'analyse de cette notion a été menée

en détails au cours de la présentation des bases conceptuelles du système d'assistance à la synthèse, au Paragraphe 3.4.4.3. Il s'agit en effet, comme pour l'adéquation à la tâche, d'une composante du modèle de synthèse.

Il s'agit alors, par le biais de l'étude utilisateurs, de mener une estimation à la fois quantitative et qualitative de cette pertinence interprétationnelle, selon ses diverses dimensions qui ont été relevées au Paragraphe 3.4.4.3. Ces multiples axes d'évaluation font l'objet de la partie Résultats de la requête du questionnaire d'évaluation. En cours de séance de test, les utilisateurs ont en général aussi exprimé un certain nombre de remarques relevant de cette composante de l'évaluation, qu'il s'agisse de commentaires écrits ou oraux. L'analyse de ces résultats va donc être menée en deux étapes.

Tout d'abord, après un rappel des éléments évalués, va être présentée une analyse quantitative des résultats des questionnaires. Ensuite, une analyse qualitative des remarques des usagers va être proposée.

5.4.5.2 Analyse quantitative

La section «Résultats de la requête» du questionnaire regroupe 11 affirmations organisées selon les dimensions de la pertinence interprétationnelle : intuitivité, informativité, utilité, suggestivité et navigabilité. Ces diverses dimensions de la pertinence interprétationnelle ont été identifiées au Paragraphe 3.4.4.3. Elles sont rappelées au sein du Tab. 5.9, avec les affirmations correspondantes et les codes associés à chaque affirmation, pour présentation compacte des résultats.

Les résultats de cette partie du questionnaire, correspondant là encore à une valeur entre 1 (pas d'accord) et 5 (tout à fait d'accord) pour chaque affirmation et pour chaque utilisateur sont présentés en Annexe L. Ces données quantitatives ont servi de base à la construction des boîtes à moustaches présentées Fig. 5.22.

Ces résultats montrent là aussi un jugement plutôt positif des usagers en ce qui concerne la pertinence interprétationnelle. Toutes les affirmations présentent une appréciation supérieure à la moyenne. Une observation plus détaillée des diverses dimensions permet de tirer des conclusions plus fines.

Premièrement, l'intuitivité présente globalement les moins bonnes évaluations de la pertinence interprétationnelle. Comme l'a remarqué l'un des utilisateurs, la présentation sous forme de grille correspond à un autre mode de pensée par rapport aux présentations habituelles, ce qui nécessite un temps d'adaptation. Cette adaptation est jugée comme rapide : «une fois qu'on a vu une fois comment ça marche, c'est facile à comprendre», comme l'a noté un autre usager. Mais la durée limitée de la séance de test a induit que celle-ci a majoritairement été consacrée à cette

Tab. 5.9: Dimensions de la pertinence interprétationnelle - Les dimensions envisagées pour la pertinence interprétationnelle sont présentées ici avec des exemples d'affirmations pour lesquelles l'usager devra donner un niveau d'approbation dans le cadre d'un questionnaire.

Dimension de la pertinence interprétationnelle	*Exemples d'affirmations*	*Code*
Intuitivité	- La relation entre la requête et le résultat affiché est facilement compréhensible.	Int1
	- L'interprétation du résultat de la requête ne pose pas de problème.	Int2
Informativité	- La forme du rendu (tableau) est pertinente.	Inf1
	- Le résultat proposé est porteur de sens.	Inf2
	- Le résultat proposé apporte de nouvelles informations par rapport au problème posé.	Inf3
	- Je sais quand il y a des erreurs de traitement.	Inf4
Utilité	- Le document de synthèse affiché m'aide à répondre au problème exprimé dans la requête.	U1
	- Le résultat pourrait être directement utilisé par exemple dans une publication.	U2
Suggestivité	- Les résultats affichés m'aident à envisager de nouvelles requêtes.	S1
	- Les résultats affichés suggèrent d'autres analyses avec d'autres outils.	S2
Navigabilité	- La navigation au sein des informations (grille et fiches associées) est facile.	N

adaptation à la présentation, conduisant à une évaluation de l'intuitivité mitigée, en particulier pour les utilisateurs dont l'expertise biologique ou l'habitude de l'outil informatique est plus faible.

Deuxièmement, l'informativité est par contre plutôt bien jugée, mis à part peut-être la communication des erreurs de traitement. Les utilisateurs ayant été confrontés à des erreurs de traitement ont apprécié que l'information soit donnée, mais l'explicitation de l'erreur reste très technique, par un affichage mis en forme d'exceptions JAVA, ce qui déroute les usagers ayant peu de connaissances informatiques. La forme de grille est jugée comme pertinente, porteuse de sens et apportant de nouvelles informations sur le problème étudié. Sur ce dernier point, seuls l'anatomopathologiste et l'un des étudiants ont un avis plus mitigé. Dans le cadre de la session de test, les études menées correspondent à des problèmes biologiques dont les résultats sont communément connus. Ils ont avoué des difficultés à s'abstraire du fait que dans les cas précis étudiés, aucune information nouvelle n'était présentée, pour imaginer le potentiel intérêt sur des jeux de données moins communs.

Ensuite, l'utilité est elle aussi plutôt bien appréciée. L'utilisation directe de la grille comme illustration d'une publication n'est pas forcément bien envisagée, par contre, les utilisateurs trouvent qu'elle aide à la résolution du problème, à l'exception de l'anatomopathologiste, là aussi parce que les problèmes étudiés n'étaient pas

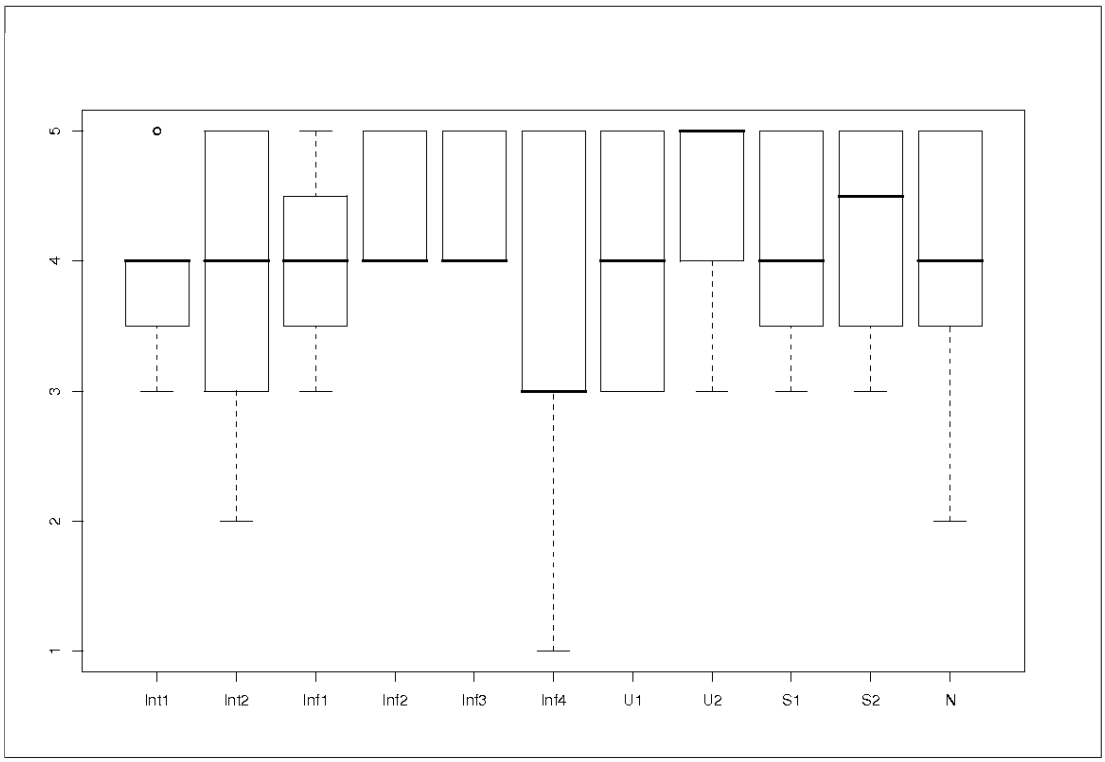


Fig. 5.22: Boîtes à moustaches pour les questions de pertinence interprétationnelle - Ces boîtes correspondent aux résultats de la partie Résultats de la requête du questionnaire. Les diverses questions sont représentées par leur code tel que défini dans le Tab. 5.9.

originaux.

En ce qui concerne la suggestivité, certains utilisateurs se sont montrés très enthousiastes, exprimant que le document de synthèse «ouvre plein de perspectives et donne plein d'idées» ou que «ça fait un support pour penser et ça donne envie d'essayer d'autres choses ou de creuser plus loin». Seuls l'anatomopathologiste et un étudiant sont moins convaincus de cette suggestivité, peut-être parce que le document de synthèse induit un raisonnement qui sort du cadre réflexif traditionnel de la profession (pour l'anatomopathologiste) ou qui leur a été enseigné (pour l'étudiant).

Enfin la navigabilité est jugée satisfaisante, sauf par l'anatomopathologiste, dont les difficultés sont peut-être liées à son habitude modérée de l'outil informatique.

Ces résultats quantitatifs sont associés à des remarques, écrites ou orales, qui apportent des informations complémentaires sur l'interprétation du document de synthèse et en particulier permettent de définir des manques et des extensions possibles. Leur étude est présentée dans le prochain paragraphe.

5.4.5.3 Analyse qualitative

5.4.5.3.1 Introduction

Comme dans le cas de l'adéquation à la tâche, les évaluations quantitative menées à partir des réponses au questionnaire sont associées à des remarques, inscrites en bas de questionnaire ou orales, dont certaines sont focalisées sur les problématiques d'interprétation du document de synthèse. En tant que telles, ces remarques relèvent de la pertinence interprétationnelle et servent de base à l'évaluation qualitative menée ici.

Les divers commentaires des usagers ont pu être classés en deux catégories, qui seront explorées plus en détails par la suite : éléments facilitant l'exploitation du document de synthèse et extensions du document qui peuvent être porteuses d'informations utiles dans un objectif d'appréhension des données.

5.4.5.3.2 Éléments facilitant l'exploitation du document de synthèse

Parmi les remarques formulées par les usagers au cours de leur interprétation des documents de synthèse résultant des exemples d'études menées au cours de la session de test, certaines peuvent être considérées comme des éléments visant à faciliter l'exploitation du document.

Tout d'abord, un élément essentiel a été omis lors de la conception du document de synthèse, dont le manque a été unanimement noté : une légende indiquant la correspondance entre les couleurs des cases de la grille et les valeurs du concept du domaine choisi comme but de l'étude. En effet, actuellement, il faut aller consulter pour chaque couleur la fiche individuelle d'un individu et y rechercher l'élément défini comme but pour réaliser cette correspondance, ce qui est peu pratique. De plus, une telle légende serait essentielle si la grille était utilisée comme illustration d'une publication.

Ensuite, toujours à des fin de facilitation de l'interprétation, il a été noté que dans les cas de comparaison, présenter les titres des différents niveaux de groupement dans des cases aux couleurs différentes, une couleur par niveau, permettrait d'appréhender au premier coup d'œil les zones de grille attribuées à chaque niveau de groupement. Conjointement, les éléments variables de ces titres pourraient être présentés en une couleur différente du reste, permettant là aussi d'identifier les différentes valeurs d'un concept, correspondant à différents groupes, plus facilement.

De plus, il a été remarqué qu'il serait intéressant de proposer une vue conjointe de plusieurs documents de synthèse, correspondant à des études différentes. Actuelle-

ment, les documents de synthèse sont toujours présentés dans le même onglet, et l'ouverture d'un nouveau document provoque le remplacement du précédent. L'une des suggestions est d'ouvrir les documents à chaque fois dans un nouvel onglet, en donnant la possibilité de fermer les onglets. Dans le même esprit, il a été noté comme potentiellement utile de proposer une vue réduite du document correspondant à la requête d'origine, lors des reformulations.

En fait, ce type de besoins correspond à ce que je considère comme des études composées, par exemple la comparaison en fonction de la classe d'âge de l'évolution du nombre de ganglions envahis en fonction du nombre de ganglions observés en visualisant la composante N du stade. Ce type d'études a été envisagé dans le cadre de de la conception originelle du système d'assistance à la synthèse, et certains éléments de développement, tels la forme des requêtes et des modèles de tâches, ainsi que certains composants, ont été développés pour gérer ces études composées, mais leur support n'est pas encore complet et demanderait une analyse et des développements complémentaires.

Enfin, le document de synthèse a été proposé comme pouvant servir d'illustration à des publications ou comme source de données pour d'autres outils, tels que des logiciels de fouille de données. Actuellement, ces rôles du document de synthèse sont encore peu supportés, et il faudrait proposer des fonctionnalités d'exportation, sous forme d'image pour une illustration, et sous forme de fichier texte de type CSV pour les données.

5.4.5.3.3 Extensions au document de synthèse

Conjointement aux suggestions d'éléments qui permettraient de faciliter l'exploitation du document de synthèse, des extensions, porteuses d'informations complémentaires ou permettant d'explorer plus avant le jeu de données, ont aussi été exprimées.

Tout d'abord, il a été proposé de mettre en place, pour chaque groupe d'une comparaison, un affichage de statistiques descriptives simples de type nombre total d'individus, moyenne, écart-type, variance, minimum ou maximum du concept choisi comme but de l'étude s'il est quantitatif, etc. Ces informations pourraient être accessibles par un lien vers une fiche récapitulative présent dans la case de titre du groupe, ou dans une info-bulle affichée lors du passage de la souris sur ce même titre. Le même type de fonctionnalités pourraient être en place pour les évolutions, au niveau des cases d'un individu moyen représentant plusieurs individus.

Ensuite, il a été suggéré de proposer des liens vers des outils de fouille de données plus avancés, et en particulier des outils statistiques. Cette problématique avait été évoquée dès avant les tests utilisateurs. Aussi, une étude préliminaire de l'intégra-

tion de statistiques simples, sous forme de graphiques de type boîtes à moustaches ou histogrammes, a été menée dans le cadre du stage de fin d'année d'un étudiant de Master 2 Biologie et Informatique. Cette intégration a été proposée comme passant par l'exécution de routines R[4] à partir de l'interface sur des données issues du document maître qui sert de base à l'affichage. Mais l'ensemble des problèmes posés par cette intégration n'ont pas encore été résolus.

Enfin, une vision encore plus avancée de cette idée de manipulation de la grille documentaire évoque de mettre en place un système permettant de réaliser un focus ou une sélection sur une partie de la grille, et de permettre de nouvelles manipulations basées sur la sélection d'individus correspondants, par exemple une nouvelle requête ou des analyses statistiques.

5.4.5.4 Une pertinence interprétationnelle encourageante

L'étude de la troisième composante de l'évaluation utilisateurs, consacrée à la notion de pertinence interprétationnelle, a permis de relever un certain nombre de points.

Premièrement, une étude quantitative des résultats des questionnaires pour la partie Résultats de la requête a révélé une satisfaction générale des usagers, malgré quelques réserves.

Ensuite, une analyse qualitative des remarques des utilisateurs a permis de mettre en exergue des manques ou des évolutions intéressantes pour le document de synthèse. En particulier, il faudrait intégrer une légende, différencier l'affichage des titres des groupes selon le niveau de groupement et la valeur ayant servi à grouper pour les comparaisons, proposer une vue conjointe de plusieurs études, qui a été évoquée comme correspondant à des requêtes composées, et permettre une exportation sous forme d'image ou de fichier texte de données. Des évolutions intéressantes seraient l'affichage de quelques statistiques descriptives, des liens vers des outils de fouille de données, et l'amélioration de l'interactivité de la grille documentaire, par exemple par la possibilité d'utiliser une partie de grille comme focus pour une nouvelle étude ou une fouille de données.

[4] http ://www.r-project.org/

5.4.6 Performances

5.4.6.1 Introduction

Dans le cadre du développement d'un prototype, les performances du système, en particulier en ce qui concerne les temps de traitement et la réactivité aux commandes utilisateur, sont rarement jugées comme critiques. En effet, il s'agit en général de valider un concept, de permettre aux utilisateurs de se faire une première idée de la proposition de système qui devrait répondre aux besoins qu'ils ont exprimés.

Ici, cette notion de performances n'est pas non plus d'une importance capitale, puisque l'objectif est surtout d'estimer l'intérêt des propositions qui ont été faites aux chapitres précédents, qu'il s'agisse du modèle de synthèse ou du prototype qui opérationnalise ce modèle.

Une évaluation de performances permet toutefois, dans une perspective d'évolution du prototype vers un système plus complet, de pointer dès ce stade précoce du développement les éventuels problèmes de performances à résoudre en priorité. En effet, le prototype, ainsi que les tests utilisateurs qui viennent d'être décrits semblent retenir l'intérêt des usagers potentiels, et l'on peut envisager que des travaux en ce sens se poursuivent.

Il paraît alors pertinent de mener une petite étude informelle de performance, pour identifier les points faibles éventuels et suggérer des solutions possibles, objet du prochain paragraphe.

5.4.6.2 Évaluation des performances

L'évaluation des performances est basée sur des chronométrages réalisés de manière grossière lors des tests en cours de développement et lors de l'étude utilisateurs, pour les études libres. Au total, ces mesures concernent une vingtaine d'études différentes.

Pour chaque étude, a été évalué le temps écoulé pour :

- ⋆ Présenter la page d'accueil
- ⋆ Présenter le premier écran de saisie de requête (informations générales),
- ⋆ Présenter l'écran de saisie de contraintes expérimentales,
- ⋆ Présenter l'écran de saisie des besoins,
- ⋆ Présenter l'écran de reformulation de requête
- ⋆ Exécuter la requête,
- ⋆ Afficher le document de synthèse,

⋆ Afficher la fiche individuelle d'un patient,

⋆ Afficher la fiche individuelle d'une donnée histologique.

Ces mesures ont été réalisées localement sur ma machine de développement, un Pentium IV dual-core à 1.83GHz avec 1Go de RAM, pour éliminer les effets de latence réseau.

Un résumé des résultats est présenté dans le Tab. 5.10.

Tab. 5.10: Estimations de performances du système - Pour les divers éléments présentés, une fourchette de temps écoulé est proposée.

Élément mesuré	*Temps minimum*	*Temps maximum*
Affichage : Page d'accueil	1s	3s
Affichage : Requête - Informations générales	1s	3s
Affichage : Requête - Contraintes expérimentales	2s	4s
Affichage : Requête - Besoins	1s	3s
Affichage : Requête - Reformulation	12s	20s
Exécution : Requête	2s	2 min
Affichage : Document de synthèse	2s	10s
Affichage : Fiche patient	3s	5s
Affichage : Fiche histologique	3s	6s

Ces mesures font apparaître pour la plupart des éléments des délais d'affichage de quelques secondes, qui sont compatibles avec un système interactif, même si des améliorations pourraient être envisagées. Par contre, deux éléments ont des temps de traitements qui posent problème : l'exécution de la requête, soit le processus de synthèse en tant que tel, et l'affichage des formulaires de reformulation.

Il convient alors de s'interroger sur les causes de ces temps élevés et de suggérer des possibilités d'amélioration, réflexion menée dans le prochain paragraphe.

5.4.6.3 Critique et suggestions d'amélioration

Les deux éléments posant problème au niveau performances, l'affichage du formulaire de reformulation de requête et le traitement de la requête, vont être détaillés successivement.

En ce qui concerne l'affichage du formulaire de reformulation de requête, celui-ci repose sur des éléments logiciels développés au sein du framework Orbeon Forms. Or, il s'agissait pour moi d'une technologie nouvelle, que j'ai apprise au cours du développement du prototype. La génération du formulaire de reformulation à partir d'un fichier de requête a été réalisée de manière simpliste et non optimale à une période où ma maîtrise du framework était encore faible. Ce sont principalement

ces erreurs de développement qui expliquent les délais d'affichage importants, et une reprise du code correspondant pourraient facilement améliorer les performances.

En ce qui concerne le traitement de la requête, une analyse plus fine du processus de synthèse, et en particulier d'exécution d'une instance de tâche, est nécessaire, afin d'identifier les composants dont le traitement est le plus long. Des exécutions manuelles, présentant des traces des composants en cours d'exécution, ont permis d'identifier le composant constituant la liste des groupes à intégrer comme le premier responsable des temps de traitement élevés.

En effet, ce composant détermine, pour chaque niveau de groupement dans les cas de comparaison, et pour chaque valeur selon chacune des deux dimensions pour les cas d'évolution, toutes les combinaisons possibles, en se basant sur les connaissances du domaine d'étude. Puis il vérifie en base de données quels groupes ne sont pas vides. Ainsi, pour une évolution du nombre de ganglions envahis en fonction du nombre de ganglions observés, concepts dont les valeurs vont de 0 à 80, avec un pas ou taille de classe de 1, ce composant exécute $80x80 = 6400$ requêtes SQL.

Afin d'améliorer ces performances, il vaudrait mieux exécuter une requête SQL pour extraire la liste des valeurs distinctes pour chaque dimension, vérifier qu'elles sont bien en conformité avec les connaissances du domaine, puis utiliser ces listes réduites de valeurs au lieu de se baser uniquement sur les connaissances expérimentales.

Ainsi, ces quelques chronométrages ont permis d'identifier les éléments qui posent des problèmes de performances et des pistes de correction ont pu être évoquées.

5.4.7 Une étude utilisateurs généralement positive

L'étude utilisateurs qui a été menée auprès des usagers potentiels du système, recrutés au sein de l'équipe du projet TMA-Explorer a permis, sur la base d'un scénario de test, l'acquisition tout à la fois quantitative et qualitative d'un certain nombre de métriques.

En matière d'utilisabilité, d'adéquation à la tâche et de pertinence interprétationnelle, les résultats quantitatifs, obtenus sur la base d'un questionnaire d'évaluation, ont permis de noter une certaine satisfaction générale des usagers. Ces mesures doivent toutefois être interprétées avec précaution, car leur valeur statistique est peu significative, étant donné la taille du panel réduite.

Les évaluations quantitatives ont été affinées par une évaluation qualitative, sur la base de commentaires oraux et écrits. Ces commentaires ont permis de noter un effet de l'apprentissage sur l'utilisation du prototype, et en particulier la formulation

de requête, qui reste une étape intellectuellement critique pour l'utilisation de l'outil.

Des limites à l'expressivité de la requête ont aussi pu être identifiées. De plus, l'observation du document de synthèse a donné lieu à nombre de suggestions intéressantes, tant pour faciliter son interprétation que pour l'étendre, dans une perspective d'exploration des données, et en particulier de changement de focus ou d'analyse statistique.

Ensuite, une variabilité de perception du système selon les utilisateurs, et en particulier selon leur expertise biologique et informatique a été notée. Cette variabilité permet de valider la notion d'archétype qui a été définie au sein du modèle de synthèse mais n'a pas été intégrée au prototype, par manque de temps.

Enfin, des mesures de performances, consistant en des chronométrages des différents affichages et traitements, ont permis d'identifier des éléments du prototype posant problème : l'affichage du formulaire de reformulation de requête et l'exécution d'une instance de tâche. Les causes de ces manques de performances ont pu être identifiées et des solutions proposées.

5.5 Une évaluation porteuse d'enseignements

Ce chapitre a été consacré à une validation du modèle de synthèse proposé au Chapitre 3. Cette validation a été menée par l'intermédiaire d'expérimentations conduites avec le prototype basé sur le modèle qui a été présenté au Chapitre 4. Ces expérimentations ont été envisagées selon une perspective d'évaluation qualité classique en ingénierie logicielle.

Une revue rapide de quelques travaux dans le domaine de l'évaluation logicielle et de l'évaluation de sites Web, qui sont tous deux d'intérêt dans le cadre d'une application Web telle que le système proposé, ont permis de cerner quelques directions d'évaluation pertinentes pour le prototype. En particulier, quelques métriques et procédures d'évaluation ont été choisies.

Tout d'abord, des études de cas ont permis d'apporter un point de vue diagnostic, orienté vers les fonctionnalités du système. Ces études de cas ont été consacrées à l'étude de deux exemples de problèmes biologiques, relevant des deux tâches prototypiques disponibles au sein du prototype : comparaison et évolution. Elles ont permis de montrer que les études envisagées peuvent être exprimées sous forme de requêtes structurées, donnant un premier aperçu de l'expressivité de la requête.

Les documents de synthèse construits suite à l'exécution des instances de tâches correspondantes ont permis de tirer des conclusions biologiques ou de mettre en

avant des données incohérentes. Ceci montre ainsi leur utilité dans l'appréhension de la collection de données, que ce soit pour valider des hypothèses dans le contexte d'une démarche expérimentale classique ou pour préparer une fouille de données. Ces documents ont aussi conduit à la conception de nouvelles études ou à la proposition d'utilisation d'autres outils, en particulier statistiques, illustrant leur suggestivité.

Ensuite, une étude utilisateurs, conduite auprès d'un panel d'usagers potentiels recrutés au sein de l'équipe du projet, a permis d'apporter un point de vue usage, orienté vers les perceptions des utilisateurs. Menée sur la base d'un scénario de test guidant l'interaction de chaque utilisateur avec le système, cette étude a permis l'acquisition de données quantitatives, par les résultats d'un questionnaire, et qualitative, par des commentaires oraux et écrits des utilisateurs du panel.

Consacrée à l'évaluation de l'utilisabilité, de l'adéquation à la tâche et de la pertinence interprétationnelle, elle a révélé un certain enthousiasme des utilisateurs, perceptible tout à la fois dans les évaluations en général supérieures à la moyenne et les nombreuses suggestions d'améliorations et extensions du prototype.

Cette étude utilisateurs a aussi été prétexte à une évaluation de performances, qui a permis d'identifier deux composantes du système posant problème : la reformulation de requête et un composant qui construit la liste des groupes possibles. Des suggestions de corrections ont été proposées.

Ces ensembles de suggestions d'évolutions et propositions d'améliorations ouvrent un ensemble de perspectives sur le projet, qui sont présentées en accompagnement d'un état de lieux des travaux réalisés, dans le chapitre suivant.

CHAPITRE

6 Conclusion et Perspectives

L'objet de ma thèse a tout d'abord été l'identification d'un problème d'appréhension de données associé aux technologies à haut débit comme les Tissue MicroArrays. Une solution, la notion de synthèse, qui a été replacée dans le contexte des disciplines auxquelles elle peut se rattacher, a été introduite. Cette proposition a été explicitée par un modèle, qui a été replacé dans le cadre des modèles de Recherche d'Information courants, et dont les diverses composantes ont été décrites plus précisément. Ce modèle a été opérationnalisé au sein d'un prototype, développé dans le cadre du projet TMA-Explorer. Une validation expérimentale du modèle a étéentreprise, par des études de cas et tests utilisateurs menés avec le prototype. Dans ce cadre, cette conclusion vise à apporter une vision globale sur les travaux réalisés, tout en ouvrant de nouvelles perspectives, dérivées des suggestions des utilisateurs ou envisagées dès le début des travaux.

6.1 Bilan des travaux

Afin de porter à son terme la présentation de mes travaux proposée ici, il convient d'effectuer un état des lieux des réalisations, qu'il s'agisse de l'analyse du problème de synthèse, l'avancée du développement, ou les conclusions qui peuvent en être tirées. Ce bilan des travaux réalisés dans le cadre de ma thèse peut être abordé selon un axe théorique, autour de la définition d'une notion de synthèse et d'un modèle de Recherche d'Information orientée tâche, et un axe pratique, focalisé sur le prototype développé et son expérimentation.

D'un point de vue théorique, une notion de synthèse, présentée comme une solution au problème d'appréhension des données rencontré par les chercheurs recourant à des technologies de traitement en masse d'échantillons ou échangeant des données avec d'autres équipes, a été introduite. Cette synthèse a été envisagée comme un processus regroupant des activités de sélection et agrégation d'informations, d'organisation conceptuelle puis organisation structurelle des informations et de présentation du document de synthèse ainsi construit, le tout réalisé dans un objectif précis, pour un public particulier.

L'analyse de ce problème de synthèse a été menée selon deux axes. Dans un premier temps, les champs disciplinaires d'intérêt pour la synthèse ont été identifiés. Il s'agit de domaines permettant l'appréhension des données, comme la fouille de données ou la Visualisation d'Information, de domaines permettant l'exploration des informations, comme la Recherche d'Information, et de domaines permettant la représentation d'entités, tels que l'utilisateur, pour les systèmes adaptatifs, ou les connaissances du domaine et le problème à résoudre, pour l'Intelligence Artificielle.

Dans un second temps, le processus de synthèse en tant que tel a été exploré dans un cadre très particulier, le domaine applicatif des Tissue MicroArrays. Ce point de vue très restrictif a conduit à une limitation du champ d'investigation à des types de problèmes de synthèse particuliers, ceux se basant sur les données, et à des ensembles informationnels spécifiques, des documents structurés regroupant des données quantitatives ou qualitatives. Ces restrictions ont permis de laisser de côté les problèmes liés au traitement du texte libre, qu'il s'agisse de problèmes de Recherche d'Information en tant que telle, ou d'organisation d'entités en fonction de valeurs de concepts extraites de textes, etc.

De plus, ce domaine applicatif se prête bien à la métaphore de la grille utilisée au cœur du document de synthèse. Cette notion de grille fait en effet partie du champ expérimental des TMA et le recours à une métaphore proche apporte une aide ancrée dans le réel pour défricher un champ scientifique relativement vierge, la synthèse d'informations.

Dans ce cadre précis, un modèle de synthèse, basé sur un paradigme de Recherche

d'Information orientée tâche, a pu être défini. D'un point de vue pratique, le modèle de synthèse a servi de base à la définition d'une architecture fonctionnelle pour un système d'assistance à la synthèse envisagé dans un premier temps dans le domaine applicatif des TMA. Ce prototype a été développé dans un objectif de simplicité, afin de valider le modèle conceptuel proposé, tout en prenant en compte des prérequis techniques, tels que l'intégration potentielle au sein d'une plateforme dédiée à la technologie des TMA.

Cet attachement à la simplicité a conduit au choix d'une représentation de l'ensemble des entités impliquées dans la synthèse sous forme XML, ainsi qu'à la mise en place d'une architecture à composants spécifique, où les composants ont tous une interface identique et communiquent par l'intermédiaire d'un tableau noir. Dans le même esprit, les heuristiques mis en place restent très simples, tout en laissant des possibilités de complexification, et la notion d'archétype utilisateur a été laissée de côté.

Surtout, le prototype a été limité dans sa couverture. D'un part, seules deux tâches prototypiques ont été mises en place, l'une relevant des problématiques de comparaison, l'autre des problématiques d'évolution. Le nombre de composants, qui permettent la résolution de tâches élémentaires et sont combinés pour résoudre une tâche prototypique, est donc fortement limité. D'autre part, le domaine applicatif a été réduit à un cœur essentiel, qu'il s'agisse des connaissances du domaine ou des données utilisées.

Enfin quelques problèmes ont pu être identifiés. Au niveau purement technique, un composant de constitution de groupes et la construction du formulaire de reformulation de requête présentent des performances assez médiocres et devraient être améliorés. Au niveau conceptuel, des limites à l'expressivité de requête ont pu être notées, comme l'impossibilité de représenter des éléments composés ou l'absence d'un notion de type OR booléen.

Au final, les tests menés dans le cadre de la validation expérimentale du modèle et du prototype ont pu montrer de premiers indices plutôt encourageants quant à la validité du concept de synthèse dans le cadre posé, c'est-à-dire l'exploration des données TMA. Ainsi, dans le cadre des études de cas, les problèmes biologiques d'exemple ont pu être exprimés sous forme de requête et le document de synthèse a tout à la fois apporté des informations intéressantes quant au problème étudié, et permis de suggérer une exploration plus poussée des données, par une analyse statistique ou de nouvelles requêtes. Malgré la faible taille du panel impliqué dans l'étude utilisateurs, les résultats du questionnaire, les suggestions d'améliorations et évolutions du prototype, semblent montrer un certain intérêt des usagers potentiels.

Ces constats suggèrent qu'il serait intéressant d'aller plus loin dans l'exploration tant du concept de synthèse que de l'amélioration et évolution du prototype. Ces diverses perspectives vont faire l'objet de la prochaine section.

6.2 Perspectives

6.2.1 Amélioration du prototype

6.2.1.1 Intégration des éléments laissés de côté

Le prototype de système d'assistance à la synthèse qui a été développé implante une version simplifiée de l'architecture qui a été déduite du modèle de synthèse. En effet, dans le temps imparti pour ma thèse, il ne m'a pas été possible de réaliser un système complet. Par rapport à ce qui a été proposé dans les bases conceptuelles de la synthèse, et en particulier parmi les composantes du modèle de synthèse, le prototype n'en opérationnalise donc qu'une partie.

D'une part, seules deux tâches prototypiques, une relevant de la problématique de comparaison, et l'autre relevant de la problématique d'évolution, ont été intégrées au prototype et seuls les composants pouvant être utilisés pour la résolution de ces tâches prototypiques ont été développés. L'amélioration du prototype pourrait alors passer par l'ajout d'autres tâches prototypiques, impliquant l'ajout des composants nécessaires à leur exécution.

En particulier, il serait intéressant d'intégrer une tâche de type distribution, pour couvrir l'ensemble des catégories de tâches. Cette intégration impliquerait sans doute une analyse plus fine de la problématique de distribution, et surtout de son adéquation avec une représentation sous forme de grille, qui mérite d'être affinée. En particulier, la compacité de la visualisation d'une distribution, telle qu'envisagée, reste limitée. Intégrer une tâche de type distribution ne serait pas alors un simple problème technique mais permettrait de progresser dans l'approfondissement de la notion de modèle de tâche et de requête et de revenir sur la métaphore de la grille et son lien à la tâche.

De plus, la notion de tâche composée a été évoqué dans le cadre des suggestions faites par les utilisateurs. Cette composition de tâches, en permettant la construction de documents de synthèse correspondant par exemple à la comparaison d'évolutions, apporterait une flexibilité complémentaire à la formulation de requête. Une simple esquisse de cette idée fait déjà apparaître des difficultés impactant des éléments majeurs du système de synthèse. La forme de la requête serait altérée, la possibilité de composition devrait être prise en compte dans le modèle de tâche, les composants résolvant certaines sous-tâches élémentaires se verraient complexifiés : la constitution de groupes d'individus est sans doute très différente de la constitution de groupes de grilles documentaires correspondant à d'autres instances de tâches.

D'autre part, une notion d'archétype utilisateur, correspondant à des catégories d'usagers aux connaissances sur le domaine d'étude et modes de raisonnement sur ce

domaine similaires, a été introduite dans le modèle de synthèse. L'étude utilisateur menée avec le prototype a aussi permis de donner des indices sur l'existence de ces archétypes. En effet, malgré la faible taille du panel, l'effet de l'apprentissage sur l'efficacité des utilisateurs avec l'outil a paru comme liée aux connaissances biologiques et informatiques des usagers. De plus, l'utilisateur ayant le moins de connaissances biologiques a exprimé un malaise vis à vis du domaine d'étude et noté la trop grande complexité de la taxonomie du domaine par rapport à ses connaissances. La notion d'archétype semble donc bien couvrir un phénomène existant. Il serait intéressant de l'étudier plus avant et d'intégrer une personnalisation basée sur des archétypes dans le prototype, ainsi que cela était prévu à l'origine.

Enfin, au sein du prototype, seuls des types de concepts simples ont été pris en compte dans les connaissances du domaine. En effet, les éléments quantitatifs, qualitatifs ou booléens ne requièrent pas le recours à des méthodes d'accès plus complexes que de simples correspondances exactes (requêtes de type SELECT en base de données). Mais à plus ou moins long terme, il faudrait pouvoir prendre en compte du texte libre. En effet, au sein des dossiers cliniques des patients, certains éléments, tels que des comptes-rendus de visite, correspondent à des textes rédigés par le médecin. De même, les commentaires de l'anatomopathologiste sur les images de lames histologiques ne consistent pas uniquement en annotations par des mots-clés. La prise en compte de tels textes induisent l'extension du système d'assistance à la synthèse vers de vraies problématiques de Recherche d'Information, et l'ajout de composants permettant la sélection ou l'organisation des items selon des concepts dont la valeur peut être extraite de textes.

6.2.1.2 Intégration de suggestions des utilisateurs

Les tests utilisateurs conduits dans le cadre de la validation expérimentale du prototype ont conduit à l'expression d'un ensemble de commentaires, pour certains consistant en suggestions d'amélioration du système. Ces divers éléments ont été répertoriés au Paragraphe 5.4.4.3 en ce qui concerne l'expression de la requête, et au Paragraphe 5.4.5.3 pour les remarques relevant de l'interprétation du document de synthèse.

Ainsi, au niveau de la formulation de l'étude à réaliser, quelques limites à l'expressivité de la requête ont été notées.

Tout d'abord, il faudrait permettre la saisie d'éléments composés. Pour ce faire, il faudrait premièrement étendre la représentation des connaissances du domaine pour permettre la définition de types de relations supplémentaires entre concepts. Ceci permettrait de décrire des éléments tels que le pourcentage de cellules marquées en marqueur Ki67 par exemple. Ces relations entre concepts permettraient aussi de mettre en place un contrôle de la saisie, par neutralisation d'une partie de la

taxonomie en fonction des choix qui ont déjà été faits, évitant ainsi la construction de requêtes incohérentes.

Ensuite, il faudrait permettre la saisie d'alternatives, par exemple la sélection d'individus selon un critère de sexe ou d'année de naissance (les patients de sexe féminin ou nés avant 1950). Ceci implique l'intégration d'une notion correspondant au OR booléen, qui n'est supportée que de manière marginale et automatique, en particulier quand des critères de sélection portent sur le même concept.

Enfin, certains rôles de la requête devraient être optionnels, permettant une sélection de tous les individus, ou un tri par défaut. Au niveau pratique, il faudrait mettre en place un système d'exécution des composants plus souple. Par exemple, dans le cas d'une absence de critères de tri pour une tâche de comparaison, le composant qui ordonne les individus ne serait exécuté que si le critère est défini. Les composants qui dépendent de son résultat verraient le paramètre correspondant à la source de leurs données modifiée à l'exécution, pour utiliser à la place la source de données du composant d'ordonnancement.

Au niveau de l'interprétation du document de synthèse, quelques améliorations et extensions ont aussi été proposées.

Premièrement, l'interprétation du document serait grandement facilitée par quelques additions simples : une légende indiquant une correspondance entre les couleurs de cases et les valeurs du concept défini comme but de l'étude, l'utilisation d'un code couleur dans les titres des groupes des comparaisons pour différencier les groupes et niveaux de groupement, la visualisation simultanée de plusieurs études ou vues patient et histologie, au sein d'onglets multiples, des fonctionnalités d'exportation de la grille documentaire à d'autres formats.

Deuxièmement, des extensions au document de synthèse pourraient utilement compléter la visualisation. Par exemple, des calculs d'analyse statistique descriptive simples pourraient être menés sur chaque groupe d'individus (groupes d'une comparaison, ou ensemble d'individus représentés par un individu moyen pour une évolution) et les résultats présentés de manière contextuelle à partir de la grille documentaire. De plus, des liens vers des outils de fouille de données plus traditionnels, qu'il faudrait intégrer au système, seraient les bienvenus. Enfin une manipulation plus avancée de la grille, permettant des focus, zooms, sélections serait d'un intérêt tout particulier, que ce soit pour mener de nouvelles études avec l'outil sur le jeu de données constitué par les éléments sélectionnés, ou pour exploiter les données correspondantes avec un outil de fouille de données.

6.2.2 Ouverture vers d'autres domaines applicatifs

6.2.2.1 Introduction

Bien que le modèle de synthèse ait été conçu en gardant à l'esprit le domaine applicatif des TMA, celui-ci a tout de même été envisagé comme relativement générique. En particulier, l'idée sous-jacente à sa mise en place et surtout à la conception et au développement du prototype de système d'assistance à la synthèse a toujours été la possibilité d'utiliser le système dans d'autres domaines applicatifs.

Par exemple, au sein de l'équipe, les travaux menés sur les mécanismes d'oncogenèse ne se basent pas uniquement sur l'étude de ce processus au niveau protéomique, par analyse de l'expression de diverses molécules selon le tissu en utilisant la technologie des TMA. Ainsi, certains membres de l'équipe s'intéressent aux problématiques d'activation des gènes et à l'identification des gènes possiblement impliqués dans la régulation du cycle cellulaire. Ils se basent sur des données acquises sur des puces à ADN par d'autres équipes, données qui sont publiquement disponibles et indiquent pour environ 12 000 gènes un niveau d'activation mesuré toutes les heures sur une période de 24h.

L'idée qui a été émise est que ces données, qui sont accessibles au sein d'une base de données construite dans l'équipe, pourraient être utilisées au sein du prototype. Une réflexion sur ce thème est en cours. Mais ce domaine applicatif reste très proche de celui des TMA, avec toujours une thématique biologique. Surtout la correspondance entre la représentation virtuelle, la grille documentaire, et un objet physique, dans ce nouveau cas la matrice de la puce à ADN, reste conceptuellement évidente.

Il semblerait alors pertinent de mener un essai de passage à un autre domaine applicatif très différent, où la notion de grille n'a pas d'existence réelle, afin de cerner les difficultés d'interprétation du document de synthèse qui pourraient émerger dans ce cadre. Afin de faciliter la construction de la taxonomie de domaine, un jeu de données réduit, publiquement disponible, impliquant un nombre limité de concepts serait préférable.

Ces données étant disponibles sous un format facilement exploitable en tant que données d'exemple pour Treemap, le choix s'est porté sur les résultats des élections présidentielles américaines de 2000 et 2004. Les constructions qui ont été nécessaires pour proposer ce second domaine applicatif et le résultat d'un exemple d'étude réalisé dans ce domaine, montrant ainsi la facilité de passage à un autre domaine applicatif, vont être présentés ici.

6.2.2.2 Inclusion du nouveau domaine en pratique

Bien que l'ajout d'un second domaine applicatif ait été prévu dès le début de la conception et du développement du prototype, cet ajout implique quelques constructions et paramétrages particuliers, qui sont présentés ici.

Tout d'abord, il faut disposer de données stockées au sein d'une base de données PostGreSQL. Le fichier Treemap des résultats des élections présidentielles américaines a donc été transformé pour être chargé dans une nouvelle base très simple, constituée de quatre tables et une vue. Cette vue permet l'accès à toutes les données et servira de base au traitement de synthèse, comme le font les vues patient et histologie dans le domaine des TMA.

Une taxonomie du domaine, organisant les concepts présents dans la base, doit alors être mise en place, sur le même modèle que la taxonomie du domaine des TMA. La taxonomie du domaine d'étude, qui reste très simpliste, est présentée Fig. 6.1.

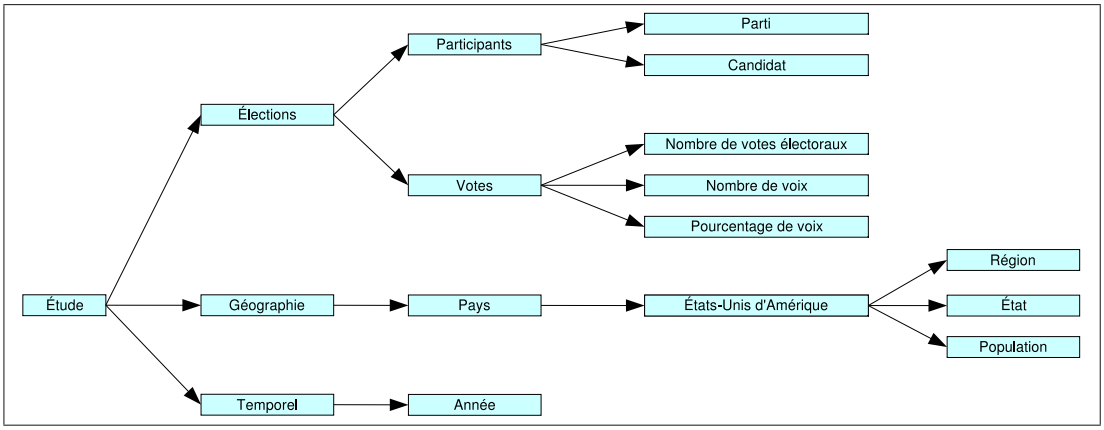


Fig. 6.1: Taxonomie simplifiée du domaine d'étude pour les élections américaines - Les différents concepts incarnés par des champs dans la base de données construite à partir du fichier de Treemap ont été organisés au sein d'une taxonomie simplifiée du domaine applicatif des élections présidentielles américaines.

Chacun des concepts s'est vu associer un fichier XML descriptif, sur le modèle présenté dans le domaine des TMA en Annexe B.

Enfin, les fichiers de règles XSLT décrivant comment construire les diverses vues annexes sur les données, par exemple des vues parti, candidat, géographie ou résultats, auraient dû être mis en place. Dans le cadre de ce simple test, cela n'a pas été le cas, majoritairement par manque de temps. Mais ceci aurait été tout à fait possible.

Quelques paramétrages supplémentaires sont enfin nécessaires. Les paramètres de connexion à la base de données doivent être ajoutés à un fichier de configuration. Le domaine des élections doit être ajouté dans la liste des domaines applicatifs possibles

qui est utilisée pour construire l'interface de saisie de requête, et en particulier la liste déroulante de choix du domaine.

La mise en place d'un nouveau domaine applicatif, si celui-ci présente des ensembles de données quantitatives ou qualitatives, reste donc relativement simple, même si elle peut s'avérer longue à réaliser quand l'espace documentaire est complexe. La question qui se pose alors est celle de l'intérêt potentiel de l'outil d'assistance à la synthèse dans un autre domaine, question abordée au prochain paragraphe.

6.2.2.3 Un exemple d'étude sur les élections américaines

Afin d'estimer l'intérêt du système d'assistance à la synthèse dans un domaine très différent du domaine des TMA, une étude de cas rapide a été menée dans le domaine des élections présidentielles américaines. Pour cette étude de cas, l'exemple de problématique qui a été choisi est la «comparaison du pourcentage de voix en fonction de l'année, du candidat et de la région, pour les élections tenues depuis 2000, trié par état».

Dans un premier temps, il s'agit de proposer une formulation de cette problématique sous forme de requête structurée. Cette formulation passe par une structuration en rôles, illustrée par le Tab. 6.1.

Tab. 6.1: Formulation informelle de l'exemple d'étude sur les élections : «comparaison du pourcentage de voix en fonction de l'année, du candidat et de la région, pour les élections tenues depuis 2000, trié par état».

Élément du modèle	*Description*
Généralités :	
- Tâche	Comparaison
- Titre	Exemple d'illustration d'une tâche de type comparaison sur les données Élections US
- Description	Comparaison du pourcentage de voix en fonction de l'année, du candidat et de la région
- Domaine	Élections US
Besoins :	
- But	Pourcentage de voix
- Critères d'inclusion	Année supérieure ou égale à 2000
- Critères de groupement	Année - Candidat - Région
- Critères de tri	État

Il apparaît ici aussi, que même dans un autre domaine d'étude, il a été proposé un exemple de problématique potentiellement intéressant qu'il a été possible d'exprimer sous forme de requête structurée. La tâche prototypique de comparaison est donc sans doute adéquate dans d'autres domaines applicatifs que celui des TMA.

La grille documentaire résultant de l'exécution de l'instance de tâche d'exemple est présentée Fig. 6.2.

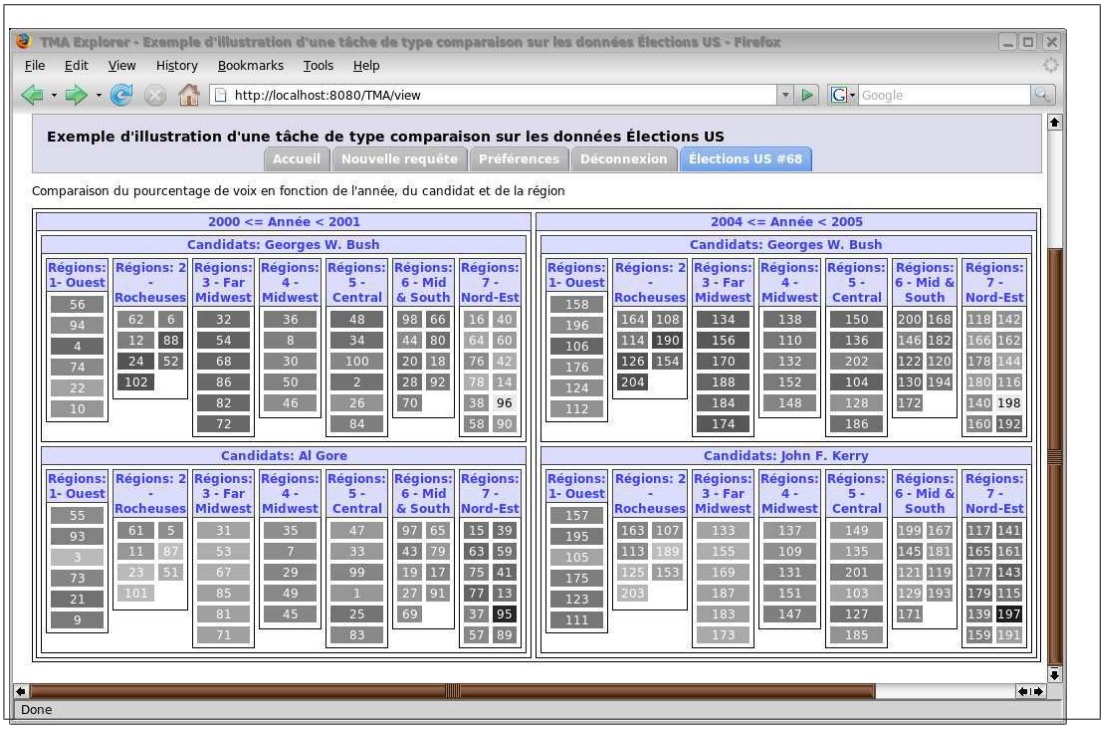


Fig. 6.2: Grille d'un document de synthèse dans le domaine applicatif des élections américaines - Cette grille correspond au résultat de l'étude portant sur la «comparaison du pourcentage de voix en fonction de l'année, du candidat et de la région».

Au sein de cette grille, les éléments sont tout d'abord organisés en deux groupes correspondant aux deux dernières années d'élections présidentielles : 2000 et 2004. Ensuite, ces groupes sont chacun divisés entre les deux candidats principaux : Georges W. Bush en haut et le candidat démocrate en bas, Al Gore à gauche pour l'année 2000 et John F. Kerry à droite pour l'année 2004. Chacun de ces groupes est ensuite divisé en fonction de la région des États-Unis considérée. Au sein de la zone correspondant à chaque région, chaque case correspond à un état dont la couleur de fond correspond au pourcentage de voix obtenu par le candidat considéré cette année là, de 0 (blanc) à 100% (noir).

Une observation plus fine de la grille fait apparaître un certain nombre de faits connus. Tout d'abord, une observation case par case, donc état par état, fait apparaître des cases plus foncées, donc correspondant à un pourcentage de voix plus élevé, en plus grand nombre pour Georges W. Bush que pour ses adversaires, et c'est bien lui qui a gagné ces deux élections.

De plus, une observation région par région révèle les tendances démocrates ou républicaines de certaines régions américaines. Ainsi, par exemple, l'ouest et le nord-est sont plutôt démocrates, avec des cases plus foncées pour Al Gore ou John F.

Kerry que pour Georges W. Bush. C'est le contraire pour des régions très traditionalistes et à tendance plus républicaine, comme les Rocheuses ou le Nord-ouest.

Il semble donc ici aussi que l'interprétation de la grille documentaire permet la révélation d'informations intéressantes, dans un domaine applicatif totalement différent. Il paraît alors légitime de considérer que la notion de synthèse telle qu'elle a été envisagée dans ma thèse pourrait être étendue à d'autres domaines applicatifs où l'appréhension d'un gros volume de données pose problème.

6.2.3 Extension de la notion de synthèse

Le modèle de synthèse proposé a été conçu comme réponse à un problème d'appréhension de données posé par des technologies telles que celle des Tissue MicroArrays. En tant qu'implantation opérationnelle de ce modèle, le prototype de système d'assistance à la synthèse se place parmi les diverses méthodes d'exploitation des données, en bout de chaîne de la technologie des TMA présentée au Paragraphe 1.3.2.

Mais cette problématique d'appréhension des données n'est pas la seule qui, dans le cadre de la technologie des TMA, pourrait relever de la problématique de synthèse. En effet, le processus de synthèse conduit à la construction d'un document dont l'élément principal est une grille organisant de manière compacte les individus pertinents dans le cadre de l'étude à réaliser. Or, la technique expérimentale des TMA présente une autre étape relevant de cette problématique : la conception du plan de fabrication du bloc TMA, en tout début de chaîne de la technologie, tel qu'illustré Fig. 6.3.

En effet, les différentes composantes de la synthèse se retrouvent dans la conception de blocs TMA :

- ⋆ Requête : de la même façon que la synthèse est guidée par une étude représentée par une requête structurée, la conception d'un bloc TMA est guidée, en fonction d'une étude à réaliser ou afin de compléter le pool de blocs TMA existants avec des blocs incluant de nouveaux patients,
- ⋆ Problème de sélection : la construction du plan de fabrication du bloc induit, comme la synthèse, une étape de sélection : la sélection des carottes de tissu à intégrer au bloc, qui passe par une sélection des blocs biopsie à représenter et une définition de coordonnées de prélèvement pour chaque bloc biopsie,
- ⋆ Problème d'organisation : la conception du bloc induit elle aussi une étape d'organisation : l'organisation des carottes à prélever au sein du ou des blocs TMA receveurs,
- ⋆ Problème de présentation : la construction du plan de fabrication requiert une étape de présentation : la mise en forme d'un document lisible par le technicien pour une construction manuelle du bloc, ou lisible par le logiciel de contrôle, pour une construction automatique.

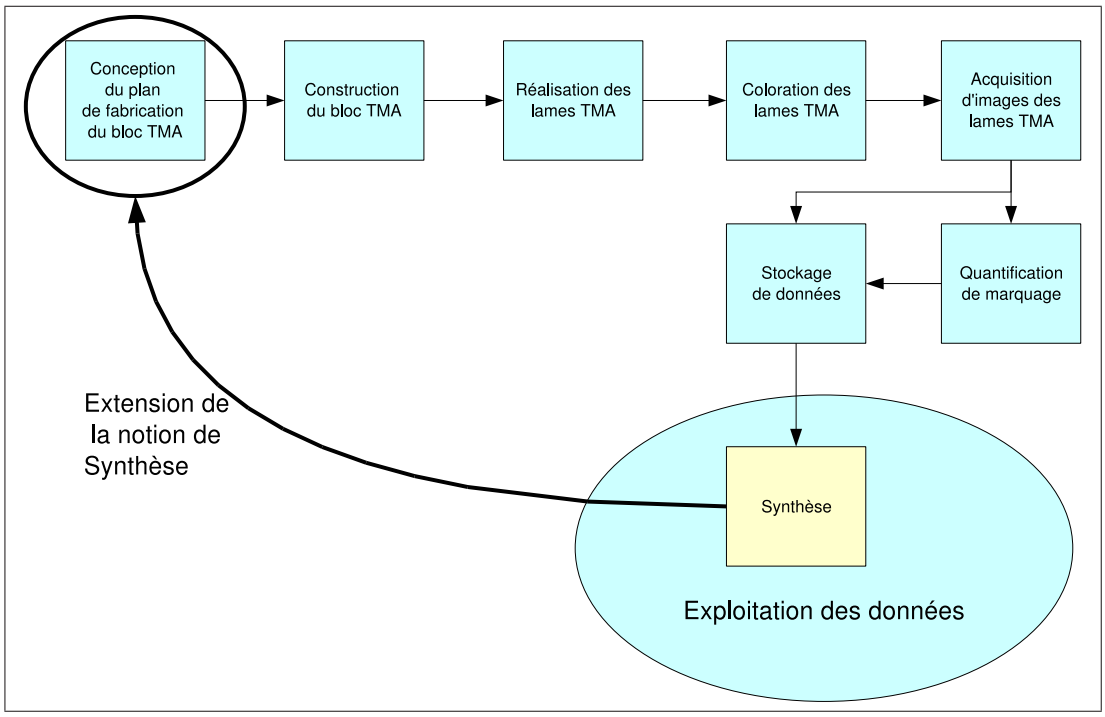


Fig. 6.3: Synthèse dans le contexte de la chaîne de la technologie des TMA - Ce schéma identifie, au sein de la chaîne de la technologie des TMA, la place originellement dévolue à la synthèse et montre où elle pourrait éventuellement intervenir..

Le concept de synthèse semble donc bien convenir pour représenter la conception d'un plan de fabrication d'un bloc TMA. Mais l'intégration de cette problématique de conception peut nécessiter quelques adaptations du modèle.

On peut déjà noter qu'il faudrait sans doute inclure de nouvelles catégories de tâches. Ces nouvelles catégories devraient permettre soit de décliner les tâches prototypiques existantes selon un point de vue conception de bloc ou exploration de données, soit de représenter des tâches de conception de blocs dans l'esprit des catégories évoquées dans la liste de [Kajdacsy-Balla et al., 2007] et rappelées au Paragraphe 3.2.3.2.

De plus, dans le cas d'une conception de bloc, les connaissances expérimentales du domaine formaliseraient l'expertise technique de l'équipe et prendraient une place prépondérante dans le processus de synthèse. Le formalisme actuellement adopté pour ces connaissances peut alors s'avérer trop limité. De plus, ainsi qu'il a été indiqué au Paragraphe 1.3.4.2, ces connaissances sont encore floues et leur formalisation pourrait s'avérer difficile.

Quels que soient les problèmes potentiellement posés par le recours à la notion de synthèse pour concevoir le plan de fabrication d'un bloc TMA, considérer cette nouvelle utilisation du paradigme de synthèse suggère son utilisation dans d'autres contextes que l'exploitation de données. En effet, la conception de blocs TMA fait

partie d'une problématique plus vaste : la conception d'expériences.

Cette problématique de conception d'expérience implique en général, en fonction d'une étude à réaliser, le choix de protocoles et matériels adéquats, en prenant en compte d'éventuelles incompatibilités entre produits ou méthodes, des considérations de temps passé ou de coût, etc. Dans ce cadre, ce ne sont plus les données qui doivent être explorées et appréhendées, mais les connaissances expérimentales, qui deviennent tout à la fois moyen pour arriver à une solution et objet de l'étude.

La notion de synthèse ne se limite alors plus à l'étude d'un seul type d'éléments, les données, mais peut s'étendre à tout champ assez vaste pour que l'appréhension de l'information, bien que formalisée, soit difficile. Les perspectives d'exploration de la notion de synthèse deviennent alors d'une richesse et d'une variété importante, qui en font un concept particulièrement intéressant.

Bibliographie

[Aamodt and Plaza, 1994] Aamodt, A. and Plaza, E. (1994). Case-Based Reasoning : Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*, 7(1) :39–59.

[Arentz and Øhrn, 2004] Arentz, W. A. and Øhrn, A. (2004). Multidimensional Visualization and Navigation in Search Results. In Negoita, M. G., Howlett, R. J., and Jain, L. C., editors, *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems – KES 2004*, volume 3213 - I of *Lecture Notes in Computer Science*, pages 620–629, Wellington, New Zealand. Springer.

[Atkinson et al., 2001] Atkinson, T., Pensy, K., Nicholas, C., Ebert, D. S., Atkinson, R., and Morris, C. J. (2001). Case study : Visualization and Information Retrieval Techniques for Network Intrusion Detection. In Ebert, D. S., Favre, J. M., and Peikert, R., editors, *Proceedings of the 3rd Joint IEEE TCVG - EUROGRAPHICS Symposium on Visualization – VisSym 2001*, Ascona, Switzerland.

[Aula, 2003] Aula, A. (2003). Query Formulation in Web Information Search. In Isaìas, P., editor, *Proceedings of the IADIS International Conference WWW/Internet 2003 – ICWI 2003*, pages 403–410, Algarve, Portugal.

[Baeza-Yates and Ribeiro-Neto, 1999] Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. ACM Press / Addison-Wesley.

[Bartlett and Toms, 2005] Bartlett, J. C. and Toms, E. G. (2005). How is Information used ? Applying task analysis to understanding information use. In Vaughan,

L., editor, *Proceedings of the 35th Annual Conference of the Canadian Association for Information Science*, London, ON, Canada.

[Bates, 1989] Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5) :407–424.

[Belkin et al., 2001] Belkin, N. J., Cool, C., Kelly, D., Lin, S.-J., Park, S. Y., Perez-Carballo, J., and Sikora, C. (2001). Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing & Management*, 37(3) :403–434.

[Belkin et al., 1982] Belkin, N. J., Oddy, R. N., and Brooks, H. M. (1982). ASK For Information Retrieval : Part I. Background and Theory. *Journal of Documentation*, 38(2) :61–71.

[Berman et al., 2003] Berman, J. J., Edgerton, M. E., and Friedman, B. A. (2003). The tissue microarray data exchange specification : a community-based, open source tool for sharing tissue microarray data. *BMC Medical Informatics and Decision Making*, 3 :5–13.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*.

[Besse et al., 2001] Besse, P., Gall, C. L., Raimbault, N., and Sarpy, S. (2001). Data mining et statistique. *Journal de la Société Française de Statistique*, 142(1) :5–36.

[Blake and Pratt, 2002] Blake, C. and Pratt, W. (2002). Collaborative Information Synthesis. In Toms, E., editor, *Proceedings of the 65th Annual Meeting of the American Society for Information Science and Technology – ASIST 2002*, Philadelphia, PA, USA.

[Bloesch and Halpin, 1997] Bloesch, A. C. and Halpin, T. A. (1997). Conceptual Queries using ConQuer-II. In Embley, D. W. and Goldstein, R. C., editors, *Proceedings of the 16th International Conference on Conceptual Modeling – ER '97*, volume 1331 of *Lecture Notes in Computer Science*, pages 113–126, Los Angeles, CA, USA. Springer.

[Bontas et al., 2004] Bontas, E. P., Tietz, S., and Schrader, T. (2004). Experiences Using Semantic Web Technologies to Realize an Information Retrieval System for Pathology. In Tolksdorf, R. and Eckstein, R., editors, *Berliner XML Tage 2004*, pages 82–93, Berlin, Germany. XML-Clearinghouse.

[Borlund, 2003] Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10) :913–925.

[Bova et al., 2001] Bova, G. S., Parmigiani, G., Epstein, J. I., Wheeler, T., Mucci, N. R., and Rubin, M. A. (2001). Web-based tissue microarray image data analysis : initial validation testing through prostate cancer Gleason grading. *Human Pathology*, 32(4) :417–427.

[Bret, 2006] Bret, D. (2006). Mémoire d'ingénieur - TMA-Explorer : Plate-forme d'assistance pour l'utilisation de la technique des puces à tissus (Tissue Micro Arrays TMA). Master's thesis, Conservatoire National des Arts et Métiers - Centre Régional Rhône-Alpes - Centre d'enseignement de Grenoble.

[Brown, 1999] Brown, C. M. (1999). Information seeking behavior of scientists in the electronic information age : astronomers, chemists, mathematicians, and physicists. *Journal of the American Society for Information Science and Technology*, 50(10) :929–943.

[Brusilovsky, 1996] Brusilovsky, P. (1996). Methods and Techniques of Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 6(2-3) :87–129.

[Brusilovsky, 2003] Brusilovsky, P. (2003). From Adaptive Hypermedia to the Adaptive Web. In Szwillus, G. and Ziegler, J., editors, *Mensch & Computer 2003 : Interaktion in Bewegung*, Stuttgart, Germany. Teubner.

[Bubendorf et al., 2001] Bubendorf, L., Nocito, A., Moch, H., and Sauter, G. (2001). Tissue microarray (TMA) technology : miniaturized pathology archives for high-throughput in situ studies. *Journal of Pathology*, 195(1) :72–79.

[Bush, 1945] Bush, V. (1945). As We May Think. *The Atlantic Monthly*.

[Camp et al., 2000] Camp, R. L., Charette, L. A., and Rimm, D. L. (2000). Validation of tissue microarray technology in breast carcinoma. *Laboratory Investigation*, 80(12) :1943–1949.

[Campos et al., 2006] Campos, R., Dias, G., and Nunes, C. (2006). WISE : Hierarchical Soft Clustering of Web Page Search Results Based on Web Content Mining Techniques. In *Proceedings of the 2006 IEEE / WIC / ACM International Conference on Web Intelligence – WI 2006*, pages 301–304, Hong Kong, China. IEEE Computer Society.

[Card and Mackinlay, 1997] Card, S. K. and Mackinlay, J. (1997). The structure of the information visualization design space. In Moorhead, R. and Johnston, N., editors, *Proceedings of the 1997 IEEE Symposium on Information Visualization – InfoVis '97*, pages 92–99, Phoenix, AZ, USA.

[Carey et al., 2003] Carey, M., Heesch, D. C., and Rüger, S. M. (2003). Info Navigator : A Visualization Tool for Document Searching and Browsing. In *Proceedings of the 9th International Conference Distributed Multimedia Systems – DMS'03*, pages 23–28, Miami, FL, USA.

[Catarci et al., 2004] Catarci, T., Dongilli, P., Mascio, T. D., Franconi, E., Santucci, G., and Tessaris, S. (2004). An Ontology Based Visual Tool for Query Formulation Support. In de Mántaras, R. L. and Saitta, L., editors, *Proceedings of the 16th Eureopean Conference on Artificial Intelligence – ECAI'2004, including Prestigious Applicants of Intelligent Systems – PAIS 2004*, pages 308–312, Valencia, Spain. IOS Press.

[Chen, 2006] Chen, J. (2006). Visual Inquiry of Spatio-Temporal Multivariate Patterns. In *Doctoral Colloquium at the IEEE Symposium on Visual Analytics Science and Technology – VAST 2006*, Baltimore, MD, USA.

[Chen et al., 2004] Chen, W., Reiss, M., and Foran, D. J. (2004). A prototype for unsupervised analysis of tissue microarrays for cancer research and diagnostics. *IEEE Transactions on Information Technology in Biomedecine*, 8(2) :89–96.

[Chevallet et al., 2005] Chevallet, J.-P., Lim, J.-H., and Vasudha, R. (2005). SnapToTell : A Singapore Image Test Bed for Ubiquitous Information Access from Camera. In Losada, D. E. and Fernández-Luna, J. M., editors, *Proceedings of the 27th European Conference on IR Research, Advances in Information Retrieval – ECIR 2005*, volume 3408 of *Lecture Notes in Computer Science*, pages 530–532, Santiago de Compostela, Spain. Springer.

[Choo et al., 1999] Choo, C. W., Detlor, B., and Turnbull, D. (1999). Information Seeking on the Web - An Integrated Model of Browsing and Searching. In Woods, L., editor, *Proceedings of the 62nd Annual Meeting of the American Society for Information Science – ASIS'99*, volume 36, pages 3–16, Washington, DC, USA.

[Cigarrán et al., 2005] Cigarrán, J. M., Peñas, A., Gonzalo, J., and Verdejo, F. (2005). Evaluating Hierarchical Clustering of Search Results. In Consens, M. P. and Navarro, G., editors, *Proceedings of the 12th International Conference on String Processing and Information Retrieval – SPIRE 2005*, volume 3772 of *Lecture Notes in Computer Science*, pages 49–54, Buenos Aires, Argentina. Springer.

[Cockburn and McKenzie, 2002] Cockburn, A. and McKenzie, B. (2002). Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments. In Wixon, D., editor, *Proceedings of the SIGCHI conference on Human factors in computing systems : Changing our world, changing ourselves – CHI'02*, pages 203–210, Minneapolis, MN, USA. ACM Press.

[Cohen and Casanova, 2001] Cohen, J. and Casanova, X. (2001). L'écran efficace : Une approche cognitive des objets graphiques. *Documentaliste*, 38(5-6) :272–283.

[Cohen and Brodlie, 2004] Cohen, M. and Brodlie, K. (2004). Focus and Context for Volume Visualization. In Lever, P. G., editor, *Proceedings of the Theory and Practice of Computer Graphics 2004 Conference – TPCG'04*, pages 32–39, Bournemouth, United Kingdom.

[Conway et al., 2006] Conway, C. M., O'Shea, D., O'Brien, S., Lawler, D. K., Dodrill, G. D., O'Grady, A., Barrett, H., Gulmann, C., O'Driscoll, L., Gallagher, W. M., Kay, E. W., and O'Shea, D. G. (2006). The development and validation of the Virtual Tissue Matrix, a software application that facilitates the review of tissue microarrays on line. *BMC Bioinformatics*, 7 :256–267.

[Cool and Spink, 2001] Cool, C. and Spink, A. (2001). Issues of context in information retrieval (IR) : an introduction to the special issue. *Information Processing & Management*, 38(5) :605–611.

[Corkill, 1991] Corkill, D. (1991). Blackboard Systems. *AI Expert*, 6(9) :40–47.

[Cortés et al., 2000] Cortés, U., López-Navidad, A., Vázquez-Salceda, J., Vázquez, A., Busquets, D., Nicolás, M., Lopes, S., Vázquez, F., and Caballero, F. (2000). Carrel : An Agent Mediated Institution for the Exchange of Human Tissues among Hospitals for Transplantation. Technical Report LSI-00-33-R, Software Department. UPC.

[Côté et al., 2005] Côté, M.-A., Suryn, W., Laporte, C. Y., and Martin, R. A. (2005). The Evolution Path for Industrial Software Quality Evaluation Methods Applying ISO/IEC 9126 : 2001 Quality Model : Example of MITRE's SQAE Method. *Software Quality Control*, 13(1) :17–30.

[Crubézy and Musen, 2004] Crubézy, M. and Musen, M. A. (2004). Ontologies in Support of Problem Solving. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 321–342. Springer.

[Cugini et al., 2000] Cugini, J. V., Laskowski, S., and Sebrechts, M. M. (2000). Design of 3D visualization of search results : evolution and evaluation. In Erbacher, R. F., Chen, P. C., Roberts, J. C., and Wittenbrink, C. M., editors, *Visual Data Exploration and Analysis VII*, volume 3960 of *Proceedings of SPIE*, pages 198–210, San Jose, CA, USA.

[Curbera et al., 2002] Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., and Weerawarana, S. (2002). Unraveling the Web Services Web : An Introduction to SOAP, WSDL, and UDDI. *IEEE Internet Computing*, 06(2) :86–93.

[Demichelis, 2005] Demichelis, F. (2005). *On information organization and information extraction for the study of gene expressions by tissue microarray technique.* PhD thesis, DIT - University of Trento.

[Dervin, 1983] Dervin, B. (1983). An overview of sense-making research : concepts, methods, and results to date. In *International Communication Association Annual Meeting*, Dallas, TX, USA.

[Diaz et al., 2004] Diaz, L. K., Gupta, R., Kidwai, N., Sneige, N., and Wiley, E. L. (2004). The use of TMA for interlaboratory validation of FISH testing for detection of HER2 gene amplification in breast cancer. *Journal of Histochemistry and Cytochemistry*, 52(4) :501–507.

[Ding et al., 2002] Ding, Y., Fensel, D., Klein, M. C. A., and Omelayenko, B. (2002). The semantic web : yet another hip ? *Data & Knowledge Engineering*, 41(2–3) :205–227.

[Dongilli et al., 2004] Dongilli, P., Franconi, E., and Tessaris, S. (2004). Semantics Driven Support for Query Formulation. In Haarslev, V. and Möller, R., editors, *Proceedings of the 2004 International Workshop on Description Logics – DL2004*, volume 104 of *CEUR Workshop Proceedings*, Whistler, BC, Canada. CEUR-WS.org.

[Driessen et al., 2006] Driessen, S., Jacobs, J., and Huijsen, W.-O. (2006). Combining query and visual search for knowledge mapping. In *Proceedings of the 10th International Conference on Information Visualization – IV '06*, pages 216–224, London, United Kingdom. IEEE Computer Society.

[Eick, 2005] Eick, S. G. (2005). Information Visualization at 10. *IEEE Computer Graphics and Applications*, 25(1) :12–14.

[Ellis, 1989] Ellis, D. (1989). A behavioral approach to information retrieval system design. *Journal of Documentation*, 45(3) :171–212.

[Faith et al., 2004] Faith, D. A., Isaacs, W. B., Morgan, J. D., Fedor, H. L., Hicks, J. L., Mangold, L. A., Walsh, P. C., Partin, A. W., Platz, E. A., Luo, J., and Marzo, A. M. D. (2004). Trefoil factor 3 overexpression in prostatic carcinoma : prognostic importance using tissue microarrays. *Prostate*, 61(3) :215–227.

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3) :37–54.

[Fejzo and Slamon, 2001] Fejzo, M. S. and Slamon, D. J. (2001). Frozen tumor tissue microarray technology for analysis of tumor RNA, DNA, and proteins. *American Journal of Pathology*, 159(5) :1645–1650.

[Fensel et al., 2003] Fensel, D., Motta, E., van Harmelen, F., Benjamins, V. R., Crubézy, M., Decker, S., Gaspari, M., Groenboom, R., Grosso, W. E., Musen, M. A., Plaza, E., Schreiber, G., Studer, R., and Wielinga, B. J. (2003). The Unified Problem-Solving Method Development Language UPML. *Knowledge and Information Systems*, 5(1) :83–131.

[Fensel et al., 2001] Fensel, D., van Harmelen, F., Horrocks, I., McGuinness, D. L., and Patel-Schneider, P. F. (2001). OIL : An Ontology Infrastructure for the Semantic Web. *IEEE Intelligent Systems*, 16(2) :38–45.

[Ferragina and Gulli, 2005] Ferragina, P. and Gulli, A. (2005). A personalized search engine based on web-snippet hierarchical clustering. In *Special interest tracks and posters of the 14th international conference on World Wide Web – WWW'05*, pages 801–810, Chiba, Japan. ACM Press.

[Fox et al., 2006] Fox, E. A., Neves, F. D., Yu, X., Shen, R., Kim, S., and Fan, W. (2006). Exploring the computing literature with visualization and stepping stones & pathways. *Communications of the ACM*, 49(4) :52–58.

[Frawley et al., 1992] Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1992). Knowledge Discovery in Databases : An Overview. *AI Magazine*, 13(3) :57–70.

[Friedman, 1997] Friedman, J. (1997). Data mining and statistics : What's the connection ? In Scott, D. W., editor, *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics – Interface'97*, Houston, TX, USA.

[Friendly and Kwan, 2003] Friendly, M. and Kwan, E. (2003). Effect ordering for data displays. *Computational statistics & data analysis*, 43(4) :509–539.

[Furtado et al., 2003] Furtado, E., Furtado, J., Limbourg, Q., Vanderdonckt, J., Silva, W., Rodrigues, D., and Taddeo, L. (2003). A Layered Approach for Designing Multiple User Interfaces from Task and Domain Models. In Jacko, J. and Stephanidis, C., editors, *Proceedings of the 10th International Conference on Human-Computer Interaction HCI International'2003*, volume 1, pages 103–107, Heraklion, Greece. Lawrence Erlbaum Associates, Mahwah.

[Garlatti and Iksal, 2003] Garlatti, S. and Iksal, S. (2003). A semantic Web approach for adaptive hypermedia. In Bra, P. D., Davis, H., Kay, J., and Schraefel, M., editors, *Workshop on Adaptive Hypermedia and Adaptive Web-Based Systems – AH 2003*, Budapest, Hungary – Johnstown, PA, USA – Nottingham, United Kingdom.

[Gediga et al., 1999] Gediga, G., Hamborg, K.-C., and Düntsch, I. (1999). The IsoMetrics usability inventory : An operationalisation of ISO 9241/10. *Behavior and Information Technology*, 18 :151–164.

[Glover et al., 2002] Glover, E. J., Tsioutsiouliklis, K., Lawrence, S., Pennock, D. M., and Flake, G. W. (2002). Using web structure for classifying and describing web pages. In *Proceedings of the Eleventh International World Wide Web Conference – WWW2002*, pages 562–569, Honolulu, HI, USA. ACM.

[Goebel and Gruenwald, 1999] Goebel, M. and Gruenwald, L. (1999). A Survey of Data Mining and Knowledge Discovery Software Tools. *SIGKDD Explorations*, 1(1) :20–33.

[Goldschmidt, 1986] Goldschmidt, P. G. (1986). Information synthesis : a practical guide. *Health Services Research*, 21(2 Pt 1) :215–237.

[Groß-Hardt, 2002] Groß-Hardt, M. (2002). Concept based querying of semistructured data. In Tolksdorf, R. and Eckstein, R., editors, *Proceedings zum Workshop XML Technologien für das Semantic Web, XSW 2002*, volume 14 of *Lectures Notes in Informatics*, pages 79–92, Berlin, Germany. GI.

[Großjohann et al., 2002] Großjohann, K., Fuhr, N., Effing, D., and Kriewel, S. (2002). A User Interface for XML Document Retrieval. In Schubert, S. E., Reusch, B., and Jesse, N., editors, *Informatik bewegt : Informatik 2002 - 32. Jahrestagung der Gesellschaft für Informatik e.v. (GI)*, volume 19 of *Lectures Notes in Informatics*, Dortmund, Germany. Springer.

[Gruber, 1993] Gruber, T. R. (1993). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In Guarino, N. and Poli, R., editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Padova, Italy. Kluwer Academic Publishers.

[Guarino, 1998] Guarino, N. (1998). Formal Ontology and Information Systems. In Guarino, N., editor, *Proceedings of the 1st International Conference on Formal*

Ontologies in Information Systems – FOIS'98, pages 3–15, Trento, Italy. IOS Press.

[Hearst, 2006] Hearst, M. A. (2006). Clustering versus faceted categories for information exploration. *Communications of the ACM*, 49(4) :59–61.

[Heidorn and Cui, 2000] Heidorn, P. B. and Cui, H. (2000). The Interaction of Result Set Display Dimensionality and Cognitive Factors in Information Retrieval Systems. In *Proceedings of the 63rd Annual Meeting of the American Society for Information Science – ASIS 2000*, pages 258–270, Chicago, IL, USA.

[Hucka et al., 2003] Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Novère, N. L., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J., and Forum, S. B. M. L. (2003). The systems biology markup language (SBML) : a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4) :524–531.

[Hunter et al., 2004] Hunter, J., Falkovych, K., and Little, S. (2004). Next Generation Search Interfaces - Interactive Data Exploration and Hypothesis Formulation. In Heery, R. and Lyon, L., editors, *Proceedings of the 8th European Conference on Research and Advanced Technology for Digital Libraries – ECDL 2004*, volume 3232 of *Lecture Notes in Computer Science*, pages 86–98, Bath, United Kingdom.

[Ingwersen et al., 2005] Ingwersen, P., Järvelin, K., and Belkin, N. J., editors (2005). *Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context - IRiX 2005*, Salvador, Brazil. ACM.

[Ivory et al., 2001] Ivory, M. Y., Sinha, R. R., and Hearst, M. A. (2001). Empirically validated web page design metrics. In *Proceedings of the SIGCHI conference on Human factors in computing systems – CHI'01*, pages 53–60, Seattle, WA, USA. ACM Press.

[Jansen et al., 2000] Jansen, B. J., Spink, A., and Saracevic, T. (2000). Real life, real users, and real needs : a study and analysis of user queries on the web. *Informantion Processing & Management*, 36(2) :207–227.

[Järvelin and Ingwersen, 2004] Järvelin, K. and Ingwersen, P. (2004). Information seeking research needs extension toward tasks and technology. *Information Research*, 10(1).

[Johnson et al., 2003] Johnson, F., Griffiths, J., and Hartley, R. (2003). Task dimensions of user evaluations of information retrieval systems. *Information Research*, 8(4).

[Jourdan et al., 2003] Jourdan, F., Sebbagh, N., Comperat, E., Mourra, N., Flahault, A., Olschwang, S., Duval, A., Hamelin, R., and Flejou, J.-F. (2003). Tissue microarray technology : validation in colorectal carcinoma and analysis of p53, hMLH1, and hMSH2 immunohistochemical expression. *Virchows Archiv*, 443(2) :115–121.

[Kajdacsy-Balla et al., 2007] Kajdacsy-Balla, A., Geynisman, J. M., Macias, V., Setty, S., Nanaji, N. M., Berman, J. J., Dobbin, K., Melamed, J., Kong, X., Bosland, M., Orenstein, J., Bayerl, J., Becich, M. J., Dhir, R., Datta, M. W., and Resource, C. P. C. T. (2007). Practical aspects of planning, building, and interpreting tissue microarrays : the Cooperative Prostate Cancer Tissue Resource experience. *Journal of Molecular Histology*, 38(2) :113–121.

[Kallioniemi et al., 2001] Kallioniemi, O. P., Wagner, U., Kononen, J., and Sauter, G. (2001). Tissue microarray technology for high-throughput molecular profiling of cancer. *Human Molecular Genetics*, 10(7) :657–662.

[Keim, 2002] Keim, D. A. (2002). Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1) :1–8.

[Kim et al., 2002] Kim, K., Carroll, J. M., and Rosson, M. B. (2002). An Empirical Study of Web Personalization Assistants : Supporting End-Users in Web Information Systems. In *2002 IEEE CS International Symposium on Human-Centric Computing Languages and Environments – HCC 2002*, pages 60–62, Arlington, VA, USA. IEEE Computer Society.

[Kim et al., 2005] Kim, R., Demichelis, F., Tang, J., Riva, A., Shen, R., Gibbs, D. F., Mahavishno, V., Chinnaiyan, A. M., and Rubin, M. A. (2005). Internet-based Profiler system as integrative framework to support translational research. *BMC Bioinformatics*, 6 :304–314.

[Kobsa, 2001] Kobsa, A. (2001). Generic User Modeling Systems. *User Modeling and User-Adapted Interaction*, 11(1) :49–63.

[Koenemann and Belkin, 1996] Koenemann, J. and Belkin, N. J. (1996). A case for interaction : a study of interactive information retrieval behavior and effectiveness. In Bilger, R., Guest, S., and Tauber, M. J., editors, *Proceedings of the SIGCHI conference on Human factors in computing systems : common ground – CHI96*, pages 205–212, Vancouver, BC, Canada.

[Kohonen et al., 1996] Kohonen, T., Oja, E., Simula, O., Visa, A., and Kangas, J. (1996). Engineering applications of the self-organizing map. *Proceedings of the IEEE*, 84(10) :1358–1384.

[Kononen et al., 1998] Kononen, J., Bubendorf, L., Kallioniemi, A., Bärlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M. J., Sauter, G., and Kallioniemi, O. P. (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medecine*, 4(7) :844–847.

[Koutri et al., 2002] Koutri, M., Daskalaki, S., and Avouris, N. (2002). Adaptive interaction with web sites : an overview of methods and techniques. In Groumpos, P. P., editor, *Proceedings on the 4th International Workshop on Computer Science and Information Technologies – CSIT 2002*, Patras, Greece.

[Kroeker, 2004] Kroeker, K. L. (2004). Seeing data : new methods for understanding information. *IEEE Computer Graphics and Applications*, 24(3) :6–12.

[Kuhlthau, 1991] Kuhlthau, C. C. (1991). Inside the search process : Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5) :361–371.

[Kules and Shneiderman, 2004] Kules, B. and Shneiderman, B. (2004). Categorized graphical overviews for web search results : An exploratory study using U. S. government agencies as a meaningful and stable structure. In McCoy, S. and Hess, T., editors, *Proceedings of the 3rd Pre-ICIS Annual Workshop on HCI Research in MIS – HCI/MIS'04*, pages 20–24, Washington, DC, USA.

[Kunz, 2003] Kunz, C. (2003). SERGIO - An Interface for Context Driven Knowledge Retrieval. In Cunningham, P., Cunningham, M., and Fatelnig, P., editors, *Proceedings of eChallenges - Building the Knowledge Economy : Issues, Applications, Case Studies*, Bologna, Italy. IOS Press.

[Larkin and Simon, 1987] Larkin, J. H. and Simon, H. A. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, 11(1) :65–100.

[Laublet et al., 2002] Laublet, P., Reynaud, C., and Charlet, J. (2002). Sur quelques aspects du web sémantique. In *Assises du GDR 13*, Nancy, France. Cépadues.

[Lee and Lee, 2005] Lee, K. and Lee, S. J. (2005). A Quantitative Software Quality Evaluation Model for the Artifacts of Component Based Development. In Mandoiu, I. and Zelikovsky, A., editors, *Proceedings of the Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks – SNPD-SAWN'05*, pages 20–25, Towson University, MD, USA.

[Liu et al., 2005] Liu, C. L., Montgomery, K. D., Natkunam, Y., West, R. B., Nielsen, T. O., Cheang, M. C. U., Turbin, D. A., Marinelli, R. J., van de Rijn, M., and Higgins, J. P. T. (2005). TMA-Combiner, a simple software tool to permit analysis of replicate cores on tissue microarrays. *Modern Pathology*, 18(12) :1641–1648.

[Liu et al., 2002a] Liu, C. L., Prapong, W., Natkunam, Y., Alizadeh, A., Montgomery, K., Gilks, C. B., and van de Rijn, M. (2002a). Software tools for high-throughput analysis and archiving of immunohistochemistry staining data obtained with tissue microarrays. *American Journal of Pathology*, 161(5) :1557–1565.

[Liu et al., 2002b] Liu, H., Lieberman, H., and Selker, T. (2002b). GOOSE : A Goal-Oriented Search Engine with Commonsense. In Bra, P. D., Brusilovsky, P., and Conejo, R., editors, *Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems – AH 2002*, volume 2347 of *Lecture Notes in Computer Science*, pages 253–263, Malaga, Spain. Springer.

[Lloyd et al., 2004] Lloyd, C. M., Halstead, M. D. B., and Nielsen, P. F. (2004). CellML : its future, present and past. *Progress in Biophysics & Molecular Biology*, 85(2–3) :433–450.

[Lugli et al., 2004] Lugli, A., Tornillo, L., Mirlacher, M., Bundi, M., Sauter, G., and Terracciano, L. M. (2004). Hepatocyte paraffin 1 expression in human normal and neoplastic tissues : tissue microarray analysis on 3,940 tissue samples. *American Journal of Clinical Pathology*, 122(5) :721–727.

[Mäkelä et al., 2005] Mäkelä, E., Hyvönen, E., and Sidoroff, T. (2005). View-Based User Interfaces for Information Retrieval on the Semantic Web. In Bernstein, A., Androutsopoulos, I., Degler, D., and McBride, B., editors, *Proceedings of the ISWC-2005 Workshop on End User Semantic Web Interaction*, Galway, Ireland.

[Manley et al., 2001] Manley, S., Mucci, N. R., Marzo, A. M. D., and Rubin, M. A. (2001). Relational database structure to manage high-density tissue microarray data and images for pathology studies focusing on clinical outcome : the prostate specialized program of research excellence model. *American Journal of Patholoy*, 159(3) :837–843.

[McIlraith et al., 2001] McIlraith, S. A., Son, T. C., and Zeng, H. (2001). Semantic Web Services. *IEEE Intelligent Systems*, 16(2) :46–53.

[Meho and Tibbo, 2003] Meho, L. I. and Tibbo, H. R. (2003). Modeling the information-seeking behavior of social scientists : Ellis's study revisited. *Journal of the American Society for Information Science and Technology*, 54(6) :570–587.

[Mich et al., 2003] Mich, L., Franch, M., and Gaio, L. (2003). Evaluating and Designing Web Site Quality. *IEEE Multimedia*, 10(1) :34–43.

[Mizzaro, 1997] Mizzaro, S. (1997). Relevance : The Whole History. *Journal of the American Society for Information Science*, 48(9) :810–832.

[Mizzaro, 1998] Mizzaro, S. (1998). How many relevances in information retrieval ? *Interacting with Computers*, 10(3) :303–320.

[Mußler and Reiterer, 2001] Mußler, G. and Reiterer, H. (2001). Visual Information Retrieval for the WWW. In Smith, M. J., Salvendy, G., Harris, D., and Koubek, R. J., editors, *Proceedings of the 2001 HCI International Conference, Usability Evaluation and Interface Design*, volume 1, pages 1150–1154, New Orleans, LA, USA. Lawrence Erlbaum.

[Ngu and Wu, 1997] Ngu, D. S. W. and Wu, X. (1997). SiteHelper : a localized agent that helps incremental exploration of the World Wide Web. *Computer*

Networks and ISDN Systems, 29(8) :1249–1255.

[Nguyen et al., 2006] Nguyen, A.-T., Denos, N., and Berrut, C. (2006). Modèle d'espaces de communautés basé sur la théorie des ensembles d'approximation dans un système de filtrage hybride. In Savoy, J., editor, *Conférence Francophone en Recherche d'Information et Applications – CORIA 2006*, pages 303–241, Lyon, France.

[Noirhomme-Fraiture and Rouard, 1997] Noirhomme-Fraiture, M. and Rouard, M. (1997). Zoom Star : a solution to complex statistical object representation. In Howard, S., Hammond, J., and Lindgaard, G., editors, *Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction – INTERACT'97*, volume 96 of *IFIP Conference Proceedings*, pages 100–101, Sydney, Australia.

[Noy et al., 2001] Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R. W., and Musen, M. A. (2001). Creating Semantic Web Contents with Protégé-2000. *IEEE Intelligent Systems*, 16(2) :60–71.

[Olsina et al., 2001] Olsina, L., Lafuente, G., and Rossi, G. (2001). *Web Engineering - Managing Diversity and Complexity of Web Application Development*, volume 2016 of *Lecture Notes In Computer Science*, chapter Specifying Quality Characteristics and Attributes for Websites, pages 266–278. Springer-Verlag, London, United Kingdom.

[Packeisen et al., 2002] Packeisen, J., Buerger, H., Krech, R., and Boecker, W. (2002). Tissue microarrays : a new approach for quality control in immunohistochemistry. *Journal of Clinical Pathology*, 55(8) :613–615.

[Patel et al., 2006] Patel, A. A., Gilbertson, J. R., Parwani, A. V., Dhir, R., Datta, M. W., Gupta, R., Berman, J. J., Melamed, J., Kajdacsy-Balla, A., Orenstein, J., Becich, M. J., and Resource, C. P. C. T. (2006). An informatics model for tissue banks – lessons learned from the Cooperative Prostate Cancer Tissue Resource. *BMC Cancer*, 6 :120–138.

[Rabinovich et al., 2006] Rabinovich, A., Krajewski, S., Krajewska, M., Shabaik, A., Hewitt, S. M., Belongie, S., Reed, J. C., and Price, J. H. (2006). Framework for parsing, visualizing and scoring tissue microarray images. *IEEE Transactions on Information Technology in Biomedecine*, 10(2) :209–219.

[Rao and Su, 2004] Rao, J. and Su, X. (2004). A Survey of Automated Web Service Composition Methods. In Cardoso, J. and Sheth, A. P., editors, *Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition - SWSWPC 2004 - Revised Selected Papers*, volume 3387 of *Lecture Notes in Computer Science*, pages 43–54, San Diego, CA, USA. Springer.

[Reix, 2003] Reix, R. (2003). Evaluation des sites Web : Nouvelles pratiques Anciennes théories. In *Actes du 8ème Colloque de l'AIM (Association Information et Management)*, Grenoble, France.

[Ricordel and Demazeau, 2000] Ricordel, P.-M. and Demazeau, Y. (2000). From Analysis to Deployment : A Multi-agent Platform Survey. In Omicini, A., Tolksdorf, R., and Zambonelli, F., editors, *Revised Papers from the First International Workshop on Engineering Societies in the Agent World – ESAW 2000*, volume 1972 of *Lecture Notes in Computer Science*, pages 93–105, Berlin, Germany. Springer.

[Rosse and Mejino, 2003] Rosse, C. and Mejino, J. L. V. (2003). A reference ontology for biomedical informatics : the Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36(6) :478–500.

[Rubin et al., 2002] Rubin, M. A., Dunn, R., Strawderman, M., and Pienta, K. J. (2002). Tissue microarray sampling strategy for prostate cancer biomarker analysis. *American Journal of Surgical Pathology*, 26(3) :312–319.

[Ruthven and Lalmas, 2003] Ruthven, I. and Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(2) :95–145.

[Sacile, 1995] Sacile, R. (1995). Using CommonKADS to Build an Expertise Model for Breast Cancer Prognosis and Therapy. Rapport de recherche RR-2737, INRIA, Sophia Antipolis, France.

[Salton and Buckley, 1990] Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4) :288–297.

[Saracevic, 1975] Saracevic, T. (1975). Relevance : a review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6) :321–343.

[Saracevic, 1999] Saracevic, T. (1999). Information Science. *Journal of the American Society for Information Science*, 50(12) :1051–1063.

[Schmidt et al., 2004] Schmidt, C., Parashar, M., Chen, W., and Foran, D. J. (2004). Engineering a Peer-to-Peer Collaboratory for Tissue Microarray Research. In *Proceedings of the Second International Workshop on Challenges of Large Applications in Distributed Environments – CLADE'04*, pages 64–73, Honolulu, Hawaii, USA.

[Sebrechts et al., 1999] Sebrechts, M. M., Cugini, J. V., Laskowski, S. J., Vasilakis, J., and Miller, M. S. (1999). Visualization of search results : a comparative evaluation of text, 2D, and 3D interfaces. In Gey, F., Hearst, M., and Tong, R., editors, *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR '99*, pages 3–10, Berkeley, CA, USA. ACM Press.

[Shaknovich et al., 2003] Shaknovich, R., Celestine, A., Yang, L., and Cattoretti, G. (2003). Novel relational database for tissue microarray analysis. *Archives of Pathology and Laboratory Medicine*, 127(4) :492–494.

[Sharma-Oates et al., 2005] Sharma-Oates, A., Quirke, P., and Westhead, D. R. (2005). TmaDB : a repository for tissue microarray data. *BMC Bioinformatics*, 6 :218–225.

[Shneiderman, 1992] Shneiderman, B. (1992). Tree Visualization with Treemaps : 2-d Space-Filling Approach. *ACM Transactions on Graphics*, 11(1) :92–99.

[Shneiderman, 2002] Shneiderman, B. (2002). Inventing discovery tools : combining information visualization with data mining. *Information Visualization*, 1(1) :5–12.

[Simony-Lafontaine, 2007] Simony-Lafontaine, J. (2007). *Bouleversements architecturaux induits dans la muqueuse colique normale et tumorale par la transformation maligne et la progression tumorale : approche morphologique.* PhD thesis, Université Joseph Fourier, Grenoble, France.

[Skupin and Fabrikant, 2003] Skupin, A. and Fabrikant, S. (2003). Spatialization Methods : A Cartographic Research Agenda for Non-Geographic Information Visualization. *Cartography and Geographic Information Science*, 30(2) :99–119.

[Soenksen, 2003] Soenksen, D. G. (2003). Automated Microscopic Inspection of Tissue Microarrays Using Virtual Microscopy. *Genomics & Proteomics Technology*, pages 28–31.

[Soshnikov, 2002] Soshnikov, D. (2002). An Architecture of Distributed Frame Hierarchy for Knowledge Sharing and Reuse in Computer Networks. In Zakharevich, V. G. and Taratoukhine, V., editors, *Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems – ICAIS'02*, pages 115–119, Divnomorskoe, Russia. IEEE Computer Society.

[Spink et al., 2002] Spink, A., Wilson, T. D., Ford, N., Foster, A., and Ellis, D. (2002). Information-seeking and mediated searching. Part 1. Theoretical framework and research design. *Journal of the American Society for Information Science and Technology*, 53(9) :695–703.

[Stuckenschmidt et al., 2004] Stuckenschmidt, H., de Waard, A., Bhogal, R., Fluit, C., Kampman, A., van Buel, J., van Mulligen, E. M., Broekstra, J., Crowlesmith, I., van Harmelen, F., and Scerri, T. (2004). A Topic-Based Browser for Large Online Resources. In Motta, E., Shadbolt, N., Stutt, A., and Gibbins, N., editors, *Proceedings of the 14th International Conference on Engineering Knowledge in the Age of the Semantic Web – EKAW 2004*, volume 3257 of *Lecture Notes in Computer Science*, pages 433–448, Whittlebury Hall, United Kingdom. Springer.

[Sutton, 1994] Sutton, S. A. (1994). The Role of Attorney Mental Models of Law in Case Relevance Determinations : An Exploratory Analysis. *Journal of the American Society for Information Science*, 45(3) :186–200.

[Szyperski, 2003] Szyperski, C. (2003). Component technology : what, where, and how ? In *Proceedings of the 25th International Conference on Software Engineering – ICSE'03*, pages 684–693, Washington, DC, USA. IEEE Computer Society.

[Tavanti and Lind, 2001] Tavanti, M. and Lind, M. (2001). 2D vs 3D, Implications on Spatial Memory. In Andrews, K., Roth, S., and Wong, P. C., editors, *Proceedings of the IEEE Symposium on Information Visualization 2001 – INFOVIS'01*, pages 139–145, San Diego, CA, USA. IEEE Computer Society.

[Tian, 2004] Tian, J. (2004). Quality-Evaluation Models and Measurements. *IEEE Software*, 21(3) :84–91.

[Torhorst et al., 2001] Torhorst, J., Bucher, C., Kononen, J., Haas, P., Zuber, M., Köchli, O. R., Mross, F., Dieterich, H., Moch, H., Mihatsch, M., Kallioniemi, O. P., and Sauter, G. (2001). Tissue microarrays for rapid linking of molecular changes to clinical endpoints. *American Journal of Pathology*, 159(6) :2249–2256.

[Tran-Thuong and Roisin, 2003] Tran-Thuong, T. and Roisin, C. (2003). Structured Media for Authoring Multimedia Documents. *Series in Machine Perception and Artificial Intelligence*, 55 :293–314.

[Trousse, 2000] Trousse, B. (2000). Evaluation of the Prediction Capability of a User Behaviour Mining Approach For Adaptive Web sites. In *Proceedings of the 6th RIAO Conference – Content-Based Multimedia Information Access*, Paris, France.

[Tullis and Steton, 2004] Tullis, T. S. and Steton, J. N. (2004). A Comparison of Questionnaires for Assessing Website Usability. In *Proceedings of the Usability Professional's 2004 Conference – UPA 2004*, Minneapolis, MN, USA.

[Turing, 1950] Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59 :433–460.

[van Zwol and Apers, 2001] van Zwol, R. and Apers, P. (2001). Webspace query formulation : an overview. CTIT Technical Report TR-CTIT-01-46, Centre for Telematics and Information Technology (CTIT), Enschede, the Netherlands.

[Viti et al., 2007] Viti, F., Merelli, I., Galizia, A., D'Agostino, D., Clematis, A., and Milanesi, L. (2007). Tissue MicroArray : a Distributed Grid Approach for Image Analysis. *Studies in Health Technology and Informatics*, 126 :291–298.

[Wang et al., 2002] Wang, H., Wang, H., Zhang, W., and Fuller, G. N. (2002). Tissue microarrays : applications in neuropathology research, diagnosis, and education. *Brain Pathology*, 12(1) :95–107.

[Wattenberg, 2005] Wattenberg, M. (2005). A Note on Space-Filling Visualizations and Space-Filling Curves. In *Proceedings of the 2005 IEEE Symposium on Information Visualization – INFOVIS '05*, pages 24–29, Minneapolis, MN, USA. IEEE Computer Society.

[Wilson, 1999] Wilson, T. (1999). Models in information behaviour research. *Journal of Documentation*, 55(3) :249–270.

[Wong, 2003] Wong, B. (2003). A Study of the Metrics applied to the Software Evaluation Framework 'SEF'. In Khoshgoftaar, T. M., Geleyn, E., and Nguyen,

L., editors, *Proceedings of the Third International Conference on Quality Software – QSIC'03*, pages 52–58, Dallas, TX, USA.

[Wu et al., 2003] Wu, A.-W., Gu, J., Ji, J.-F., Li, Z.-F., and Xu, G.-W. (2003). Role of COX-2 in carcinogenesis of colorectal cancer and its relationship with tumor biological characteristics and patients' prognosis. *World Journal of Gastroenterology*, 9(9) :1990–1994.

[Xu et al., 2006] Xu, G., Wu, C., and Du, X. (2006). Sentences, Hierarchical Clustering for Shopping Search. In *Proceedings of the 2006 International Conference on Semantics, Knowledge and Grid – SGK 2006*, pages 47–50, Guilin, China. IEEE Computer Society.

Publications

Revues Nationales

J. Bourbeillon, C. Garbay C. and F. Giroud. Une perspective analytique pour la recherche d'information. Application : conception et évaluation de Tissue Microarrays. *Revue ISI - Ingéniérie des Systèmes d'Information, Numéro spécial Systèmes d'Information Spécialisés, Hermès Sciences*, 11(5) :109-135, 2006.

Congrès Internationaux

J. Bourbeillon, C. Garbay, J. Simony-Lafontaine and F. Giroud. Multimedia Data Management To Assist Tissue MicroArrays Design. In : Silvia Miksch, Jim Hunter, Elpida Keravnou (Eds.) : *Artificial Intelligence in Medicine. Proceedings of the 10th Conference on Artificial Intelligence in Medicine, AIME 05.* Aberdeen, Scotland, 23-27 July 2005.

F. Giroud, J. Simony-Lafontaine, M.-P. Montmasson, R. Heus, **J. Bourbeillon** and C. Garbay. The TMAs-Explorer Platform : an integrated tool including a virtual TMA concept. In : *Abstract Book of the 7th European Congress on Virtual Microscopy and 1st International Congress on Virtual Microscopy.* Poznan, Poland, July 8-11 2004.

Congrès Nationaux

J. Bourbeillon, C. Garbay and F. Giroud. Génération de documents multimédia adaptatifs dans une perspective analytique. In : *Actes de la conférence INFORSID 05*. Grenoble, France, 24-27 mai 2005.

J. Bourbeillon, C. Garbay and F. Giroud. Gestion de connaissances et de données dans l'aide à la conception de Tissue Microarrays. *RNTI E5 - Extraction des connaissances : Etat et perspectives*. Rédacteurs invités : Florence Cloppet, J-Marc Petit, Nicole Vincent. Paris, France, 18 janvier 2005.

Annexes

Annexe

A Exemple de fichier XML de modèle de tâche

Une vue partielle du modèle de tâche pour une tâche de comparaison est proposée Listing A.1. Après un rappel de la tâche considérée au sein de la balise <Task>, ce modèle est décomposé en trois <Part> qui chacune correspond à un des sous-problèmes majeurs de la tâche de synthèse : sélection, organisation et présentation. Au sein du Listing A.1, seules les opérations les plus importantes de sélection sont présentées. Au sein de chaque <Part>, la hiérarchie des sous-tâche élémentaires est représentée par une hiérarchie de balises <Section>, <SubSection> et <Category>. Une tâche élémentaire est représentée par une balise <Component>. Pour chaque <Component>, sont listés paramètres et cibles.

Les paramètres sont représentés par une balise <Param>. Chaque paramètre, en plus de son nom, indique au moins une source, qui définit l'origine de la spécialisation : «Query» pour la requête, «Ontology» pour les représentations de connaissances et «Blackboard» pour le tableau noir. L'attribut path indique le chemin de l'élément dans la hiérarchie des représentations de connaissances ou dans le fichier de requête ou du blackboard. De manière optionnelle, un paramètre peut inclure des balises <Alternate>, décrivant les sources supplémentaires possibles, dont l'utilisation a été présentée précédemment, ou des balises <Additional> quand plusieurs sources doivent être utilisées conjointement pour spécialiser un paramètre.

Les cibles correspondent à la destination du résultat du traitement réalisé par le composant et sont indiquées par des balises <Target>. Elles se limitent à un nom (attribut name) et un chemin de destination au sein du tableau noir (attribut path).

Un modèle de tâche est donc un document XML décrivant de façon hiérarchisée les sous-tâches à résoudre ainsi que leurs entrées et sorties. Cette hiérarchie correspond

à une version élaguée pour une tâche particulière de la taxonomie de l'ensemble des sous-tâches élémentaires possibles.

Listing A.1: Modèle de tâche partiel pour une tâche de type comparaison - Dans ce listing, sont présentés les éléments principaux du modèle de tâche correspondant à la première catégorie de sous-tâches à résoudre : les opérations de sélection.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<TaskModel>
  <Task>Comparison</Task>
    <Part name="Selection">
      <Section name="Elements">
        <SubSection name="UserDefined">
          <Category name="SelectionCriteria">
            <Component name="CriteriaApplication">
              <Param name="Criteria" source="Query" path="Needs/Criteria"/>
              <Target name="Items" path="Selection/ListItems" />
            </Component>
          </Category>
        </SubSection>
      </Section>
      <Section name="Groups">
        <SubSection name="Constitution">
          <Category name="UserDefined">
            <Component name="GroupingCriteria">
              <Param name="Groups" source="Query" path="Needs/Group"/>
              <Target name="Groups" path="BlackBoard/Selection/ListGroups" />
            </Component>
            <Component name="GroupsBuilding">
              <Param name="Groups" source="BlackBoard"
                  path="Selection/ListGroups"/>
              <Param name="Items" source="BlackBoard"
                  path="Selection/CleanListItems"/>
              <Target name="Views" path="Selection/ViewsUsed" />
              <Target name="Items" path="BlackBoard/Selection/ListGroupedItems" />
              <Target name="Counts" path="BlackBoard/Selection/ListGroupsCounts"/>
            </Component>
          </Category>
        </SubSection>
      </Section>
      <Section name="Validity">
        <SubSection name="Objects">
          <Category name="MissingData">
            <Component name="MissingDataRemoval">
              <Param name="Items" source="BlackBoard" path="Selection/ListItems"/>
              <Param name="Requirements" source="Query" path="Needs/Goal">
                <Additional source="Query" path="Needs/Group" />
                <Additional source="Query" path="Needs/Criteria" />
                <Additional source="Query" path="Needs/Ordering" />
              </Param>
              <Target name="CleanListItems" path="Selection/CleanListItems" />
            </Component>
          </Category>
        </SubSection>
      </Section>
      <Section name="Finalisation">
        <SubSection name="Documents">
          <Category name="Factual">
            <Component name="FactualDocumentWriting">
              <Param name="Groups" source="Query" path="Needs/Group"/>
              <Param name="Data" source="BlackBoard"
                  path="Selection/ListGroupedItems"/>
              <Param name="Counts" source="BlackBoard"
                  path="Selection/ListGroupsCounts"/>
              <Param name="Lang" source="BlackBoard" path="Presentation/Language"/>
              <Target name="File" path="Selection/FactualDocumentLocation" />
            </Component>
          </Category>
        </SubSection>
      </Section>
    </Part>
   [...]
</TaskModel
```

ANNEXE

B Exemple de fichier XML de concept de la taxonomie du domaine d'étude

Un exemple de représentation d'un concept de la taxonomie du domaine d'étude est présenté Listing B.1 pour une notion qualitative : la notion d'organe. Elle est décomposée en :

- ⋆ Titre et description : en plusieurs langues, ils permettent l'identification de l'élément et une explication de ce qu'il représente,
- ⋆ Contenu : cette partie du fichier permet la description de l'élément. Elle inclut un type, qui permet de spécifier la nature de l'élément et conduit à des traitements différents par les composants résolvant le problème de synthèse. Pour l'instant, les types possibles sont «quantitatif» (correspondant à des nombres) et «qualitatif» (représentant des listes de valeurs textuelles prédéfinies) ; des types «date» et «texte» sont à l'étude. Ici, l'organe est de type qualitatif, ce qui implique de lister les valeurs possibles ainsi que leur titre en plusieurs langues et une valeur par défaut,
- ⋆ Accès Base de Données : cette section permet un mapping avec la base de données, en spécifiant la vue et le champ correspondant à l'élément.

Listing B.1: Portion de fichier représentant une entité du domaine d'étude - Chaque entité est décrite par une balise <Element>. Il inclut un titre et une description en plusieurs langues, une partie <Contents> et une partie <DBAccess>. La partie <Contents> permet de définir le type d'entité (quantitative ou qualitative), ici qualitative pour l'organe; la liste des valeurs possibles telles que trouvées dans la base de données, ainsi que le titre correspondant en diverses langues; une valeur par défaut. La partie <DBAccess> permet de réaliser un mapping entre l'élément de taxonomie décrit et le champ correspondant en base de données.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>

<Element name="Organ">
  <Title lang="en">Organ</Title>
  <Title lang="fr">Organe</Title>
  <Description lang="en">This item defines the organ in which the lesion is
     located.</Description>
  <Description lang="fr">Cet élément définit l'organe dans lequel la lésion est
     localisée.</Description>
  <Contents>
    <Type>Qualitative</Type>
    <Values>
      <Value name="brain">
        <Title lang="en">Brain</Title>
        <Title lang="fr">Cerveau</Title>
      </Value>
      <Value name="breast">
        <Title lang="en">Breast</Title>
        <Title lang="fr">Sein</Title>
      </Value>
      <Value name="colon">
        <Title lang="en">Colon</Title>
        <Title lang="fr">Côlon</Title>
      </Value>
      <Value name="liver">
        <Title lang="en">Liver</Title>
        <Title lang="fr">Foie</Title>
      </Value>
      <Value name="lung">
        <Title lang="en">Lung</Title>
        <Title lang="fr">Poumon</Title>
      </Value>
      [...]
      <Default>colon</Default>
    </Values>
  </Contents>
  <DBAccess>
    <View>patient</View>
    <Field>organ</Field>
  </DBAccess>
</Element>
```

ANNEXE

C Exemple de fichier XML de modèle de requête

Un fichier de modèle de requête, décrivant les éléments de formulaire nécessaires pour une étude de type «Comparaison», est présenté dans le Listing C.1. Ce modèle de requête décrit les deux parties de formulaire qui sont spécifiques de la tâche considérée.

La partie <FormPart name="Needs" ontology="Study"> correspond à la partie «Besoins» de la requête structurée et indique la taxonomie à utiliser (ici ontology="Study" pour le domaine d'étude). Elle définit pour chaque rôle le titre correspondant en diverses langues. Elle spécifie si plusieurs éléments peuvent être inclus dans le rôle (multi="1" pour oui et multi="0" pour non) et un type éventuel pour définir si une valeur (type="value") ou un pas (type="range") doivent compléter la sélection d'un élément de taxonomie.

La partie <Settings> correspond à la partie «Contraintes expérimentales» de la requête structurée et indique la taxonomie à utiliser (ici ontology="Experimental" pour les connaissances expérimentales). Elle décrit la hiérarchie de contraintes expérimentales qui peuvent être spécifiées, ainsi que l'élément de taxonomie correspondant, qui sert de source de valeurs possibles pour la saisie.

Listing C.1: Modèle de requête pour une tâche de type comparaison - Ce modèle décrit les deux parties du formulaire de saisie qui sont spécifiques de la tâche : les besoins (partie <FormPart name="Needs" >) et les contraintes expérimentales (partie <Settings>).

```
<QueryDef>
  <Task>Comparison</Task>
  <FormPart name="Needs" ontology="Study">
    <Title xml:lang="fr">Besoins</Title>
    <Title xml:lang="en">Needs</Title>
    <Section name="Goal">
      <Title xml:lang="fr">But</Title>
      <Title xml:lang="en">Goal</Title>
    </Section>
    <Section name="Criteria" multi="1" type="value">
      <Title xml:lang="fr">Critères de sélection</Title>
      <Title xml:lang="en">Selection criteria</Title>
    </Section>
    <Section name="Group" multi="1" type="range">
      <Title xml:lang="fr">Critères de groupement</Title>
      <Title xml:lang="en">Grouping criteria</Title>
    </Section>
    <Section name="Ordering" multi="1">
      <Title xml:lang="fr">Critères de tri</Title>
      <Title xml:lang="en">Ordering</Title>
    </Section>
  </FormPart>
  <Settings>
    <Section name="Geometry">
      <SubSection name="LengthWidthRatio"  source="Experimental/Conception/
          SpecificParams/Organisation/Geometry/LengthWidthRatio" />
      <SubSection name="VirtualSlideWidth" source="Experimental/Conception/
          SpecificParams/Organisation/Geometry/VirtualSlideWidth" />
      <SubSection name="VirtualSlideLength" source="Experimental/Conception/
          SpecificParams/Organisation/Geometry/VirtualSlideLength" />
    </Section>
    <Section name="Contents">
      <SubSection name="Language"source="Experimental/Conception/
          SpecificParams/Presentation/Contents/Language" />
      <SubSection name="Colour">
        <SubSection name="MinColour" source="Experimental/Conception/
            SpecificParams/Presentation/Contents/Colour/MinColour" />
        <SubSection name="MaxColour" source="Experimental/Conception/
            SpecificParams/Presentation/Contents/Colour/MaxColour" />
      </SubSection>
    </Section>
  </Settings>
</QueryDef>
```

Annexe

D Processus de saisie de requête au sein de l'interface

La construction du fichier XML de requête structurée est réalisée suite à la saisie par l'utilisateur, au sein d'une interface construite à partir du modèle de requête, des éléments pertinents pour chaque rôle de chaque partie de la requête structurée. Cette saisie est réalisée au cours d'un enchaînement de formulaires contraints par les connaissances incluses dans l'archétype de l'utilisateur.

Après le choix de l'élément de menu «Nouvelle requête», l'utilisateur est dirigé sur un premier formulaire de création de requête, dont une copie d'écran est présentée Fig. D.1. Ce formulaire correspond à la partie «Généralités » de la requête et permet la saisie de titre et description de la requête, le choix du domaine d'étude (ici les TMA) et de la tâche (ici comparaison).

Cliquer sur le bouton de création provoque la construction de deux formulaires spécifiques de la tâche de synthèse spécifiée.

Le permier formulaire, dont une copie d'écran est présentée Fig. D.2, rappelle les «Généralités » et permet la spécification d'éléments de configuration, soit des contraintes expérimentales. La liste des contraintes qui peuvent être spécifiées est construite à partir de la taxonomie des connaissances expérimentales, selon les instructions de la partie <Settings> du modèle de requête. Les valeurs proposées dans les listes déroulantes sont issues des fichiers XML de description d'une entité de la taxonomie. En haut de page, un lien «Besoins» donne accès au troisième formulaire illustré par la Fig. D.3.

Ce formulaire est décomposé en un ensemble de cadres, chacun correspondant à un rôle, tels que décrits dans la partie <Needs> du modèle de requête. Pour chaque

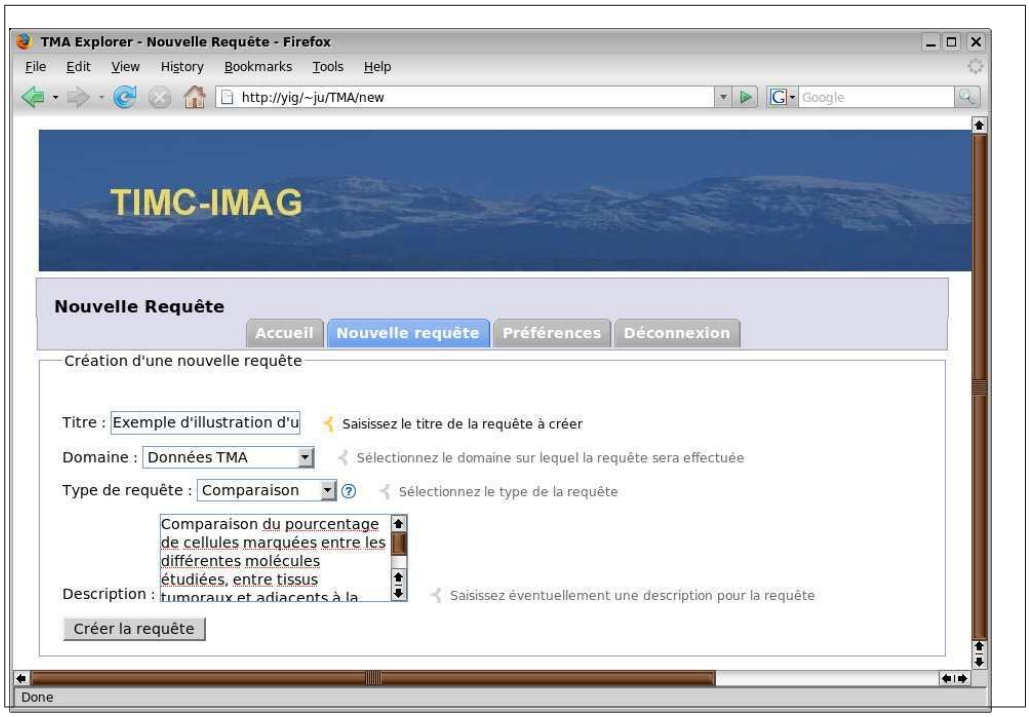


Fig. D.1: Saisie des «Généralités» de la requête - Ce formulaire permet la spécification du titre et de la description de la requête, du domaine d'étude et de la tâche.

rôle, des menus déroulants apparaissant en cascade au fur et à mesure des sélections qui y sont faites permettent un parcours de la taxonomie des connaissances du domaine d'étude et la construction du chemin de l'entité choisie. Au niveau feuille, possibilité est donnée de définir des attributs de l'élément tels qu'opérateur, valeur ou pas.

Une fois tous les rôles complétés, le bouton «Exécuter la requête» permet tout à la fois la sauvegarde des saisies sous forme d'un fichier de requête tel que présenté dans le paragraphe précédent et l'ajout de la requête à la liste des requêtes à traiter. Ce traitement implique tout d'abord une opérationnalisation du modèle de tâche, introduite dans la prochaine section.

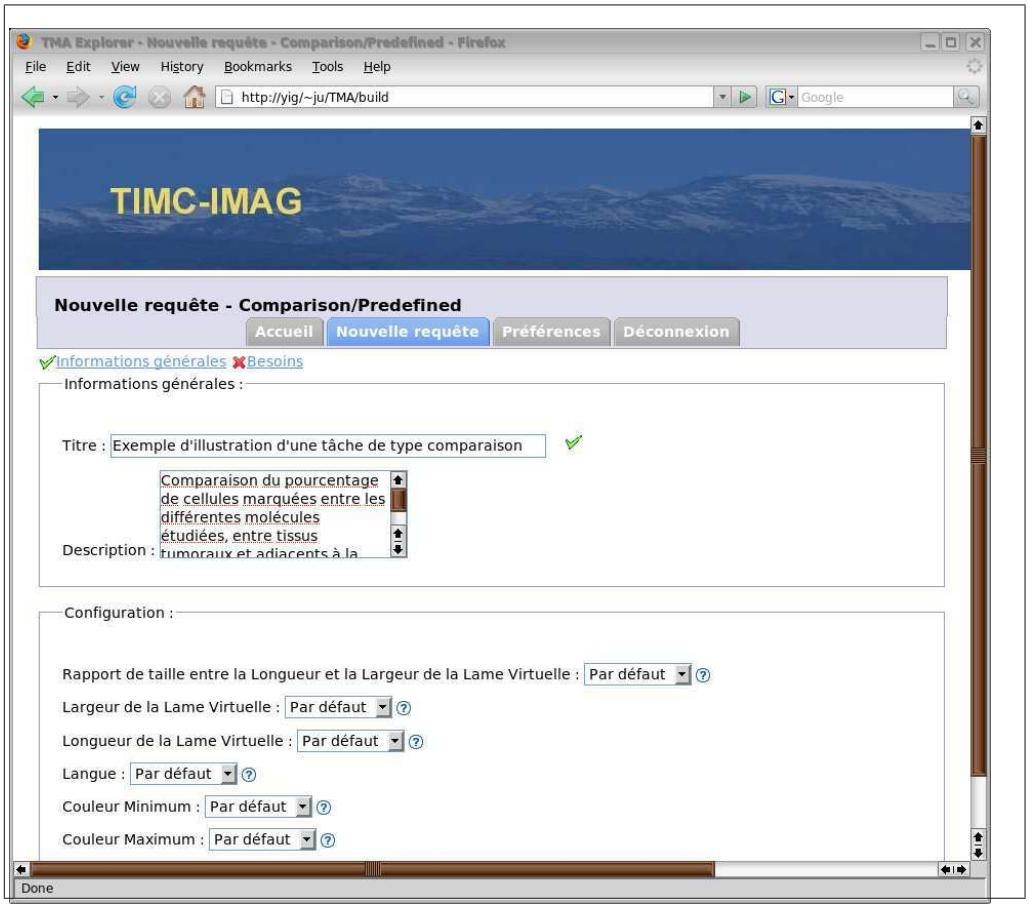


Fig. D.2: Rappel des «Généralités» et saisie des contraintes expérimentales - Ce formulaire rappelle le titre et la description précédemment saisis et permet la spécification d'éléments de configuration, qui seront stockés dans la partie <Means> du fichier XML de la requête.

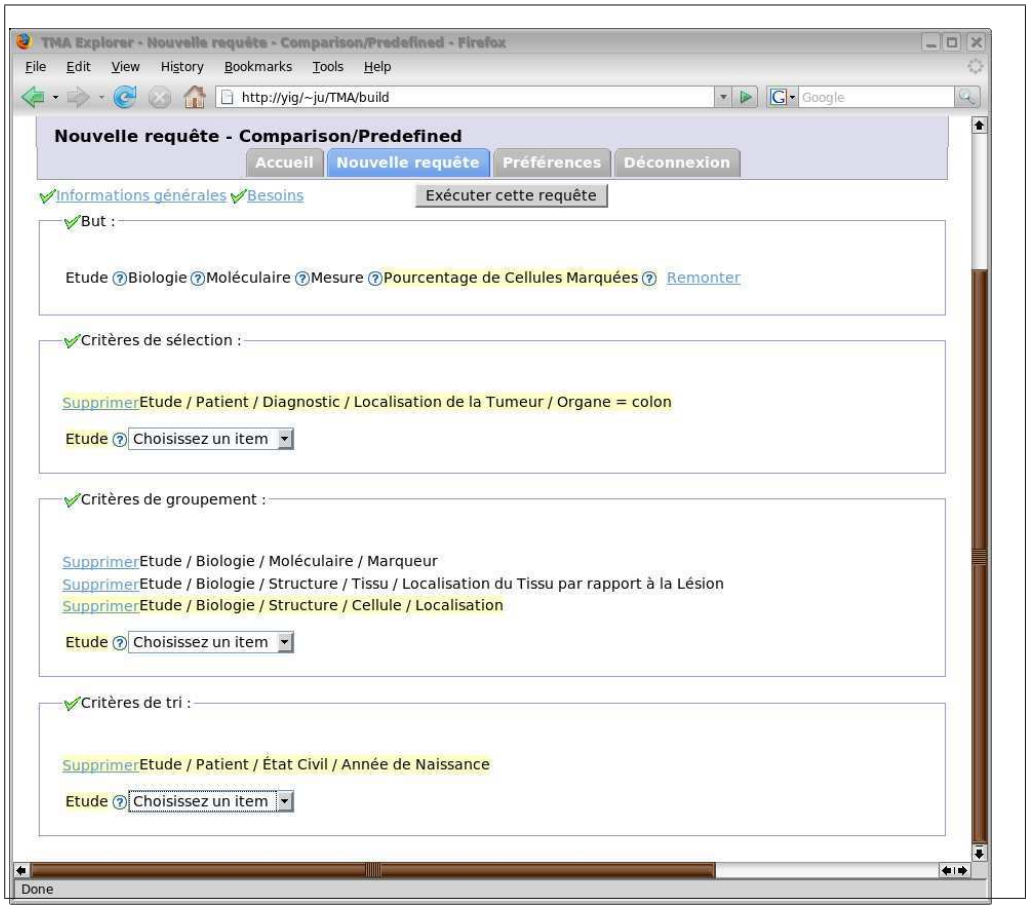


Fig. D.3: Saisie de requête - Un modèle de requête, spécifique de la tâche choisie par l'utilisateur, est utilisé au sein du framework Orbeon pour construire un formulaire de saisie de requête que l'utilisateur complète avec des informations issues de son archétype. Les données saisies et des informations utilisateur sont alors enregistrées dans un fichier de requête.

ANNEXE

E Exemple de fichier XML de requête structurée

Le fichier de requête, présenté Listing E.1 est structuré en deux parties : une partie d'en-tête et une partie <Task>. L'en-tête regroupe les éléments faisant partie des Généralités telles que le domaine applicatif considéré, le titre et la description de la requête. S'y ajoutent un identifiant numérique pour la requête, le nom de l'utilisateur qui a réalisé la saisie, et un timestamp repésentant l'heure et la date de saisie. La partie <Task> identifie la tâche de synthèse considérée et regroupe besoins (partie <Needs>) et contraintes expérimentales (partie <Means>).

Au sein de la partie <Needs>, chaque rôle de la requête structurée est représenté par une <Section>, dont l'attribut «name» définit le rôle. Ainsi par exemple, la <Section name=«Goal»> contient le but de l'étude. Chaque entité ayant un rôle donné est représentée par un <Element>. Ces éléments correspondent à des entités de la taxonomie du domaine d'étude et sont représentés par leur chemin dans cette taxonomie. Ils peuvent inclure des attribut «operator» et «value», pour fixer la valeur d'une variable, ou «range» pour définir une taille de classe pour des variables quantitatives.

Au sein de la partie <Means>, les contraintes expérimentales sont classées en sections et sous-sections. Les éléments ont un attribut «source»à la fonction similaire au «name» de la partie <Needs>, mais sinon fonctionnent sur le même principe, avec comme taxonomie de base la taxonomie des connaissances expérimentales pour les contraintes de type Matériel, et la librairie de composants pour les contraintes de type Méthode.

Listing E.1: Exemple de fichier de requête structurée - Ce fichier reflète la formulation de l'étude décomposée au sein du Tab. 4.2.

```
<?xml version="1.0" encoding="utf-8"?>
<Query xmlns:ev="http://www.w3.org/2001/xml-events"
       xmlns:xf="http://www.w3.org/2005/02/xpath-functions"
       xmlns:xhtml="http://www.w3.org/1999/xhtml"
       xmlns:xxforms="http://orbeon.org/oxf/xml/xforms"
       xmlns:xforms="http://www.w3.org/2002/xforms"
       xmlns:xi="http://www.w3.org/2001/XInclude"
       domain="TMA"
       id="2">
 <User>ju</User>
 <Timestamp>1186049369</Timestamp>
 <Title lang="fr">Exemple d'illustration d'une tâche de type comparaison</Title>
 <Description lang="fr">Comparaison du pourcentage de cellules marquées entre les
    différentes molécules étudiées, entre tissus tumoraux et adjacents à la
    tumeur, en fonction de la localisation intracellulaire du marquage, chez les
    patients atteints d'un cancer du côlon</Description>
 <Task path="Comparison/Predefined">
  <Part name="Needs">
   <Section name="Goal">
    <Element name="Study/Biology/Molecular/Measure/PercentMarkedCells"/>
   </Section>
   <Section name="Criteria">
    <Element name="Study/Patient/Diagnostic/TumorLocation/Organ" operator="="
        value="colon"/>
   </Section>
   <Section name="Group">
    <Element name="Study/Biology/Molecular/Marker"/>
    <Element name="Study/Biology/Structure/Tissue/LocationTowardsLesion"/>
    <Element name="Study/Biology/Structure/Cell/Location"/>
    <Element name="Study/Biology/Structure/Cell/Location"/>
    <Element name="Study/Biology/Structure/Cell/Location"/>
   </Section>
   <Section name="Ordering">
    <Element name="Study/Patient/PersonalData/BirthYear"/>
   </Section>
  </Part>
  <Part name="Means">
   <Section name="Geometry">
    <SubSection name="LengthWidthRatio"
      source="Experimental/Conception/SpecificParams/Organisation/Geometry/
      LengthWidthRatio"/>
    <SubSection name="VirtualSlideWidth"
      source="Experimental/Conception/SpecificParams/Organisation/Geometry/
      VirtualSlideWidth"/>
    <SubSection name="VirtualSlideLength"
      source="Experimental/Conception/SpecificParams/Organisation/Geometry/
      VirtualSlideLength"/>
   </Section>
   <Section name="Contents">
    <SubSection name="Language"
      source="Experimental/Conception/SpecificParams/Presentation/Contents/
      Language"/>
    <SubSection name="Colour">
     <SubSection name="MinColour"
       source="Experimental/Conception/SpecificParams/Presentation/Contents/
       Colour/MinColour"/>
     <SubSection name="MaxColour"
       source="Experimental/Conception/SpecificParams/Presentation/Contents/
       Colour/MaxColour"/>
     </SubSection>
   </Section>
    [...]
  </Part>
 </Task>
</Query>
```

Annexe

F Exemple de fichier XML d'instance de tâche

Le Listing F.1 présente un exemple de modèle spécialisé. Dans cette instance de tâche, deux ensembles sont discernables.

Tout d'abord, trois <Part> sont à peu près identiques aux <Part> du modèle d'origine présenté en Annexe A. Seuls les paramètres spécialisés à partir de la requête, des préférences utilisateur ou des connaissances sont différents : leur attribut path devient une chaine de caractères de type Params/ParamX où X correspond à un numéro.

Ensuite, la balise <BlackBoard> recense les informations de spécialisation. Elle consiste un ensemble de balises <Group>, chacune correspondant à un paramètre, dont les <Elements> correspondent à la spécialisation du paramètre. L'attribut name est de type ParamX, où X correspond au numéro de l'attribut path du paramètre spécialisé dans la première partie du fichier.

Ainsi, par exemple, le paramètre Criteria du composant CriteriaApplication, qui permet la sélection des patients pertinents a pour path Params/Param0. Sa spécialisation est donc indiquée par le <Group> à l'attribut name Param0, soit une localisation de tumeur dans l'organe côlon.

Listing F.1: Vue partielle du modèle de tâche opérationnalisé - Ce fichier se décompose en deux ensembles : l'un constitué de trois balises <Part> est similaire au modèle d'origine. Le second, au sein d'une balise <BlackBoard>, définit un ensemble de balises <Group>. Chacune correspond à la spécialisation d'un paramètre, la correspondance ayant lieu entre l'attribut path dans le premier ensemble, et l'attribut name du groupe dans le second.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<OperationalModel>
  <Task>Comparison</Task>
  <Part name="Selection">
    <Section name="Elements">
      <SubSection name="Predefined">
        <Category name="SelectionCriteria">
          <Component name="CriteriaApplication">
            <Param path="Params/Param0" name="Criteria"/>
            <Target path="BlackBoard/Selection/ListItems" name="Items"/>
          </Component>
        </Category>
      </SubSection>
    </Section>
    <Section name="Groups">
      <SubSection name="Constitution">
        <Category name="Predefined">
          <Component name="GroupingCriteria">
            <Param path="Params/Param2" name="Groups"/>
            <Target path="BlackBoard/Selection/ListGroups" name="Groups"/>
          </Component>
          <Component name="GroupsBuilding">
            <Param path="Selection/ListGroups" name="Groups"/>
            <Param path="Selection/CleanListItems" name="Items"/>
            <Target path="BlackBoard/Selection/ViewsUsed" name="Views"/>
            <Target path="BlackBoard/Selection/ListGroupedItems" name="Items"/>
            <Target path="BlackBoard/Selection/ListGroupsCounts" name="Counts"/>
          </Component>
        </Category>
      </SubSection>
    </Section>
    [...]
  </Part>
  <BlackBoard>
    <Group name="Param0">
      <Element value="colon" operator="="
          name="Study/Patient/Diagnostic/TumorLocation/Organ"/>
    </Group>
    <Group name="Param2">
      <Element name="Study/Biology/Molecular/Marker"/>
      <Element name="Study/Biology/Structure/Tissue/LocationTowardsLesion"/>
      <Element name="Study/Biology/Structure/Cell/Location"/>
    </Group>
   [...]
  </BlackBoard>
</OperationalModel>
```

ANNEXE

G Exemple de fichier XML de description d'un composant

Le Listing G.1 propose un exemple de fichier de description pour un composant d'application de critères de sélection sur une base de données.

Ce fichier peut être décomposé en trois parties. Au début, les titres et descriptions en plusieurs langues jouent un rôle documentaire, fournissant des informations sur le rôle du composant. Ensuite, au sein de la balise <Targets> sont regroupées toutes les cibles potentielles, identifiées par un nom et documentées par un titre et une description. De plus, la balise <Parameters> recense les paramètres et leur documentation. En plus d'un nom, chaque paramètre a des attributs minOccurs, qui indique le nombre minimum d'éléments correspondant au paramètre, et maxOccurs, qui indique le nombre maximum d'éléments possible. Enfin, l'attribut javaclass de la balise principale indique la classe JAVA principale correspondant au composant.

Listing G.1: Exemple de description de composant - Ce fichier, présenté ici pour un composant permettant l'application de critères de sélection dans une base de données, a tout à la fois un rôle documentaire, par les descriptions qu'il contient, et pratique, en indiquant la classe JAVA principale correspondant au composant.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>

<Component name="CriteriaApplication"
    javaclass="docsyn.comp.selec.scrit.CriteriaApplication">
  <Title lang="en">Application of one or several selection Criteria</Title>
  <Title lang="fr">Application d'un ou plusieurs critères de sélection</Title>
  <Description lang="en">This component selects from the database all elements
     fitting the defined criteria and writes them in the blackboard.</Description>
  <Description lang="fr">Ce composants selectionne dans la base de données tous les
     éléments correspondant aux critères définis et les écrit dans le blackboard.
  </Description>
  <Targets>
    <Target name="Items">
      <Title lang="en">List of the selected items</Title>
      <Title lang="fr">Liste des éléments sélectionnés</Title>
      <Description lang="en">This blackboard element holds the list of identifiers
          for the items fitting the selection criteria provided as parameter.
      </Description>
      <Description lang="fr">Cet élément du tableau noir reçoit la liste des
          identifiants des éléments qui satisfont les critères de sélection fournis
          en paramètre.</Description>
    </Target>
  </Targets>
  <Parameters>
    <Param name="Criteria" minOccurs="1" maxOccurs="unbounded">
      <Title lang="en">Selection criteria</Title>
      <Title lang="fr">Critères de sélection</Title>
      <Description lang="en">This parameter defines the list of selection criteria
          to apply</Description>
      <Description lang="fr">Ce paramètre définit la liste des critères de sélection
          à appliquer.</Description>
    </Param>
  </Parameters>
</Component>
```

Annexe

H Exemple de fichier XML du tableau noir

Une portion de fichier de déchargement du tableau noir est présentée List. H.1.

Ce fichier est organisé en ensemble de balises <Location>, identifiées par un chemin, qui correspondent à un emplacement de lecture ou d'écriture dans le tableau noir. Chaque <Location> peut contenir soit des balises <Value>, décrivant des valeurs, soit des balises <Group> contenant des balises <Value>, fournissant une possibilité d'organisation hiérarchique pour le résultat d'un traitement.

Listing H.1: Portion de fichier dans lequel est déchargé le tableau noir - Ce fichier, présenté ici partiellement pour la requête de test considérée, permet d'accéder aux entrées et sorties de composants, à des fins d'évaluation du système. Sont représentées ici les valeurs de longueur et largeur de grille à atteindre (issues de la requête, des préférences utilisateur ou des connaissances) ainsi qu'une partie de la hiérarchie de groupes calculée.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<BlackBoardDump>
  <Location path="BlackBoard/Organisation/GridWidth">
    <Value value="200"/>
  </Location>
  <Location path="BlackBoard/Organisation/GridLength">
    <Value value="15"/>
  </Location>
  <Location path="BlackBoard/Organisation/FactDocGroups">
    <Group name="Marqueur: Ki67 AND Localisation du Tissu par rapport à la Lésion:
        Adjacent AND Localisation: Noyau AND Localisation: Noyau">
      <Value value="Marqueur: Ki67 AND Localisation du Tissu par rapport à la
          Lésion: Adjacent AND Localisation: Noyau AND Localisation: Noyau AND
          Localisation: Noyau"/>
    </Group>
    <Group name="Marqueur: BCl2 AND Localisation du Tissu par rapport à la Lésion:
        Tumoral AND Localisation: Cytoplasme">
      <Value value="Marqueur: BCl2 AND Localisation du Tissu par rapport à la
          Lésion: Tumoral AND Localisation: Cytoplasme AND Localisation:
          Cytoplasme"/>
    </Group>
    [...]
  </Location>
  [...]
</BlackBoardDump>
```

ANNEXE

I Exemple de fichier XML de document maître

Le Listing I.1 présente un exemple partiel de document maître. Ce document est organisé en deux sections.

La section <Organisation> sert de base à la construction d'une grille documentaire. Elle inclut tout d'abord une balise <Header> qui rappelle la requête et donne des contraintes générales sur la construction de la grille. Ainsi, l'attribut primaryview définit quelle vue est la vue principale dont les identifiants sont indiqués dans les cases de la grille. L'attribut grouptitles indique si les titres des groupes construits doivent être affichés ou non.

Ensuite cette section décrit la grille documentaire à construire. Les balises <Row> permettent de construire des rectangles horizontaux. Elles contiennent des <Group>, correspondant à des rectangles verticaux, identifiés par des titres. Au niveau de <Group> le plus fin, les balises <Line> regroupent des <Item> sur une ligne de la grille. Chaque <Item> est associé à un attribut value, qui correspond à la couleur de fond de la case dans la grille. Il contient un ou plusieurs <VItem>, qui chacun spécifie l'identifiant de l'item au sein d'une vue.

La section <Data> sert à la construction des vues. Pour chaque vue, elle comporte une section <DView> identifiée par son nom. Chaque balise <DView> comporte alors un ensemble de <DItem>, identifiés par un attribut id, contenant un ensemble de balises <Field>, une pour chaque champ de la vue en base de données.

Listing I.1: Portion de document maître - La partie <Organisation> décrit la position des entités sur la grille du document de synthèse, alors que la partie <Data> indique, pour chaque vue, les informations correspondant à l'individu, qui serviront par la suite à construire les vues associées à la grille dans le document de synthèse.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>

<MasterDocument>
  <Organisation>
    <Header primaryview="patient" grouptitles="1">
      <Title>Exemple d'illustration d'une tâche de type comparaison</Title>
      <Description>Comparaison du pourcentage de cellules marquées entre les
          différentes molécules étudiées, entre tissus tumoraux et adjacents à la
          tumeur, en fonction de la localisation intracellulaire du marquage, chez
          les patients atteints d'un cancer du côlon</Description>
    </Header>
    <Row>
      <Group rowspan="3" name="Marqueur: Beta-Caténine" colspan="1">
        <Row>
          <Group rowspan="1" name="Localisation du Tissu par rapport à la Lésion:
              Adjacent" colspan="1">
            <Row>
              <Group rowspan="1" name="Localisation: Membrane" colspan="1">
                <Line>
                  <Item fg="0" value="#ffffff">
                    <VItem view="patient" id="9"/>
                    <VItem view="histology" id="423"/>
                  </Item>
                  <Item fg="0" value="#ffffff">
                    <VItem view="patient" id="16"/>
                    <VItem view="histology" id="430"/>
                  </Item>
  [...]
  </Organisation>
  <Data>
    <DView name="patient">
      <DItem id="11">
        <Field name="patientID">93-05025</Field>
        <Field name="surgeryDate">1998</Field>
        <Field name="sex">f</Field>
        <Field name="birthYear">1950</Field>
        <Field name="organ">colon</Field>
        <Field name="organArea">sigmoide</Field>
        <Field name="synchronousMetastasis">non</Field>
        <Field name="positiveNodes">9</Field>
        <Field name="observedNodes">17</Field>
        <Field name="lesionType">adénocarcinome lieberkühnien</Field>
        [...]
     </DView>
     <DView name="histology">
       <DItem id="11">
         <Field name="id">12810</Field>
         <Field name="sampleID">98-03862</Field>
         <Field name="patientID">93-05025</Field>
         <Field name="locationTowardsLesion">tumoral</Field>
         <Field name="marker">betacatenin</Field>
         <Field name="cellularLocation">nucleus</Field>
         <Field name="percentMarkedCells">30</Field>
         <Field name="stainingIntensity">2</Field>
         <Field name="stainingHeterogeneity">H3</Field>
         [...]
       </DItem>
       [...]
    </DView>
  </Data>
</MasterDocument>
```

ANNEXE

J Scenario de test pour les études utilisateur

J.1 Introduction

L'objectif de cette séance est de tester l'application Web permettant l'exploration des données TMA. Il s'agit tout à la fois d'évaluer le bon fonctionnement du système en utilisation réelle et de collecter des informations sur les impressions des utilisateurs potentiels. En tant que participant au projet, vous avez été sollicité pour participer.

La session de test est prévue pour durer environ une heure et se déroulera en deux phases. Dans un premier temps, il s'agit pour vous de réaliser un scénario de test, soit de suivre un cheminement défini au sein de l'application, scénario qui est décrit dans ce document. Dans un second temps, il vous sera demandé de remplir un questionnaire. Ce questionnaire consiste en un jeu d'affirmations pour lesquelles il vous faudra indiquer votre degré d'approbation, de 1 (pas du tout d'accord) à 5 (tout à fait d'accord), ainsi que d'un champ permettant de rédiger des commentaires non couverts par les questions précédentes.

Je vous invite tout d'abord à lire le scénario, afin de vous familiariser avec les tâches à accomplir.

Durant l'exécution du scénario de test, je me tiendrai à vos côtés pour réaliser des mesures et prendre des notes, et vous êtes invité à indiquer à voix haute tout commentaire ou réflexion que vous jugerez pertinent.

Quand vous êtes prêt à commencer, indiquez le moi.

J.2 Authentification

L'application permettant l'exploration des données TMA est une application Web, donc accessible sur internet, dont l'accès est contrôlé. Pour l'utiliser, il faut donc procéder à une authentification par nom d'utilisateur et mot de passe qui va être testée dans la première partie de cette session. Durant cette séance, vous utiliserez un compte spécifique de test. Vous pouvez maintenant accéder au système en suivant les instructions suivantes :

- ⋆ Aller sur le site correspondant au système : http ://yig.imag.fr/ ju/
- ⋆ Se connecter en utilisant les paramètres :
  - ⋆ Nom d'utilisateur : ju
  - ⋆ Mot de passe : testtma

Vous arrivez alors sur la page d'accueil. Cette page d'accueil est divisée en deux sections :

- ⋆ la partie supérieure contient un lien qui conduit vers les formulaires de saisie d'une nouvelle requête,
- ⋆ la partie inférieure liste les requêtes précédemment exécutées ainsi que des informations complémentaires les concernant et permet d'accéder à leurs résultats.

Les activités accessibles par ces deux parties de la page d'accueil vont faire l'objet des prochaines sections de la session de test.

J.3 Nouvelle requête

Afin d'explorer les données selon une nouvelle perspective, il faut tout d'abord créer une nouvelle requête. Cette fonctionnalité est accessible par le lien «Créer une nouvelle requête» de la première section de la page d'accueil ou l'onglet «Nouvelle Requête» en haut de page. Cliquer sur l'un de ces liens permet d'accéder à un premier écran de saisie qui permet :

- ⋆ la définition d'un titre pour l'étude, qui permet une description succincte,
- ⋆ le choix du domaine d'étude (TMA ou Élections, où Élections correspond à l'application jouet sur les données élections US),
- ⋆ le choix d'un type de requête (comparaison ou évolution, qui sont les deux types d'études disponibles pour le moment),
- ⋆ la définition d'une description, qui permet de prendre des notes plus complètes que le titre sur l'étude correspondante.

Afin de tester la saisie de requête, on va tout d'abord voir deux études imposées puis prendre l'opportunité de «jouer» avec le système.

J.4 Étude de la famille «comparaison»

La première étude que l'on veut réaliser pourrait être, en langue naturelle :

«Comparaison du pourcentage de cellules marquées, entre les différents marqueurs, entre tissus tumoraux et adjacents à la tumeur, en fonction de la localisation intracellulaire du marquage, chez les patients souffrant d'un cancer du côlon.»

Afin de réaliser cette étude, il faut tout d'abord compléter le formulaire décrit au paragraphe précédent de façon pertinente par rapport à l'étude puis créer la requête.

Une fois la requête créée, plusieurs onglets, matérialisés par des liens hypertextes, sont disponibles :

- ⋆ l'onglet «Informations générales» permet de relire et corriger les saisies précédentes,
- ⋆ ce même onglet donne aussi accès à une partie «Configuration» qui permet de spécifier des paramètres expérimentaux. Parmi ces paramètres expérimentaux, choisir une valeur de 50 pour la largeur de lame virtuelle (ceci permet une visualisation optimale dans le cadre de cet exemple),
- ⋆ La section «Besoins» permet de décrire l'objet de de l'étude.

Il vous faut alors décrire l'étude à réaliser dans la section «Besoins», dont les différentes composantes sont :

- ⋆ But : l'élément à étudier ; les valeurs de cet item serviront à définir les couleurs de fond des cases de la grille,
- ⋆ Critères de sélection : le ou les éléments(s) permettant de définir la population à étudier,
- ⋆ Critères de groupement : le ou les éléments(s) permettant la constitution des groupes et sous-groupes ; l'ordre des éléments dans cette section correspond à l'ordre de constitution des groupes,
- ⋆ Critères de tri : l'élément selon les valeurs duquel les individus seront ordonnés au sein de chaque sous-groupe de niveau le plus fin.

Au cours de la saisie, vous pouvez observer :

- ⋆ le système de listes déroulantes et le système d'aide associé,
- ⋆ les indicateurs visuels signalant la complétude/l'incomplétude des diverses sections de la requête,
- ⋆ la présentation de l'ontologie du domaine pour évaluer les chemins à parcourir dans la taxonomie pour définir l'étude sera ajoutée prochainement. Elle est mise à disposition sous forme papier dans le cadre de cette évaluation.

Une fois réalisées les sélections pertinentes dans les listes déroulantes permettant la définition de l'étude, vous pouvez lancer le traitement de la requête. Vous êtes

alors ramenés à la page d'accueil où une section centrale liste maintenant les requêtes en cours de traitement. Le rafraîchissement de la page n'étant pas automatique, il vous faut rafraîchir manuellement la page jusqu'à ce que votre requête apparaisse dans la liste des requêtes exécutées.

Pour accéder au résultat de la requête, il suffit de sélectionner la ligne de la requête dans la liste et de cliquer sur le lien «Voir le résultat». L'observation de ces résultats devrait permettre d'en tirer des conclusions au niveau biologique.

J.5 Étude de la famille «évolution»

La seconde étude que l'on veut réaliser pourrait être, en langue naturelle :

«Évolution du nombre de ganglions envahis, en fonction du nombre de ganglions observés, avec visualisation de la composante T du stade du cancer, chez des patients atteints d'un cancer du côlon.»

Comme pour la requête précédente, il faut accéder au formulaire de création de requête et réaliser les saisies pertinentes pour créer l'étude. Dans le cas d'une requête de type «évolution», le formulaire de saisie des «besoins» est légèrement différent et inclut :

- ⋆ But : l'élément à visualiser ; les valeurs de cet item serviront à définir les couleurs de fond des cases de la grille,
- ⋆ Axe des abscisses : l'élément dont les valeurs seront sur cet axe,
- ⋆ Axe des ordonnées : l'élément dont les valeurs seront sur cet axe,
- ⋆ Critères de sélection : le ou les éléments(s) permettant de définir la population à étudier.

Après avoir décrit l'étude, vous pouvez lancer le traitement de la requête et accéder aux résultats comme pour la requête de type «comparaison».

J.6 Études libres

Sur le modèle des études présentées, vous pouvez essayer d'exprimer quelques autres études et observer leurs résultats.

J.7 Accès aux anciennes requêtes et reformulation

La page d'accueil permet actuellement d'accéder à la liste des anciennes requêtes. Cette liste indique le numéro identificateur de la requête (incrémenté automatiquement), le domaine d'étude concerné, le titre et un état. Cet état peut être soit «Terminé», si le processus de traitement s'est déroulé sans problème, soit «Erreur !», si le traitement a échoué (en général à cause d'un bug du système, tous n'étant pas encore résolus, ou à cause d'un élément en cours de développement).

Le choix d'une ancienne requête est réalisé par sélection de la ligne correspondante. Trois actions sont alors possibles :

- ⋆ visualiser les résultats : pour les requêtes en «Erreur !» est affichée la cause de l'erreur, pour les requêtes «Terminé», le résultat du traitement,
- ⋆ ré-exécuter la requête : si les données ont changé, le résultat peut être différent,
- ⋆ reformuler la requête : les formulaires présentés lors de la saisie d'une nouvelle requête sont alors accessibles avec les champs saisis lors de la formulation précédente préremplis.

C'est cette reformulation qui va être testée ici. L'objectif est de modifier l'exemple de comparaison précédent. Il s'agit donc de le sélectionner et de modifier les éléments suivants :

- ⋆ largeur de lame virtuelle : 10,
- ⋆ limiter la sélection au marqueur β-caténine,
- ⋆ limiter la sélection aux patients dont le nombre de ganglions envahis ne soit pas égal à 0.

Observer les résultats et en tirer des conclusions biologiques.

J.8 Ajout et modification de préférences

La gestion de préférences utilisateurs par le biais de l'interface est en cours de développement et sera intégrée ultérieurement.

Annexe

K Questionnaire utilisateur

Utilisateur

Évaluation TMA Conception

Veuillez indiquer votre appréciation des affirmations suivantes entre 1 (pas d'accord) et 5 (tout à fait d'accord).

		Pas d'accord				Tout à fait d'accord
		1	2	3	4	5
Généralités						
1	De façon générale, je suis satisfait de la facilité d'utilisation du système.					
2	De façon générale, il a été facile d'apprendre à utiliser le système.					
3	La navigation au sein des diverses pages est intuitive.					
4	L'interface du système est plaisante.					
5	Je me suis senti à l'aise dans l'utilisation du système.					
6	De façon générale, je suis satisfait de l'utilisation du système.					
Formulation de requête						
1	Les tâches proposées, telles que je les comprends, correspondent à de vrais problèmes d'appréhension de données.					
2	Les tâches proposées (comparaison et évolution) sont adaptées aux problématiques courantes du domaine.					
3	La transposition d'un problème que je voudrais explorer en une requête ne pose pas de problème.					
4	L'organisation des éléments de l'interface de saisie est claire.					
5	La saisie de la requête est facile à réaliser.					
6	Le vocabulaire proposé est complet par rapport au domaine d'étude.					
7	L'organisation taxonomique du vocabulaire est adéquate.					
8	La syntaxe de la requête permet de décrire les tâches avec précision.					
9	La syntaxe de la requête permet d'exprimer toutes les tâches de comparaison ou d'évolution que je voudrais réaliser.					
10	Tous les types de tâches que je voudrais réaliser sont supportées.					
11	Il est facile d'explorer les données en modifiant la requête.					
Résultats de la requête						
1	Le résultat proposé apporte de nouvelles informations par rapport au problème posé.					
2	Le résultat pourrait être directement utilisé par exemple dans une publication.					
3	La forme du rendu (tableau) est pertinente.					
4	Le résultat proposé est porteur de sens.					
5	Le document de synthèse affiché m'aide à répondre au problème exprimé dans la requête.					
6	La relation entre la requête et le résultat affiché est facilement compréhensible.					
7	L'interprétation du résultat de la requête ne pose pas de problème.					
8	Les résultats affichés m'aident à envisager de nouvelles requêtes.					
9	Les résultats affichés suggèrent d'autres analyses avec d'autres outils.					
10	La navigation au sein des informations (grille et fiches associées) est intuitive.					
11	Je sais quand il y a eu des erreurs de traitement.					

Commentaires (pour toutes autres remarques non prises en compte dans les questions précédentes)

Annexe

L Résultats du questionnaire

Le Tab. L.1 présente les résultats du questionnaire pour la partie Généralités, correspondant à l'évaluation de l'utilisabilité.

Tab. L.1: Résultats du questionnaire pour l'utilisabilité - Ce tableau présente pour chaque utilisateur du panel la valeur entre 1 (pas d'accord) et 5 (tout à fait d'accord) attribuée à chaque affirmation.

Utilisateur	*F1*	*F2*	*C*	*A*	*N*	*S*
Novice	3	2	4	4	2	3
Étudiant 1	4	3	5	3	4	4
Étudiant 2	4	4	4	4	4	4
Technicien	5	4	4	4	5	4
Ingénieur	4	4	4	4	4	4
Doctorant imagerie	5	5	5	4	3	5
Maître de conférence	3	3	4	3	3	3
Anatomopathologiste	5	4	5	5	3	5

Le Tab. L.2 présente les résultats du questionnaire pour la partie Formulation de Requête, correspondant à l'évaluation de l'adéquation à la tâche.

Le Tab. L.3 présente les résultats du questionnaire pour la partie Résultats de la Requête, correspondant à l'évaluation de la pertinence informationnelle.

Tab. L.2: Résultats du questionnaire pour l'adéquation à la tâche - Ce tableau présente pour chaque utilisateur du panel la valeur entre 1 (pas d'accord) et 5 (tout à fait d'accord) attribuée à chaque affirmation. Les valeurs NA correspondent à une absence de réponse.

Utilisateur	*Int1*	*Int2*	*Inf1*	*Inf2*	*Inf3*	*Inf4*	*U1*	*U2*	*S1*	*S2*	*N*
Novice	2	NA	NA	NA	NA	2	4	4	4	4	4
Étudiant 1	5	5	5	5	4	4	5	4	5	5	3
Étudiant 2	5	NA	3	3	3	4	5	4	5	4	4
Technicien	4	4	4	3	3	3	4	3	4	4	5
Ingénieur	3	3	4	4	3	3	4	3	4	4	4
Doctorant imagerie	5	5	5	5	3	3	5	NA	5	5	5
Maître de conférence	5	5	4	3	2	3	4	4	5	4	3
Anatomo-pathologiste	3	4	3	4	2	NA	NA	NA	4	3	4

Tab. L.3: Résultats du questionnaire pour la pertinence informationnelle - Ce tableau présente pour chaque utilisateur du panel la valeur entre 1 (pas d'accord) et 5 (tout à fait d'accord) attribuée à chaque affirmation. Les valeurs NA correspondent à une absence de réponse.

Utilisateur	*A1*	*A2*	*A3*	*C1*	*C2*	*Exp1*	*Exp2*	*Exp3*	*Ext1*	*N1*	*N2*
Novice	3	2	4	4	4	NA	4	4	4	4	3
Étudiant 1	4	5	5	5	4	5	5	5	4	4	4
Étudiant 2	4	5	3	4	5	NA	3	5	3	3	4
Technicien	3	3	4	4	4	NA	4	5	5	5	5
Ingénieur	4	3	4	4	4	3	3	4	4	5	4
Doctorant imagerie	4	5	5	5	5	5	5	5	5	5	5
Maître de conférence	5	5	4	5	5	1	5	5	5	5	5
Anatomo-pathologiste	4	3	3	4	4	3	3	3	3	3	2

Résumé

Dans un contexte où des technologies et matériels nouveaux permettent un traitement en masse d'échantillons et où les données acquises sont de plus en plus partagées entre équipes de recherche, les scientifiques sont confrontés à un problème majeur d'exploitation de données. Plus précisément, utiliser ces données par des outils de fouille de données ou les replacer dans une démarche expérimentale classique nécessite une appréhension préalable de l'espace informationnel disponible afin de diriger le processus. Or cette appréhension de données est un problème complexe, peu supporté par les outils informatiques actuels.

L'objectif de cette thèse est de proposer une solution à ce problème d'appréhension des données scientifiques. Illustrée dans le domaine applicatif des Tissue MicroArrays, la proposition se base sur la notion de synthèse, inspirée des paradigmes de Recherche d'Information. Le modèle de synthèse envisagé, qui donne un rôle central à l'étude que le chercheur veut mener, par la notion de tâche, permet l'opérationnalisation d'un concept de Recherche d'Information orientée tâche par un prototype. Le prototype mis en place est validé par des étude de cas et une étude utilisateurs et ouvre des perspectives intéressantes d'extension du modèle ou d'extension à d'autres domaines applicatifs.

Mots clés

Appréhension de données, Synthèse de données, Recherche d'Information orienté tâche, Interface Homme/Machine, recherche en oncologie, Tissue MicroArrays, Visualisation d'Information

Abstract

In a context where new technologies and equipment allow for mass treatment of samples and where research teams share more and more acquired data, scientists are facing a major data exploitation problem. More precisely using this data through data mining tools or replacing it in a classical experimental approach require a preliminary grasp on the information space in order to direct the process. But acquiring this grasp on the data is a complex activity which is seldom supported by current software tools.

The goal of this thesis is to propose a solution to this scientific data grasp. Illustrated in the Tissue MicroArrays application domain, the proposal is based on the synthesis notion, which is inspired by Information Retrieval paradigms. The envisionned synthesis model gives a central role to the study the researcher wants to conduct through the task notion. It allows for the operationnalisation of a task-oriented Information Retrieval through a prototype. The prototype which has been developped is validated by case studies and an user study. It opens interesting prospects for the extension of the model or extensions towards other application domains.

Keywords

Data grasping, Data synthesis, task-oriented Information Retrieval, Human/Computer Interface, oncology research, Tissue MicroArrays, Information Visualisation