Systèmes d'Information Scientifique : des modèles conceptuels aux annotations sémantiques Application au domaine de l'archéologie et des sciences du vivant

Marinette Savonnet

Université de Bourgogne

Habilitation à Diriger des Recherches
Discipline : Sciences et Techniques

Marinette Savonnet

Systèmes d'Information Scientifique : des modèles conceptuels aux annotations sémantiques

Application au domaine de l'archéologie et des sciences du vivant

Soutenue le 12 septembre 2013 devant le jury composé de :

Madame le Professeur	Danielle Boulanger (Rapporteur)	Université Jean Moulin Lyon 3
Monsieur le Professeur	Bernard Espinasse (Rapporteur)	Université d'Aix-Marseille
Madame le Professeur	Chantal Reynaud (Examinatrice)	Université Paris-Sud
Madame le Professeur	Claudia Roncancio (Examinatrice)	Grenoble INP
Monsieur le Professeur	Kokou Yétongnon (Examinateur)	Université de Bourgogne
Monsieur le Professeur	Esteban Zimányi (Rapporteur)	Université Libre de Bruxelles

Laboratoire Électronique Informatique et Image (Le2i) UMR 6306
École Doctorale Sciences Pour l'Ingénieur et Microtechniques (SPIM)

Remerciements

À l'issue de la rédaction de cette Habilitation à Diriger des Recherches, je suis convaincue que l'HDR est loin d'être un travail solitaire. En effet, je n'aurais jamais pu réaliser ce travail sans le soutien de mon collègue *Éric* Leclercq.
Une pensée va vers *Marie-Noëlle* Terrasse qui a participé à tisser le fil conducteur de mon activité de recherches actuelle.

Je tiens à remercier *Kokou* Yétongnon dont la générosité, la bonne humeur et l'intérêt manifestés à l'égard de cette HDR m'ont permis de progresser.

Je tiens à remercier *Nadine* Cullot pour son soutien et son aide dans les démarches administratives.

Mes sincères remerciements vont aux Professeurs *Danielle* Boulanger, *Bernard* Espinasse et *Esteban* Zimányi qui ont répondu favorablement pour être les trois rapporteurs de cette HDR.
Je remercie également Mesdames les Professeurs *Chantal* Reynaud et *Claudia* Roncancio, qui m'ont fait l'honneur d'accepter d'être membres du jury, permettant ainsi d'avoir un jury paritaire.

Ce travail est le fruit de plusieurs collaborations scientifiques. Mes remerciements vont à ceux qui m'ont accompagné, au gré des vents tourbillonnants de la collaboration scientifique multidisciplinaire, dans cette aventure intellectuelle : *Pascale* Chevalier*, Laure* Saligny*, Christian* Sapin*, Patrick* Ducoroy*, Caroline* Truntzer *et Jean* Gascuel. J'ai toujours apprécié leurs capacités à relever les défis que je leur lançais. Sans eux ce document n'existerait pas. Ils m'ont transmis une part de leur expérience.
Je tiens à remercier celui qui a eu les doigts dans le cambouis (ou, plus précisément, sur le clavier) : *Arnaud* Millereux

Une pensée va vers mes enfants, *Thomas* et *Claire*, que je puisse leur montrer que l'on peut encore progresser.
Je dédie ce travail à mon époux *Bernard* qui a suivi non seulement les méandres de l'élaboration de ma thèse mais aussi de mon HDR.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Systèmes d'Information Scientifique

« *[...] all science, it seems, is becoming computer science.* »
George Johnson, journaliste scientifique, New-York Times mars 2001

Sommaire

Ce chapitre d'introduction donne une vue générale sur les Systèmes d'Information Scientifique, la problématique à laquelle je m'attache et la démarche employée.

Mes travaux de recherche depuis 2005 sont dans le cadre de l'eScience (*e-[nhanced] Science*). L'eScience désigne les infrastructures permettant le partage des productions scientifiques et l'exploitation *in silico* des données de la recherche dans des environnements logiciels collaboratifs permettant entre autres une communication rapide des résultats de recherche [NEO03, HTT09].

1.1 Systèmes d'Information Scientifique

Les Systèmes d'Information Scientifique ont pour but de produire de la connaissance et non pas de gérer ou contrôler une activité de production de biens ou de services comme dans les Systèmes d'Information d'entreprise. Les scientifiques sont amenés à construire leurs propres expérimentations pour vérifier et valider leurs hypothèses. Une expérimentation met en œuvre des chaînes de traitements (ou protocoles expérimentaux) plus ou moins complexes. De plus, les données, les paramètres et les traitements associés à chaque expérimentation doivent être sauvegardés afin de pouvoir être ré-exécutés plusieurs fois soit dans diverses configurations, soit avec différentes données en paramètre. En outre, la recherche scientifique est devenue fortement collaborative, impliquant des scientifiques de disciplines, d'organisations et de pays différents pour faire face à la complexité des problèmes traités. Par conséquent, les Systèmes d'Information Scientifique doivent permettre une conception collaborative et une configuration dynamique des

protocoles expérimentaux. De plus, ils doivent être capables de gérer à la fois une connaissance formelle et informelle et permettre un « cycle de découverte scientifique » (voir figure 1.1) [AKD10]. Comme pour les Systèmes d'Information d'entreprise, les Bases de Données sont au centre des mécanismes de traitement de l'information scientifique.

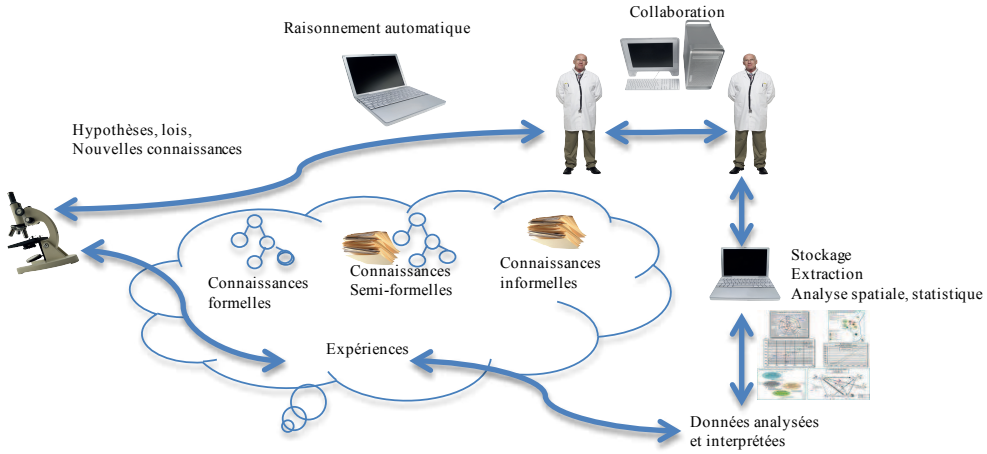


FIGURE 1.1 – Cycle de construction de la connaissance dans un Système d'Information Scientifique

1.1.1 Bases de Données scientifiques

Les Bases de Données scientifiques font l'objet de recherches constantes depuis les années 1980, cependant l'amplitude de ces recherches n'est pas comparable au domaine des Bases de Données pour les Systèmes d'Information d'entreprise [BDI+08]. En 2010, elles ont été identifiées comme un des enjeux majeurs de la recherche dans le domaine des Bases de Données [ABK+10]. Les domaines d'utilisation des Bases de Données scientifiques sont très nombreux. Ils recouvrent la physique des hautes énergies (par exemple l'expérience ATLAS [1]), l'astronomie (par exemple le projet Pan-STARRS [2]), les sciences du vivant comme la biologie (par exemple le projet Blue Brain [3]), la protéomique (par exemple le projet Human Proteome Project [4]), mais aussi des domaines en Sciences Humaines et Sociales (par exemple le projet DARIAH [5]) tels que la sociologie, l'archéologie.
Les Bases de Données scientifiques pourraient être définies dans une approche sommaire comme de vastes entrepôts de données pour lesquels des méthodes de recherche de données sur mesure seraient nécessaires. En effet, les Bases de Données scientifiques demandent des capacités de stockage importantes car le volume des données est sans rapport avec celui des Bases de Données d'entreprise. Les progrès rapides et constants des instruments d'observation scientifique (par

1. ATLAS est l'une des 5 expériences des collisionneurs LHC au CERN. Il s'agit d'un détecteur de particules ayant détecté le boson de Higgs, site Web : http://atlas.ch/
2. Pan-STARRS (*Panoramic Survey Telescope And Rapid Response System*) est un projet de relevés astronomiques qui effectue de l'astrométrie et de la photométrie d'une grande partie du ciel quasiment en continu, site Web : http://pan-starrs.ifa.hawaii.edu/public/
3. Le projet Blue Brain a pour objectif de créer un cerveau synthétique, site Web : http://bluebrain.epfl.ch/
4. Le projet Human Proteome Project est un programme mondial d'exploration du protéome humain, site Web : http:/www.hupo.org/research/hpp/
5. Le projet DARIAH (*Digital Research Infrastructure for Arts and Humanities*) a pour objectif la construction d'une infrastructure européenne, permettant de mettre en valeur, développer et soutenir la recherche dans toutes les disciplines des Sciences Humaines.

exemple un réseau de capteurs) et des outils de simulation (qui favorisent l'expérimentation *in silico*) conduisent à une véritable explosion des données, en termes de volume et de complexité [HT03]. Dans le domaine de la modélisation du climat, par exemple, l'augmentation des données est telle qu'elle devrait produire plusieurs centaines d'exabytes de données d'ici 2020. Dans le domaine de la physique des hautes énergies, les LHC (Grand collisionneur de hadrons) du CERN produisent 15 petabytes de données brutes par an. Cependant même si on remarque un usage important des ETL (*Extract Transform Load*) dans les processus qui traitent les données scientifiques, une Base de Données scientifiques n'est pas un entrepôt de données classique. D'ailleurs le besoin de construire des entrepôts de données spécifiques s'est ressenti : par exemple l'entrepôt SciBORQ [SKB11] a été construit au dessus de la Base de Données NoSQL MonetDB [6] [IGN+12].

Les Bases de Données scientifiques présentent des spécificités qui ne sont pas uniquement relatives au stockage. Contrairement aux Bases de Données d'entreprise qui sont modélisées sur des structures pré-définies par l'activité qu'elles gèrent [AWH92], les données scientifiques peuvent *a priori* n'entrer dans aucune structure pré-définie (objet même de la science) et elles sont soumises à des traitements numériques nombreux : les données brutes subissent généralement des étapes de pré-traitements pour retirer le bruit de fond et les normaliser, à l'issue de cette étape on obtient des données calibrées. Un contrôle de qualité des données calibrées est effectué pour aboutir à des données validées. Les données validées obtenues font l'objet d'analyses le plus souvent statistiques pour aboutir à des données dérivées. Lorsque les données dérivées sont associées à d'autres ensembles de données ou à la littérature du domaine on dit qu'elles sont interprétées [FJP90]. Pour gérer ces différentes étapes, assurer la traçabilité, le contrôle de la provenance et l'enchaînement des traitements, le besoin d'outils de gestion de *workflow* scientifique s'est fait rapidement sentir.

1.1.2 *Workflow* scientifique

Gil et al. [GDE+07] présentent les défis auxquels doit répondre un *workflow* scientifique : il doit être suffisamment flexible pour reproduire et enrichir les expérimentations ; être capable de gérer la provenance des données [7] ; détecter et gérer les inconsistances ainsi que les hétérogénéités induites par le fait que les données proviennent de différentes sources.
On peut considérer deux grandes catégories de *workflows* scientifiques : l'une orientée gestion et contrôle de données et l'autre orientée calcul. Les LIMS (*Laboratory Information Management System*) caractérisent la première catégorie, ils intègrent différentes fonctionnalités mises en œuvre dans un laboratoire telles que la traçabilité des échantillons, la coordination des équipements et des tâches. Ils peuvent être comparés à des ERP (*Enterprise Resource Planning*) permettant de traiter toutes les activités d'une plate-forme de recherche. Les *workflows* tels que Kepler [8] et Taverna [9] sont des exemples d'outils relevant de la seconde catégorie. Ils décrivent le protocole expérimental en offrant la possibilité de représenter, intégrer, traiter et analyser les données ; leur plate-forme d'exécution est la grille de calcul.

6. MonetDB est une Base de Données NoSQL de type Orientées Colonnes. Pour une relation R possédant k attributs, il existe k *Binary Association Tables* stockant deux attributs (OID, valeur), tous les attributs d'un même tuple ont le même OID.

7. Pour chaque donnée le *workflow* doit être capable de trouver la (les) donnée(s) dont elle est issue et comment elle a été produite [BBDH08].

8. Kepler est un *workflow* scientifique construit au-dessus de la plate-forme Ptolemy II de l'Université de Berkeley, site Web : http://kepler-project.org

9. Taverna est un *workflow* initié par l'équipe ^{my}Grid en Angleterre, utilisé principalement dans le domaine de la biologie. Les traitements dans cet environnement sont essentiellement assurés par des Services Web (par exemple le Service Web de l'outil BLAST, un Service Web qui renvoie une séquence de protéine au format FASTA). L'annotation des services permet de facilement réutiliser les *workflows* construits [OAF+04].

1.1.3 Spécificités des données scientifiques

La représentation et la gestion des données scientifiques posent des problèmes aux concepteurs de Systèmes d'Information [JO04]. En effet, les données à représenter et à gérer sont complexes car elles sont disparates, elles évoluent, elles sont hétérogènes et elles présentent souvent des problèmes de qualité. La disparité des données provient des multiples processus pouvant être pris en compte lors des expérimentations. Par exemple, une même molécule peut être vue par son nom, sa structure 2D ou la matrice d'adjacence de ses atomes. L'évolution de la connaissance fait évoluer les données scientifiques ce qui se traduit par une évolution des Systèmes d'Information. Par exemple, la Base de Données GenBank [10] voit son schéma modifié deux fois par an [NPG07]. Les Systèmes d'Information Scientifique doivent être extensibles pour intégrer et consulter des données non initialement prévues. L'hétérogénéité des données scientifiques provient de la multiplicité des sources de données dont elles sont issues [DOB95]. Leur qualité dépend de l'incertitude, l'incomplétude, l'incohérence et l'inconsistance de certaines données.

Les données scientifiques utilisent des modèles de représentation des données complexes (plusieurs centaines d'attributs) le plus souvent spatio-temporels et multi-échelles afin de représenter aussi bien une structure moléculaire que le positionnement de planètes [SOW84a, Pfa07]. Par exemple, MonetDB a étendu son modèle pour prendre en compte le caractère spatial des données [INGK07] que l'on retrouve dans des domaines comme l'astronomie, l'océanographie ou la climatologie. La Base de Données scientifique SciBD [Sto12] offre un tel modèle de données et un langage de manipulation spécifique SciQL [ZKIN11]. En revanche, des domaines comme la biologie ou la chimie préfèrent travailler avec des modèles de graphes ou de séquences. Les domaines appartenant aux humanités travaillent essentiellement avec des modèles documentaires s'appuyant le plus souvent sur des langages de la famille XML. Le modèle documentaire doit gérer du texte, des photographies, des plans, etc. L'ensemble de ces différentes données sont rarement modifiées (*no-overwrite*) une fois acquises mais doivent être intégralement conservées. De telles données représentent un grand défi pour la recherche car les techniques sont complexes et doivent passer à très grande échelle.

En résumé, les Bases de Données scientifiques sont un maillon essentiel dans les Systèmes d'Information scientifique. Elles demandent des traitements complexes non nécessairement prévisibles à l'avance et des données de qualité (aussi bien sur des données quantitatives que qualitatives). Elles doivent gérer des données très variables en terme de structure qui va au-delà de la simple hétérogénéité et des volumes très importants. Le modèle relationnel montre ses limites à prendre en compte des types de données aussi spécifiques, à offrir des structures arborescentes ou de graphe performantes pour traiter la provenance des données mais aussi à prendre en compte le caractère incertain des données scientifiques. De plus, les implémentations dans les SGBD souffrent d'une difficulté à traiter de la variabilité des données par l'utilisation des schémas relationnels [GLNS+05]. Ceci explique qu'historiquement, les grands projets de science, comme le LHC du CERN ou l'étude *Mission Planet Earth* de la NASA, ont choisi de développer leur solution *ad hoc* de gestion de données [IAJA11]. Un contre-exemple notable est le projet *Sloan Sky Survey* déployé au-dessus de SQLServer avec succès. Néanmoins, les solutions *ad hoc* devront être abandonnées car elles demandent des développements trop coûteux, trop spécifiques et trop complexes, aboutissent à des formats propriétaires avec des processus propres pour pouvoir être réutilisées ou re-créées pour chaque nouveau projet.

10. GenBank est une Base de Données de gènes, disponible publiquement, proposant les séquences nucléotidiques et leur traduction en protéines, site Web http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide

1.2 Défis auxquels nous répondons

La section précédente nous a permis de dégager quatre caractéristiques essentielles des Systèmes d'information Scientifique :

1. les Systèmes d'Information Scientifique ont pour but de **produire de la connaissance** et non de supporter ou contrôler une activité de production ou de commerce ;
2. les Systèmes d'Information Scientifique sont dits *data-intensive* car ils produisent chaque jour d'**importants volumes de données très variables pour des équipes de recherche multi-disciplinaires**. Il en résulte une très grande variabilité des structures de données qui va au-delà de la simple hétérogénéité et une variabilité des partenaires impliqués dans le Système d'Information Scientifique. La variabilité est l'essence même des Systèmes d'Information Scientifique ;
3. la production de connaissance nécessite des données complexes non nécessairement prévisibles à l'avance : le **schéma des données doit évoluer** rapidement sous l'impulsion de nouvelles découvertes et techniques d'acquisition de données ;
4. la **qualité des données** est essentielle aux Systèmes d'Information Scientifique, la validité des résultats est fortement dépendante de la qualité des données acquises. Bien que la variabilité entre les acteurs et la variabilité entre les études tendent à réduire cette qualité, les fonctionnalités des Systèmes d'Information Scientifique doivent maintenir et contrôler cette qualité.

Les Bases de Données classiques (SGBDR) ne permettent pas de répondre aux exigences d'extensibilité des Bases de Données scientifiques au niveau de la structure et des applications. Même si des couches d'abstraction permettant de réaliser le mapping objet-relationnel sont utilisées pour découpler les applications des données, l'extension de schéma a un impact important sur les applications.

Pour répondre à ces verrous, nous proposons une approche qui offre l'interopérabilité, l'extensibilité et la qualité des données grâce à un unique paradigme : **l'annotation sémantique**. Les mécanismes d'annotation, utilisés dans de nombreux domaines, permettent d'apporter aux Systèmes d'Information Scientifique l'extensibilité de la structure et la prise en compte de la multi-disciplinarité (voir figure 1.2). Tout comme la relation dans le modèle relationnel, l'annotation est une structure simple et quasi universelle qui permet de développer des composants génériques pour traiter l'extensibilité nécessaire. L'annotation est basée sur une ontologie et des règles qui permettent de découpler la connaissance du domaine de la connaissance sur les données.

1.3 Une approche basée sur un modèle formel d'annotation

La gestion de données scientifiques nécessite, au niveau du Système d'Information, un degré de souplesse qui est généralement beaucoup plus élevé que dans un Système d'Information d'entreprise. Un Système d'Information Scientifique doit être un système extensible 1) permettant l'ajout et la consultation de nouvelles connaissances et données et, 2) minimisant l'impact de l'évolution des données sur l'architecture du Système et sur les applications. La prise en compte des données non initialement prévues ne peut être traitée de manière statique, c'est-à-dire, en faisant évoluer le ou les schéma(s) à chaque nouvelle donnée et nécessite par conséquent un mécanisme dynamique d'extension de schéma.
Notre proposition est basée sur les annotations sémantiques afin de supporter l'interopérabilité, l'extensibilité et la qualité des données. Pour représenter les connaissances, nous utilisons à la fois des Bases de Données relationnelles et des langages du Web Sémantique (RDF et OWL).

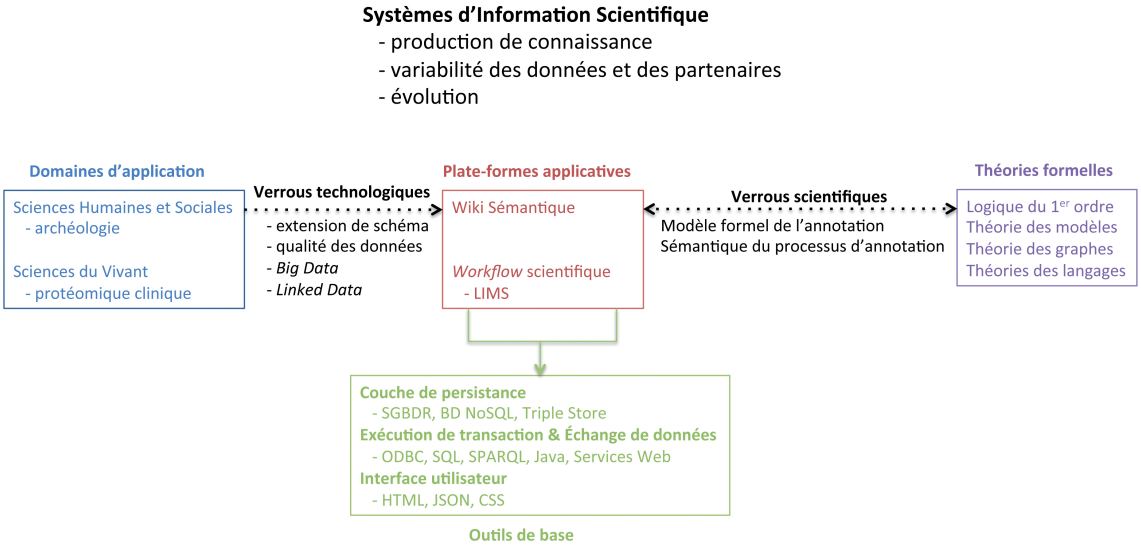


Figure 1.2 – Verrous scientifiques et technologiques traités dans les Systèmes d'Information Scientifique

1.3.1 Stockage multi-paradigmes

Les Systèmes d'Information Scientifique manipulent des données aux structures très variables qui vont au-delà de la simple hétérogénéité : séquences de nucléotides ou d'acides aminés, méta-données, textes, bibliographies, photographies provenant de différentes sources. Trois types de données doivent alors être considérés : 1) les données fortement structurées, 2) les données semi-structurées et 3) les données non structurées.
Les Bases de Données sont un bon outil pour stocker des données fortement structurées. Le modèle de données relationnelles présente deux niveaux : 1) le niveau schéma, défini pendant la phase de conception, décrit la structure d'un ensemble de données partageant des caractéristiques communes et 2) le niveau données où chaque objet obéit à une structure définie au niveau du schéma.
Une donnée semi-structurée n'a pas de schéma *a priori*. Il est implicite et véhiculé avec la donnée. Chaque objet contient son propre schéma, schéma et donnée sont confondus à un seul et même niveau. La plupart du temps, les données semi-structurées sont représentées sous la forme d'un graphe dont les feuilles contiennent les données et dont les nœuds et les arcs représentent la structure de l'ensemble. Ce type de données est généralement représenté en utilisant le format XML qui propose une structure flexible pour s'adapter à la grande variabilité des données.
La troisième catégorie est celle des données non structurées où n'est présente aucune notion de schéma de données. L'information est représentée en utilisant soit des images soit des phrases en langage naturel. Elle peut être transformée, en utilisant des techniques d'indexation et d'annotation, en données semi-structurées.

Classiquement, la conception d'une application s'appuie sur une Base de Données Relationnelles pour réaliser la couche de persistance. Bien que les Bases de Données Relationnelles sont un excellent outil, elles ne peuvent pas faire face à toutes les propriétés requises pour le stockage des données scientifiques, ni même à tous les types de traitements. En effet, dans une Base de Données Relationnelles, toutes les informations sont décrites *a priori*. Le schéma relationnel est établi très tôt lors de la phase d'analyse en s'appuyant sur une connaissance du domaine à un instant donné. Enfin, résultant de la normalisation et de l'adaptation au SGBD cible, le schéma

relationnel est en général très éloigné du modèle conceptuel ce qui rend sa compréhension et la prise en main de sa structure difficiles. Les nouvelles Bases de Données NoSQL ne s'attachent qu'à des problèmes particuliers : le volume avec le NoSQL orienté colonnes ou clé-valeur, les documents et les liens dans les bases de données documentaires. Ce type de bases de données NoSQL comme les bases de données relationnelles ne permettent pas de prendre en compte les requêtes "de type chemin" faisant appel à la récursivité, pourtant essentielles pour les analyses de données scientifiques. Les Bases de Données graphe traitent en partie le problème mais passent mal à l'échelle.

Nous rejoignons Ghosh [Gho10] et Ivanova et al. [IKM12] en proposant une infrastructure de persistance multi-paradigmes (voir figure 1.3 - Couche d'accès aux données). Les connaissances à gérer dépendant fortement du domaine à modéliser, l'infrastructure de persistance devra prendre en compte différents types de données. Par exemple les Sciences Humaines et Sociales manipulent essentiellement des documents, le mécanisme de persistance se tournera alors vers des Bases de Données NoSQL Orienté Document comme MongoDB [11] ou vers une plate-forme orientée document comme un wiki. Les annotations peuvent être stockées dans une Base de Données Relationnelles, une Base de Données NoSQL Orientées Graphe comme Neo4j [12] ou dans un Triple Store. L'infrastructure de persistance multi-paradigmes, combinant les propriétés de chaque système, permet de couvrir les attentes en termes de stockage, d'interrogation et de traitement des données scientifiques.

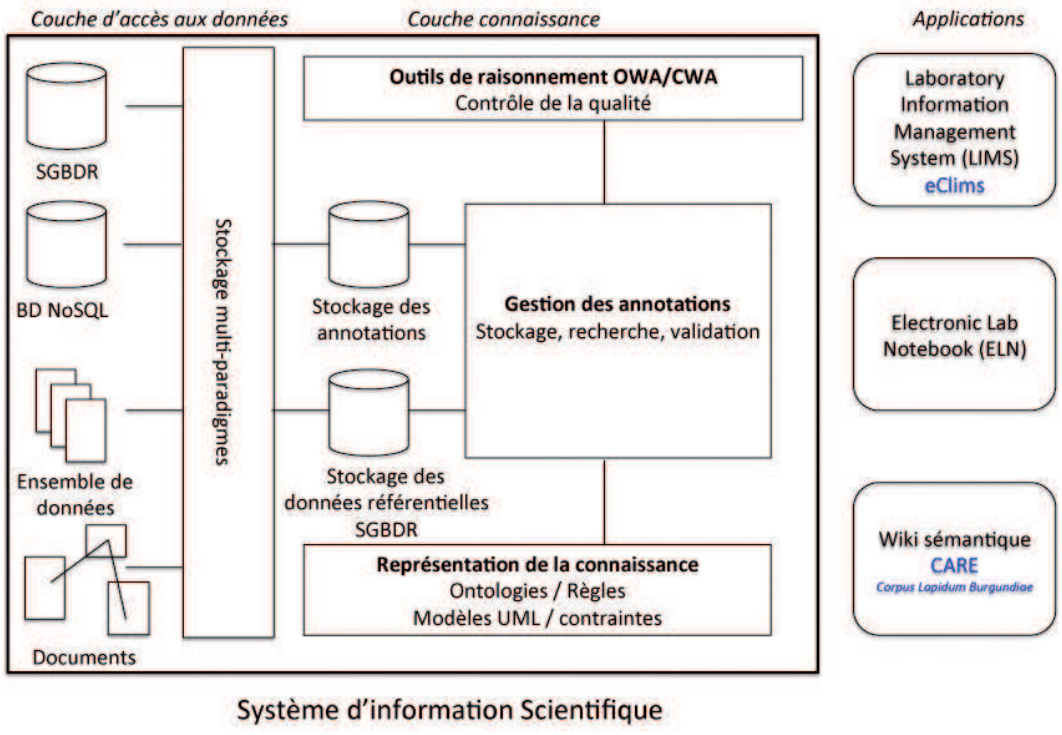


FIGURE 1.3 – Notre proposition pour les Systèmes d'Information Scientifique

11. Site Web : http://www.mongodb.org/
12. Site Web : http://neo4j.org

1.3.2 Représentation multi-paradigmes des connaissances

La couche dédiée à la gestion et au stockage de la connaissance comprend les ontologies, les données référentielles, les annotations sémantiques à bases ontologiques ainsi que des outils d'interrogation et de raisonnement (voir figure 1.3 - Couche connaissance).
Les ontologies représentent la connaissance du domaine. Les Systèmes d'Information Scientifique reposent non pas sur un domaine métier mais sur plusieurs, la connaissance du domaine est souvent construite à partir de plusieurs descriptions partielles issues du caractère multi-disciplinaire de la recherche. Chacun de ces domaines a son vocabulaire et ses pratiques qu'il convient de respecter. Par exemple, dans le domaine biomédical, on peut citer OBO RO [13] qui propose une liste exhaustive des relations qui relèvent de ce domaine, ou encore FBbi qui propose un vocabulaire concernant les techniques utilisées pour la préparation des échantillons, leur visualisation ainsi que les méthodes d'imagerie utilisées en recherche biomédicale. De telles ontologies de domaine servent de base pour construire une ontologie d'application qui contient les connaissances demandées par l'application.

Pour mettre en œuvre les liens entre les données modélisées dans des paradigmes différents et la connaissance du domaine nous employons une stratégie combinant les principes de gestion de données référentielles avec des annotations sémantiques à bases ontologiques.
Les données référentielles [DHM+08] sont fortement structurées et identifiables lors de la phase d'analyse avec un modèle UML. Elles sont reconnues par l'ensemble des partenaires de l'application, elles sont amenées à évoluer que très rarement. Nous utilisons une approche de type co-existence pour la gestion des données référentielles. Ainsi, les données importantes sont dupliquées dans un SGBDR, et les données qui ont besoin d'un modèle spécifique ou pour lesquelles il n'est pas possible de fixer un schéma, sont stockées séparément dans des systèmes de stockage spécifiques.
Des données complémentaires viennent enrichir les données référentielles. Ces données sont inhérentes à la démarche scientifique, elles s'ajoutent aux données référentielles pour répondre à une question apparaissant au cours d'une expérimentation. Les données complémentaires sont traitées sous forme d'annotations sémantiques qui fournissent un mécanisme d'extension de schéma de Base de Données. La sémantique des annotations est garantie par l'ontologie. Une annotation sémantique est une information supplémentaire qui est ajoutée à une donnée afin de la définir, de la décrire ou de la préciser, elle comporte un sujet, un prédicat et un objet modélisant respectivement la ressource annotée, la relation exprimée par l'annotation, et la ressource annotant (voir chapitre 3).

Les annotations offrent aussi une contextualisation des données, mais sans mécanisme de contrôle la qualité des annotations est incertaine. OWL et la logique de description $\mathcal{SROIQ}$ adhérent à l'hypothèse du monde ouvert où l'absence d'information ne peut pas être interprétée comme fausse mais seulement comme de l'information incomplète. Cette hypothèse est appropriée dans les Systèmes d'Information Scientifique pour les annotations sémantiques. Cependant, quand de nouvelles données doivent être intégrées dans le Système d'Information, il est préférable d'utiliser l'hypothèse du monde clos.

Pour répondre aux besoins des SIS, nous proposons une approche multi-paradigmes centrée autour des annotations sémantiques à bases ontologiques et des données référentielles. Notre approche apporte une grande extensibilité à deux niveaux : 1) au niveau intra-modèle car il est possible de rajouter des informations complémentaires sur les données existantes sans remettre en cause ou modifier les applications ; et 2) au niveau inter-modèles car il est possible de lier de manière transparente des données de différents modèles et stockées dans différents systèmes.

13. Site Web : http://www.obofoundry.org/ro/

Par ailleurs, l'utilisation des ontologies pour représenter la connaissance du domaine permet d'assurer la traçabilité, le contrôle de la qualité et l'interopérabilité sémantique.

1.4 Mise en œuvre de l'approche dans une démarche de construction de plate-forme

Les applications informatiques en eScience exploitent trois principaux niveaux de connaissance :

1. la connaissance générale du domaine souvent modélisée sous la forme de thésauri, d'ontologies (par exemple Gene Ontology en biologie ou CIDOC-CRM dans le domaine du patrimoine culturel), de standards (par exemple FuGE pour la génomique fonctionnelle) et de normes qualité ;
2. le savoir-faire exprimé par les processus métier mis en place. Généralement, le processus scientifique est le suivant : a) définition d'un protocole, b) production de données brutes, c) consolidation des données, d) analyse des données, e) archivage et publication ;
3. la base technique prenant souvent la forme d'une plate-forme technologique sophistiquée. Nous définissons une plate-forme comme un assemblage d'outils logiciels (SGBD, serveurs d'application, moteurs de *workflow*, serveurs Web, etc.) articulés autour de la description d'un fond sémantique commun et reconnu (ontologies, modèles, processus) et susceptible d'agir de façon organisée. Le choix de la plate-forme doit tenir compte des utilisateurs qui :
   - attendent d'une nouvelle application de nouveaux services mais ces mêmes utilisateurs ont des difficultés pour modifier leur pratique et leur mode de travail ;
   - viennent d'horizons variés, disposent de ressources informatiques non homogènes et peuvent avoir des niveaux de compétence très différents les uns des autres.

   Cependant les fonctionnalités offertes par les plate-formes ont de fortes conséquences sur les processus métiers et la structuration des connaissances.

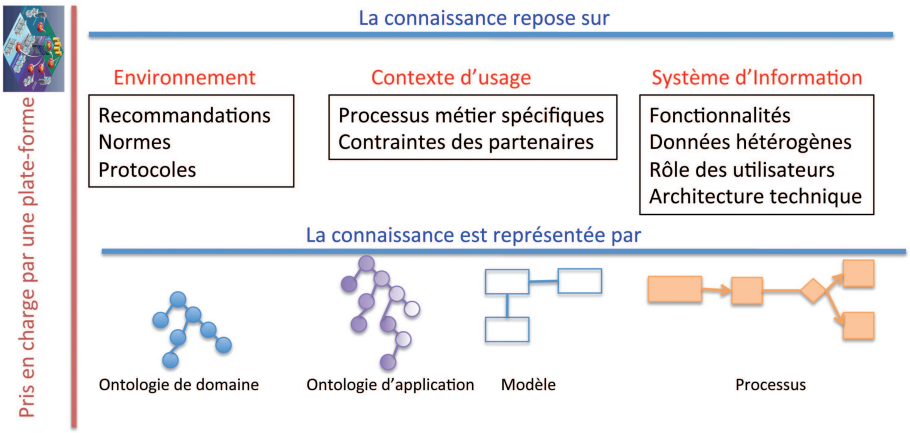


FIGURE 1.4 – Correspondances entre le Système d'Information, son utilisation et son environnement

La figure 1.4 reprend ces trois niveaux de connaissance et les met en perspective avec les différents contextes intervenant dans l'utilisation d'une application. La base technique d'une application correspond au Système d'Information. Un Système d'Information est constitué des éléments liés aux fonctionnalités qu'il propose, aux données qu'il contient, aux rôles dédiés aux

utilisateurs et à l'architecture complexe de services. La connaissance représente l'environnement pour lequel le Système d'Information a été créé, il formalise les contraintes métier et les techniques utilisées. Entre ces deux contextes, le contexte d'usage représente le savoir-faire des acteurs d'un domaine, il prend en compte les objectifs recherchés, les contraintes imposées par les différents utilisateurs mais aussi l'environnement de travail. Les ontologies de domaine sont alors spécialisées en ontologie d'application, les processus sont adaptés à l'usage que l'on veut en faire. Le contexte d'usage peut être vu comme un intermédiaire entre le contexte lié au Système d'Information et celui lié à la connaissance.

Nous avons alors pu dégager la démarche de construction de plate-formes suivante :
- étude des fonctionnalités nécessaires aux utilisateurs et des connaissances générales disponibles ;
- définition d'une organisation des termes du domaine dans une ontologie ;
- définition d'un modèle UML pour déterminer l'étendue du système en terme de données référentielles ;
- mis en place d'un couplage entre le modèle et certaines parties de l'ontologie (en particulier pour définir les domaines de valeurs de certains attributs) ;
- mise en place du modèle formel d'annotation pour gérer l'extensibilité ;
- choix d'une plate-forme correspondant aux attentes des utilisateurs : la plate-forme doit proposer un certain nombre d'outils logiciels combinables entre eux pour assurer le passage du système à l'état opérationnel. Les outils mis en place doivent contrôler précisément les modifications induites par les nouvelles connaissances sur les connaissances déjà intégrées au système. Cela nécessite en particulier de vérifier que la cohérence de l'ensemble des connaissances est préservée et que les données satisfont bien à toutes les contraintes énoncées.

Nous avons validé notre approche avec deux applications dans des domaines d'étude différents : le projet ANR CARE dans le domaine de l'archéologie et le projet *e-Clims* dans le domaine de la protéomique clinique. Le chapitre 2 présente la représentation des connaissances (ontologies et modèles UML) pour ces deux applications. Les chapitres 4 et 5 décrivent leur implémentation :
- pour le projet ANR CARE, les archéologues travaillant avec des documents, nous avons développé un wiki sémantique. Le choix d'un wiki comme plate-forme applicative permet de respecter la méthodologie de travail des archéologues. Les verrous technologiques traitent du contrôle de la qualité des relations entre les documents (données) et les ontologies ainsi que la recherche avancée d'information ;
- pour le projet *e-Clims*, le besoin d'un *workflow* scientifique nous a porté vers le développement d'un LIMS. Il met en place une Base de Données annotées pour traiter l'extensibilité de schéma et un processus de contrôle de la qualité de données lors de leur importation dans le Système d'Information Scientifique. Ce processus est construit autour d'ontologies et de modèles exécutables.

Le choix de ces deux plate-formes a permis aux utilisateurs de ces deux domaines de conserver leur mode de travail et obtenir de nouvelles fonctionnalités.

Organisation du document

La figure 1.5 montre l'organisation du mémoire en mettant en avant les grandes lignes directrices qui sont :

1. la compréhension d'un domaine qui conduit à la construction de **modèles conceptuels** offrant un cadre pour la sémantique et de **modèles exécutables** pour implémenter l'application. Nous proposons un **couplage modèles UML/ontologie** : un modèle UML représente la structure set les contraintes sur les données, une ontologie descriptive contient le vocabulaire et la définition des concepts et relations du domaine ;
2. la prise en considération des besoins inhérents aux Systèmes d'Information Scientifique (exigence d'extensibilité de la structure des données et caractère pluridisciplinaire de la recherche) repose sur une approche formelle basée sur un unique paradigme, l'**annotation sémantique** ;
3. la **validation des solutions proposées** avec deux implémentations qui ont exigé une adéquation entre la plate-forme choisie et le domaine à modéliser.

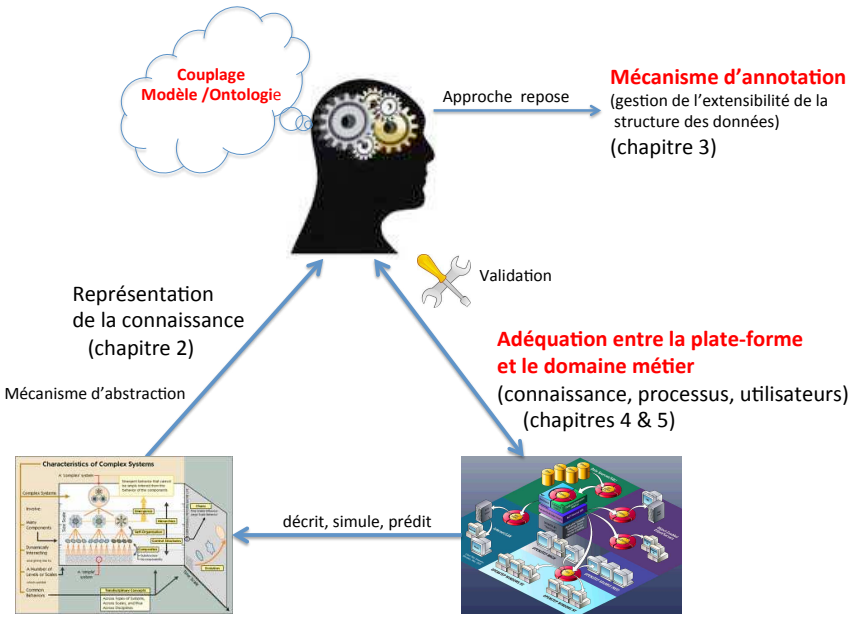


FIGURE 1.5 – Organisation du document

Chapitre 2 : représentation des connaissances Notre mode de représentation des connaissances combine à la fois une démarche descendante et une démarche ascendante.

La démarche descendante permet de :

- modéliser de façon conceptuelle l'application à l'aide d'une combinaison de modèles UML et d'ontologies,
- gérer le raffinement fonctionnel jusqu'au modèle exécutable et
- prendre en compte les préférences de l'utilisateur en vue de choisir une plate-forme répondant à ses contraintes.

La démarche ascendante projette les caractéristiques de la plate-forme sur le niveau fonctionnel.

Chapitre 3 : annotation Ce chapitre donne une définition formelle de l'annotation et de son expressivité. L'utilisation du modèle d'annotation est décrit dans deux plate-formes applicatives, le modèle répondant à un besoin propre à chaque application. Les bases formelles des technologies du Web Sémantique nous permettent de développer des mécanismes de contrôle des annotations, de qualité de données, ou de recommandation. Dans ce cadre, nous développons, par analogie avec la sémantique des langages de programmation, trois sémantiques pour le processus d'annotation (axiomatique, dénotationelle, opérationnelle) permettant des raisonnements OWA, CWA (*open/closed world assumption*).

Chapitres 4 et 5 : validation dans deux plate-formes La suite de ce travail a été la validation des idées des chapitres 2 et 3 dans deux plate-formes.

Chapitre 4 : Wiki Sémantique *WikiBridge* est un wiki sémantique développé pour le domaine des Sciences Humaines et Sociales dans le cadre de l'élaboration d'un corpus en archéologie. Cette plate-forme met en avant les aspects sémantiques *via* un assistant d'annotation s'appuyant sur une ontologie d'application. Les outils mis en jeu reposent sur les langages OWL, RDF, SPARQL et utilisent comme composants un wiki, un triple store et un raisonneur.

Chapitre 5 : LIMS Le chapitre 5 présente *eClims* un module de LIMS dans le domaine des Sciences du Vivant dans le cadre d'un projet pour la protéomique clinique. Cette plate-forme prend en charge les aspects extensibilité de la structure des données et qualité des données. Les outils mis en jeu reposent sur les langages OWL, SQL et Java et utilisent comme composants un SGBDR et des frameworks J2EE (Spring, Hibernate, GWT).

Le dernier chapitre conclut avec 1) quelques prolongements possibles dans le travail sur les annotations et 2) une extension du travail d'annotation par la fragmentation des annotations stockées dans une Base de Données Orientées Graphe.

Chapitre 2

Des modèles conceptuels aux modèles exécutables pour représenter les connaissances

« Theoretical modeling, and engagement with computing realities, are synergistic ingredients of both past and future progress. »
Turing's Titanic Machine ? S. Barry Cooper, Communications of the ACM, Vol.55, n^{o}3, Mars 2012

Sommaire

Dans ce chapitre, je donne une vue d'ensemble sur les problématiques liées à la modélisation et à la représentation des connaissances en présentant deux formes de modélisation : l'abstraction par conceptualisation où plusieurs couches de description sont ordonnées en fonction de leur

précision et l'abstraction par projection où des descriptions orthogonales de même niveau sont traitées de façon cohérente. Dans la suite du chapitre, je privilégie l'abstraction par projection. Je montre que la connaissance dans des domaines complexes et évolutifs ne peut pas être stabilisée dans les phases initiales de modélisation, mais que la complémentarité des modèles UML et des ontologies permet de travailler avec une fiabilité acceptable. Il s'agit d'associer aux données définies par un modèle UML (dites données référentielles) des annotations sémantiques basées sur une ontologie. J'illustre cette complémentarité dans deux projets : le projet ANR CARE relatif à la gestion d'un corpus des édifices religieux antérieurs à l'an Mil ; le projet *eClims* de gestion de données clinique en protéomique. En conclusion du chapitre, je compare les modèles UML et les ontologies.

2.1 Donnée, Information, Connaissance

Une **donnée** est le résultat d'observations présentée sous une forme conventionnelle en vue d'être exploitée indépendamment de sa source. La donnée est représentée par une valeur (quantitative ou qualitative) sans contexte ; elle devient information dès lors qu'elle est mise en contexte et devient porteuse de sens pour l'esprit humain qui s'en empare. L'**information** est l'interprétation des données.
Le terme **connaissance** est l'objet de nombreuses définitions provenant de diverses disciplines : philosophie, psychologie, sciences de l'information, sciences humaines et sociales, etc. La connaissance est définie par les philosophes comme un acte de pensée. Elle est liée à une représentation du monde (modèle), partagée, acquise par l'expérience dans le monde réel duquel on ne peut dissocier une dimension sociale. On appelle connaissance *a priori* une connaissance qui peut être prouvée sans référence aucune à l'expérience, par exemple, l'espace et le temps. Une loi scientifique constitue une connaissance qui est vraie par hypothèse : elle est une construction de l'esprit du scientifique pour expliquer le monde et elle peut être réfutée lorsque ses prédictions sont mises en défaut par l'expérience. La science apparaît comme une construction dynamique de modèles du monde.

En informatique, la représentation de la connaissance a pris des formes variées au long du temps : de la simple liste de termes, des modèles conceptuels et jusqu'à la notion d'ontologies. Plus la possibilité de représenter la connaissance est importante, plus les capacités de raisonnement sont grandes. La figure 2.1 présente une vue synthétique des langages de représentation de la connaissance et des capacités de raisonnement associées, sur un axe symbolisant une représentation formelle de plus en plus forte. Almeida et al. [ARF11] reprennent ce spectre et repositionnent les différents éléments les uns par rapport aux autres selon leurs utilisations (utilisation liée aux systèmes Web, liée à l'organisation de l'information dans des documents, etc.).

2.2 Comprendre un domaine

Marc Linster dans [Lin92] montre que l'interaction entre les experts du domaine, les spécialistes en ingénierie de représentation des connaissances et les outils créent la connaissance. Il a montré que le processus pour élaborer un système à base de connaissances est un processus itératif de construction de modèles qui comprend :

1. un processus de discussion entre les spécialistes en ingénierie de représentation des connaissances et les experts du domaine et
2. la construction à la fois de **modèles conceptuels** dont l'objectif est de fournir un cadre général et abstrait pour la sémantique en représentant la connaissance du domaine, sup-

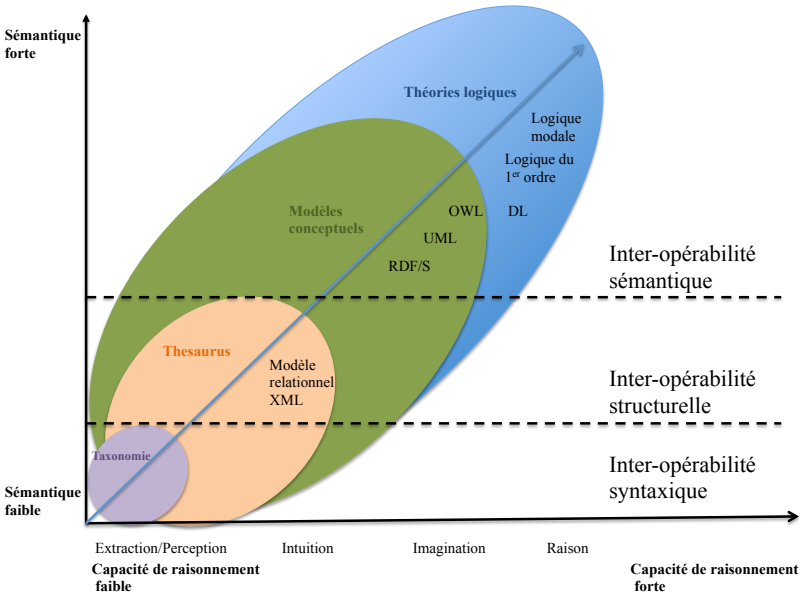


FIGURE 2.1 – Spectre des langages de représentation et des capacités de raisonnement (d'après [DOS05])

port du dialogue entre humains et de **modèles exécutables** construits pour implémenter l'application.

De même, Frantz dans [Fra95] décrit le processus de modélisation sous la forme de deux traductions successives : le monde réel est traduit en un modèle conceptuel par un mécanisme d'abstraction ; le modèle conceptuel est ensuite traduit en un modèle de simulation par un mécanisme d'implémentation [1]. Frantz propose ensuite une taxonomie des méthodes d'abstraction : 1) abstraction de la frontière du modèle en changeant les facteurs pris en compte dans le monde réel ; 2) abstraction des comportements en éliminant les changements d'états sans lien avec l'objectif de la modélisation et 3) abstraction par changement de la forme du modèle lui-même.
Le modèle a donc un double rôle : interprétatif à la lecture par les humains et un cadre formel pour une interprétation par la machine.

2.2.1 Modélisation conceptuelle comme cadre sémantique

Nous retenons comme définition pour le modèle conceptuel celle donnée par Bachimont [Bac04] : « *Un modèle conceptuel exprime les connaissances d'un domaine relatives à une tâche dans un langage de modélisation ... Une fois rendu opérationnel dans un système, il constitue un outil pour agir sur le monde, qui doit être pertinent en usage* ». Pour Gruber [Gru93], « *A conceptualisation is an abstract, simplified view of the world that we wish to represent for some purpose* ». La conceptualisation est donc un raisonnement qui consiste à identifier par abstraction les concepts essentiels du domaine de connaissances et à établir les relations entre ces concepts. Cette abstraction peut inclure des aspects structurels mais aussi comportementaux et elle n'est pas attachée à une technologie informatique.

1. [Fra95] ne traite que des modèles de simulation.

Aspects O.O.	Vues des systèmes multi-vues	Diagrammes UML2.0
Statique	Objet	Diagramme d'objet
	Physique	Diagrammes de composants, de déploiement de package
	Structurelle	Diagrammes de classes, des structures composites
Dynamique	Externe	Diagramme Use Case
	Comportementale	Diagrammes d'activités, de séquences, de description des interactions, de timing, de communication, états-transitions

Figure 2.2 – UML2.0 : aspects, vues et diagrammes

Wand et Weber [WW02] ont défini les objectifs auxquels un modèle conceptuel doit répondre : 1) permettre la communication entre développeurs et utilisateurs ; 2) aider les analystes à comprendre le domaine à modéliser ; 3) fournir de la matière au processus de conception ; 4) documenter l'application. La difficulté de la construction d'un modèle conceptuel découle de plusieurs facteurs : identifier les connaissances qu'il doit représenter, mais aussi les rôles prioritaires que l'on veut lui faire jouer en connaissant les contraintes qu'imposent ces rôles. Aussenac-Gilles [AG05] a identifié quatre rôles que peut jouer un modèle conceptuel : 1) *langage de méta-niveau permettant de s'adapter à différentes applications* : il correspond à la volonté d'adapter ou de spécialiser des primitives conceptuelles de haut niveau ; 2) *langage partagé par les acteurs de la modélisation* : en particulier entre les experts du domaine et les ingénieurs en représentation des connaissances. Le modèle sert alors à spécifier comment le système va fonctionner et sur quelle base de connaissances ; 3) *cadre d'expression des connaissances* : le modèle rassemble des connaissances identifiées comme pertinentes pour une pratique, un usage ou la réalisation d'une tâche. Il est le support pour expliciter et assurer la vérification syntaxique des connaissances ; 4) *trait d'union entre connaissances mises en œuvre et connaissances calculables* : le modèle ayant pour vocation à déboucher sur un système opérationnel, il doit être au final interprétable par la machine. Suivant les approches, le passage à un modèle exécutable peut suivre un continuum ou bien correspondre à une traduction des éléments conceptuels en nouvelles primitives.
De nombreuses propositions s'articulent autour de modèles multi-perspectives [Ren00]. Dans le cadre d'une démarche de modélisation d'un domaine d'application, UML (*Unified Modeling Language*) est largement utilisé. Par ailleurs, les ontologies fournissent un moyen efficace de représentation de la connaissance d'un domaine. Dans la suite, nous détaillons le langage UML et les ontologies.

Le langage UML2.0 est articulé autour de treize diagrammes. Ces diagrammes constituent des vues convergentes du Système d'Information. Les vues relèvent des aspects statiques ou dynamiques (voir figure 2.2).

- La **vue externe** (diagramme Use Case) décrit le système sous forme d'un ensemble de fonctionnalités (appelées *use cases*) qui sont offertes à différents types d'utilisateur (appelés *acteurs*).
- La **vue structurelle** (diagramme de classes et diagramme des structures composites) correspond à la description statique du système. Le diagramme de classes rend compte des entités en jeu ainsi que des relations "sémantiquement fortes" entre ces entités. Le diagramme des structures composites permet de mettre en valeur les interfaces entre classes.
- La **vue comportementale** (diagrammes de séquence, d'activité, de communication, de description des interactions, de timing et états-transitions) décrit la dynamique du système. Ces

diagrammes établissent (soit pour une classe isolée, soit pour plusieurs classes en interaction) les protocoles à travers lesquels les objets communiquent et interagissent.
- La **vue physique** (diagrammes de package, des composants, de déploiement) décrit le système comme un ensemble de composants (packages de données, code, documentation, etc.) positionnés sur des machines reliées par les liens physiques d'un réseau.
- La **vue objet** (diagramme d'objets) présente le système en terme d'objets reliés. C'est à la fois un "instantané" du système et une instanciation du diagramme de classes.

Evermann et al. [EW05] ont proposé près de quatre-vingts règles, basés sur l'ontologie de Bunge [Bun77], pour utiliser les diagrammes UML dans une modélisation conceptuelle. Nous donnons en exemple deux règles essentielles : 1) seule une chose réelle peut être modélisée par une classe ; 2) une classe-association ne peut pas représenter une chose réelle et donc avoir des opérations. Lu et al. [LP05] ont implémenté ces règles en leur attribuant des priorités dans l'outil *open source* ArgoUML [2].

Les ontologies sont aussi des modèles conceptuels multi-perspectives. Elles peuvent être classifiées selon l'objet de conceptualisation (voir figure 2.3) :
- l'**ontologie de représentation de connaissances** regroupe les concepts utilisés dans la formalisation des connaissances. Par exemple, l'ontologie de frames d'Ontolingua [3] définit de façon formelle les concepts utilisés dans les langages à base de frames ;
- l'**ontologie de haut niveau** est une ontologie générale. Ce type d'ontologie étudie les concepts de haute abstraction, valables dans différents domaines, tels que entité, évènement, temps, espace, relation, propriété, processus, action. Par exemple, BFO (*Basic Formal Ontology*) est une ontologie qui entre dans cette catégorie, elle a été développée en lien avec la fonderie OBO (*Open Biomedical Ontology* [4]). BFO pose les bases de descriptions spatio-temporelles : elle considère les trois dimensions géométriques et une dimension temporelle. La distinction des concepts est faite suivant l'axe temporel en fonction de la façon dont les entités s'inscrivent dans le temps [GS04] ;
- l'**ontologie générique** (ou méta-ontologie ou *core ontology*) regroupe des connaissances moins abstraites que dans les ontologies de haut niveau mais assez générales néanmoins pour être réutilisées par différents domaines. On peut citer l'ontologie méréologique de Borst [Bor97] qui définit la relation *partie-de* et ses propriétés ;
- l'**ontologie de tâche** est utilisée pour conceptualiser des tâches spécifiques comme la planification, le diagnostic ou la configuration. Ces ontologies fournissent un ensemble de termes génériques (par exemple plan, objectif, assigner, classer, sélectionner) pour décrire comment résoudre un type de problème [MVM95] ;
- l'**ontologie de domaine** décrit un vocabulaire en relation avec un domaine spécifique, elle représente les principales activités, théories et principes de base du domaine en question. Elle est réutilisable par plusieurs applications qui travaillent sur le domaine pour lequel l'ontologie a été créée car elle a été conçue de façon aussi indépendante que possible du type de manipulations qui vont être opérées sur ses connaissances. Par exemple, dans le domaine biomédical, on peut citer OBO RO [5] qui propose une liste exhaustive des relations qui relèvent de ce domaine, ou encore FBbi qui propose un vocabulaire concernant les techniques utilisées dans le domaine biomédical pour la préparation des échantillons, leur visualisation ainsi que les méthodes d'imagerie utilisées en recherche biomédicale ;
- l'**ontologie d'application** est la forme la plus spécifique, elle contient les connaissances exigées par une application particulière et elle n'est pas réutilisable. Elle peut en outre inclure

2. http://argouml.tigris.org
3. http://www-ksl.stanford.edu/knowledge-sharing/ontologies/html/frame-ontology/index.html
4. OBO foundry http://obofoundry.org/
5. Site Web : http://www.obofoundry.org/ro/

une ontologie de domaine.

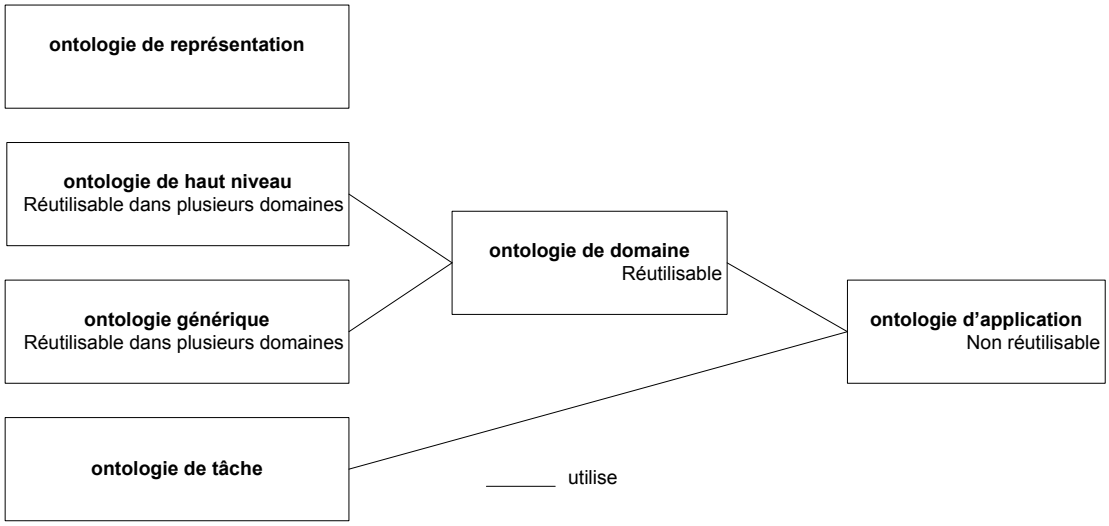


Figure 2.3 – Classification des ontologies selon l'objet de conceptualisation

De plus, un modèle conceptuel peut être complètement informel exprimé en langage naturel par exemple, semi-formel combinant langage naturel et diagrammes UML ou bien formel en utilisant par exemple la logique de description [BCM+03].

2.2.2 Modèle exécutable pour représenter l'application

Si un modèle conceptuel fournit un lien entre la perception du monde par les humains et la façon dont le monde est représenté dans un système formel, un modèle exécutable est la plus haute couche d'abstraction basée sur les langages d'implémentation [JZM08]. Un modèle exécutable peut être directement traduit dans un langage de programmation. Par exemple, *executable* UML [MB02] restreint le diagramme de classes en supprimant l'utilisation des associations de composition et d'agrégation, des attributs multivalués, si aucun attribut de la classe ne peut jouer le rôle d'identifiant, un attribut supplémentaire est ajouté permettant ainsi une traduction automatique vers le modèle relationnel par exemple. Pour les ontologies, les différents axiomes, issus de la Logique de Description choisie, ne sont pas anodins en termes de connaissances représentables, de décidabilité et de calculabilité, RDF/S est un langage qui rend l'information traitable par des programmes.

2.3 Modèle ou modèles ?

La complexification des Systèmes d'Information est principalement due au fait que les systèmes sont de plus en plus dépendants de leur environnement (sécurité, distribution) et que la cohérence des informations est de plus en plus liée aux inter-relations entre les différentes parties de l'information qu'à la structure propre à chacune de ces parties. Les spécialistes en ingénierie de représentation des connaissances opèrent donc par une suite de descriptions. Le processus unifié qui utilise UML comme notation, est basé sur la construction successive de plusieurs modèles de plus en plus concrets et de plus en plus détaillés. L'IDM [6] (Ingénierie Dirigée par

6. L'IDM est appelé MDE en anglais (Model Driven Engineering).

les Modèles) va plus loin et place les modèles –sous toutes leurs formes– au cœur du cycle de vie des applications. Pour la construction d'une ontologie, Fürst [Für02] passe par un modèle conceptuel, une ontologie semi-formelle puis une ontologie formalisée. Il indique que ce passage peut être fait de façon ascendante (des connaissances vers la description formelle) ou de façon descendante (par instanciation de la description formelle).
La modélisation d'un Système d'Information prend essentiellement deux formes :

1. **l'abstraction par conceptualisation** consiste à utiliser plusieurs "couches" de description du Système d'Information. Les "couches" étant ordonnées en fonction de leur précision et du niveau de détail pris en compte ;
2. **l'abstraction par projection** consiste à utiliser plusieurs descriptions orthogonales (c'est-à-dire si possible indépendantes les unes des autres) mais placées à un même niveau de conceptualisation et traitées de façon cohérente.

2.3.1 Abstraction par conceptualisation : raffinement fonctionnel

J'ai appliqué la conceptualisation par abstraction dans deux domaines :
– les ontologies en proposant une ontologie spatiale multi-couches [BLS+00]. Chaque couche de l'ontologie propose un ensemble de fonctionnalités spatiales. L'objectif est de prendre en compte le fait qu'un même objet spatial peut être associé à différents jeux de fonctionnalités dans différents systèmes. Par exemple, une rue peut être vue comme une surface par le service de voirie et comme un lien dans un graphe par le service en charge de la circulation urbaine.
– lors du co-encadrement de la thèse de Jean-Claude Simon « Une approche liée à la préservation des propriétés transversales pour construire une offre de services web d'un éco-système d'entreprises » soutenu en 2008. Je détaille dans la suite ce travail.

Représentation des connaissances dans une architecture orientée Services Web
Cette thèse voit un système complexe sous la forme de blocs fonctionnels très abstraits au départ où la combinaison des blocs ne suit pas une logique de construction incrémentale à partir de blocs simples mais une construction par raffinement. On obtient à l'arrivée une combinaison par imbrication de blocs spécialisés dans des blocs génériques. Ce raffinement peut être itéré autant que de besoin. Une première couche décrit le domaine du système complexe et l'application ; une deuxième couche est la résultante du raffinement par métiers du modèle applicatif décrit et une troisième couche caractérise ce dernier à travers un raffinement par entreprises (voir figure 2.4).

Niveau fonctionnel

Le niveau fonctionnel décrit les fonctionnalités offertes par l'application à travers un *workflow* d'activités progressivement raffiné.
– **Description du domaine** Ce niveau consiste d'abord à décrire les fonctionnalités accomplies par l'application à travers un *workflow* d'interactions entre blocs de haut niveau. Ensuite les différents acteurs impliqués sont identifiés pour trouver les métiers associés. Enfin, ces métiers sont projetés sur les entreprises impliquées dans le Système d'Information. La stratégie orientée-domaine que nous avons choisie pour définir les blocs fonctionnels autour desquels va s'articuler le Système d'Information consiste à choisir tout d'abord une description « gros-grain » qui présente une vision hautement simplifiée du cœur de métier. Il est indispensable de nous placer dans un cadre reconnu par la communauté qui travaille sur la décomposition métier des processus d'entreprises, ainsi nous avons choisi le travail de Van Der Aalst et al. [vdAtHKB03]. Ils ont proposé une liste de vingt patterns de *workflows* qui permettent de traiter toute activité en entreprise. Ces patterns sont classés en six grandes familles : 1)

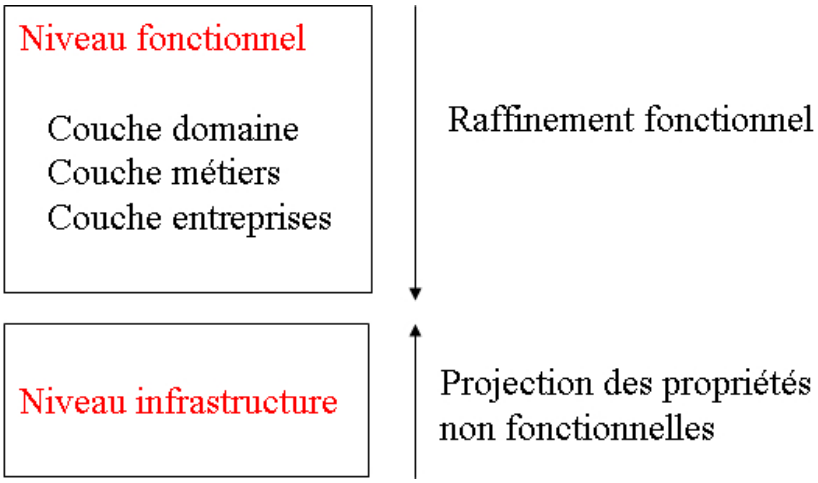


Figure 2.4 – Notre approche de représentation de connaissances par raffinement fonctionnel

contrôle élémentaire ; 2) branchements et synchronisation ; 3) structures ; 4) contrôle d'instances multiples ; 5) contrôle basé sur des états pour prendre en compte les facteurs externes et 6) annulation.
Dans cette couche, les modèles de description du *workflow* d'interactions entre blocs abstraits peuvent être décrits avec le *Business Process Modeling Notation* (BPMN), le diagramme d'activités UML ou n'importe quel autre langage de représentation de *workflow*. BPMN et UML font l'objet d'un consensus suffisant pour émettre l'hypothèse qu'une transformation sera possible, à partir de n'importe quel outil de description des activités métier. Par ailleurs, White [Whi04] propose des transformations de modèles entre les patterns de Van Der Aalst, le diagramme d'activités d'UML et la notation BPMN.

– **Description métiers** Dans cette étape, l'application cible est affinée selon la structure métiers. L'objectif est de construire des couches contenant des « sous-blocs ». Les blocs fonctionnels de haut niveau sont projetés sur des axes métiers. Le diagramme Use Case UML met en avant les acteurs et ainsi trouve les métiers concernés. Ce travail doit être fait à haut niveau d'abstraction et prendre en compte à la fois les différences de métier et celles d'appartenance à des structures différentes. Les métiers de certains acteurs dans deux entreprises différentes peuvent être identiques. Chaque bloc fonctionnel conceptuel est alors spécialisé suivant le métier concerné. Nous utilisons une approche similaire à celle des stratégies *« block-oriented »* pour le *process-mining* [Sch00] afin de construire le *workflow* sous forme de blocs imbriqués.
– **Description des entreprises** Nous instancions alors les blocs fonctionnels raffinés à la particularité de chaque entreprise. À l'issue de cette phase de travail nous disposons donc d'un ensemble d'acteurs et d'un ensemble de blocs d'activité définis sous forme de *workflows* par des activités et des transitions entre activités.

Notre méthode respecte les sept règles édictées par Mendling et al. [MRvdA10]. Ce mécanisme orienté métier de description et de projection du *workflow* d'activités peut être itéré jusqu'à obtention d'activités implémentables.
Cette méthodologie s'inscrit dans les préconisations données par un rapport européen sur l'interopérabilité des entreprises [RC06]. Ce dernier a mis en avant plusieurs défis dont 1) la mise à disposition de *modular software building blocks* correspondant à une décentralisation fonctionnelle de leurs activités métier et 2) la capacité à garantir certaines propriétés dans le cadre de la collaboration entre entreprises.

Niveau infrastructure

Ce niveau est composé de l'ensemble des services et de l'architecture matérielle.

L'ensemble des services réalise l'application raffinée au moyen des services disponibles. Il est composé de deux couches, la couche *Generic Business Process* et la couche instances. La première décrit l'application à travers une interaction de services abstraits, qui correspondent aux blocs activités déterminés au niveau fonctionnel pour décrire les processus métier cibles. Elle est obtenue à partir des résultats de l'étape de définition de l'architecture. La deuxième couche englobe les réalisations possibles du diagramme de services abstraits en utilisant les services fournis dans le Système d'Information. Ils implémentent une interface soumise aux contraintes de la réalisation.
Enfin, l'architecture matérielle décrit les caractéristiques du réseau.

Cette approche par une infrastructure à base de Services Web a été mise en œuvre dans deux thèses. L'objectif de la thèse de Jean-Claude Simon [Sim08] que j'ai co-encadrée était de traiter les propriétés non fonctionnelles telles que la sécurité et la traçabilité. La convergence métier/plate-forme répond au besoin de compléter l'organisation descendante pilotée par les aspects métier, par une organisation ascendante des Services Web et la réalité du terrain. Un des points d'achoppement est lié en effet à l'utilisation conjointe de plusieurs Systèmes d'Information et à la nécessité de limiter les ingérences d'un Système d'Information à l'autre. Techniquement, ceci a été traité par le biais de transactions incluant des mécanismes de compensation. Les transactions appellent les Services Web en garantissant à une échelle multi-systèmes la cohérence des actions effectuées. Les mécanismes de compensation permettent de restaurer un état sémantiquement cohérent lors d'invalidations de transactions. Les choix faits au niveau des Services Web consistent alors à laisser à chaque Système d'Information, la responsabilité de ses propres mécanismes de compensation sur ses Services Web.
L'objectif de la thèse de Elie Abi-Lahoud [AL10] est d'instancier cette démarche en formalisant le choix des services, selon un ensemble de contraintes données comme un problème d'optimisation de coûts.

2.3.2 Abstraction par projection : couplage Modèle / Ontologie

Les modèles UML et les ontologies présentent à la fois des similitudes et des différences [SMJ02]. Comme les modèles UML, les ontologies conceptualisent l'univers du discours au moyen de classes qui peuvent être hiérarchisés et associées à des propriétés. Les principes de base sont donc similaires. Par contre, des différences ont été identifiées entre ces deux types de modèles :
- objectif de modélisation : UML décrit l'information qui doit être représentée dans un Système d'Information particulier. Au contraire les ontologies décrivent les concepts d'un domaine indépendamment de tout Système d'Information dans lequel l'ontologie pourra être utilisée ;
- consensualité : une ontologie permet de représenter les concepts de la même façon dans tous les Systèmes d'Information d'une "communauté". Au contraire un modèle UML est spécifique à une application donnée ;
- identification des concepts : dans une ontologie les concepts et leurs propriétés sont identifiés par des URI, ce qui leur permet d'être référencés à partir de n'importe quel modèle. Au contraire la conceptualisation dans un modèle UML ne permet pas aux éléments du modèle d'être référencés à l'extérieur de l'application ;
- souplesse de description : toutes les instances des classes d'une ontologie peuvent ne pas initialiser les mêmes propriétés.

Gruber [Gru93] souligne le lien très étroit qui relie structures conceptuelles et ontologies : alors qu'un schéma conceptuel définit les relations sur des données, une ontologie définit les termes avec lesquels on représente la connaissance.

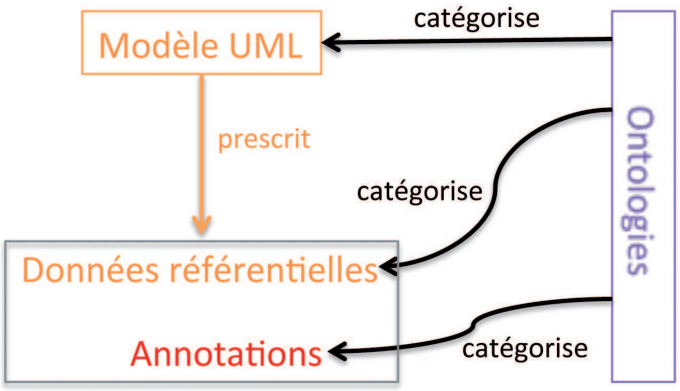


Figure 2.5 – Notre approche de représentation des connaissances par couplage Modèle / Ontologie

En résumé, un modèle UML représente la structure et les contraintes sur les données d'une application et apporte le caractère prescriptif, alors que une ontologie contient le vocabulaire et la définition des concepts et de leur relation et apporte le caractère descriptif. Nous proposons de coupler ces deux modèles pour représenter la connaissance d'un Système d'Information Scientifique. Dans le chapitre précédent, nous avons montré qu'une des caractéristique des Systèmes d'Information Scientifique est la difficulté à modéliser de façon exhaustive une étude car la nature de la recherche peut changer. La découverte de nouvelles connaissances scientifiques demande un mécanisme d'extensibilité de la structure. Comme les modèles évoluent dans le temps, les schémas tels qu'ils sont proposés en Base de Données ne sont pas utiles. Nous proposons de suivre les conseils de la gestion des données référentielles [DHM+08] pour proposer un système de persistance multi-paradigmes. L'identification de ces données porte sur trois notions principales :

- le partage concerne des données utilisées par différents systèmes ou par différents blocs fonctionnels au sein d'un même système ;
- la stabilité concerne des données étant rarement amenées à évoluer ;
- la fréquence de consultation concerne des données, servant de pivot à de nombreux processus et consultées fréquemment.

Nous proposons de les modéliser avec un modèle UML et de les compléter par des données annotées (voir figure 2.5). L'annotation a pour but d'exprimer la "sémantique" à l'intérieur d'une ressource afin d'en améliorer sa compréhension, sa recherche, sa réutilisation par d'autres utilisateurs. L'annotation se base sur une ontologie. Pour cela nous nous appuyons sur :

- la définition de Sowa des ontologies. John F. Sowa (http://www.jfsowa.com/ontology/) présente une ontologie comme un outil de catégorisation lié à un domaine d'intérêt vu par une personne ayant un objectif et à un langage de description : « *The subject of ontology is the study of the categories of things that exist or may exist in some domain. The product of such a study, called an ontology, is a catalog of the types of things that are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D* ».
- les travaux de Elvesæter et al. [EHBN06] qui prétendent que, pour la plupart des domaines d'application, des ontologies pré-existent : « *In each of these business domains [supply chain management, collaborative product development, e-procurement, portfolio management] we find domain-specific dictionaries, thesauri, nomenclatures, ...* ». De plus, ils estiment qu'une

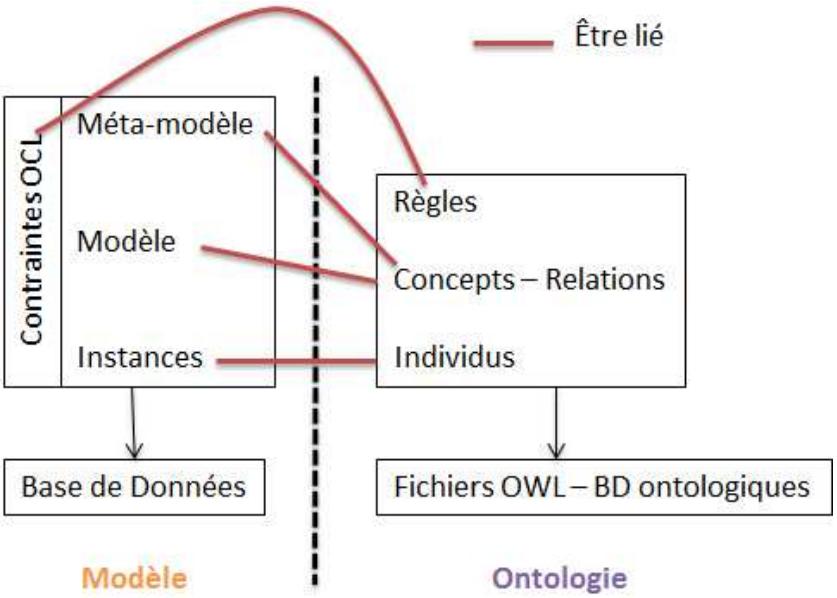


FIGURE 2.6 – Détail du couplage Modèle / Ontologie

ontologie offre une possibilité de catégorisation [7] dès le processus de modélisation : « *The models at the various levels may be semantically annotated using ontologies which help to achieve mutual understanding on all levels.* ».

Le couplage du modèle UML avec la représentation des connaissances *via* des ontologies prend les formes suivantes (voir figure 2.6) :

- les instances du modèle sont liées aux individus de la description ontologique, l'ontologie définit les domaines de valeurs de certains attributs ;
- les concepts et les relations de la description ontologique peuvent être liées aux classes et relations du modèle ;
- les concepts et les relations de la description ontologique qui touchent aux concepts de haut niveau (comme les concepts provenant des ontologies générales) sont liés aux constructeurs du modèle ;
- les règles de la description ontologique sont liées aux contraintes OCL exprimées à tous niveaux.

Notre approche tire le meilleur parti possible de la description du domaine en ayant recours à un double paradigme (modèle et ontologie) pour garantir au niveau sémantique l'exploitation de la connaissance du domaine.

La suite de ce mémoire présente mes travaux sur l'utilisation de données référentielles et données annotées dans deux domaines scientifiques.

7. Bessière et al. [BDL+98] appelle « ***catégorisation*** *l'opération consistant, à « projeter » sur le phénomène étudié la structure ensembliste du modèle* ».

2.4 Application au projet ANR CARE : gestion d'un corpus des édifices chrétiens antérieurs à l'an Mil

L'objectif du projet international CARE (*Corpus Architecturae Religiosae Europeae - IV-X saec.*) est la constitution d'un corpus des monuments chrétiens antérieurs à l'an Mil [CS08]. Il s'agit de recenser tous les édifices religieux et leurs évolutions entre le IV^e et le début du XI^e siècle. Plusieurs pays dont l'Italie, l'Espagne, la Tchéquie, la Slovaquie et la Croatie ont adhéré à ce projet. Il a commencé en France en 2008, après avoir été accepté par l'ANR (ANR-07-CORP-011), pour une durée de quatre ans. Le corpus français met l'accent sur les VII^e-VIII^e siècles et sur les décennies précédant ou suivant l'an Mil très riches en monuments. D'un point de vue organisationnel, le projet CARE prend la forme d'un réseau d'experts de différentes disciplines assurant l'alimentation du corpus et collaborant à son exploitation au moyen de travaux de recherche collectifs ou individuels.

2.4.1 Architecture générale du projet

Le projet ANR CARE comporte un volet archéologique et un volet informatique. Ces deux volets ont été conçus pour être menés de front et en étroite liaison, c'est-à-dire dans une démarche collaborative qui ne se limite pas à une prestation de service du type utilisateur-fournisseur.

Volet archéologique

Le volet archéologique vise au recensement, à la compilation et à l'analyse de la totalité des données disponibles dans les sources documentaires pour la période de référence (du IV^e au X^e siècle) et pour la majeure partie des pays européens. Il se décompose en deux tâches principales : la collecte des données (établissement et dépouillement du corpus, rédaction des fiches de dépouillement, intégration dans le corpus numérique) puis l'analyse et l'interprétation des données recueillies. La collecte et l'analyse des données porte sur environ 2 700 monuments en France.
Les données de terrain recueillies par les archéologues, complétées des sources (littéraires, d'archives, épigraphiques) font l'objet d'un dépouillement systématique. Chaque édifice ou groupe d'édifices fait l'objet d'une fiche de dépouillement. L'annexe C détaille cette fiche. Cette tâche de fond a été menée sur toute la durée du projet.
L'analyse et l'interprétation des données consistent à exploiter les fiches en liaison, le cas échéant, avec les spécialistes des autres disciplines. Cette tâche a notamment pour finalité la publication annotée des résultats du dépouillement sous forme d'ouvrages et d'une plate-forme numérique qui permet l'analyse spatio-temporelle du corpus.

Volet informatique

Le volet informatique couvre la conception et la réalisation d'outils spécifiques au projet : architecture de la plate-forme informatique, base de données, moteur de recherche sémantique. Nous avons pris en charge le pilotage et la coordination des activités de développement de la plate-forme informatique, support du travail de collecte, de structuration et d'annotation des fiches. L'objectif de cette plate-forme, fondée sur les technologies du Web 2.0 et du Web Sémantique, est de faciliter les processus d'interprétation des fiches de dépouillement et de recherche des ressources. Ces deux processus utilisent un mécanisme d'annotations. D'un point de vue technique, comme nous le détaillons dans le chapitre 4, la plate-forme est déployée sous la forme d'un **wiki sémantique**, *WikiBridge*. Le développement du wiki sémantique nécessite la construction d'une ontologie d'application propre à l'espace thématique et répondant aux

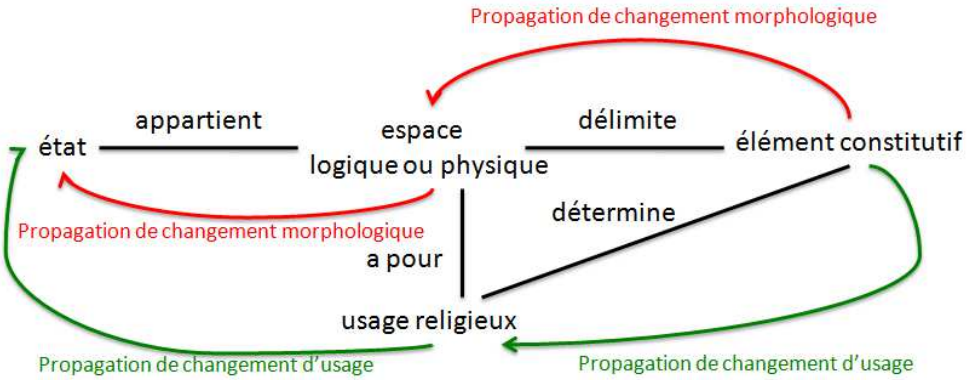


FIGURE 2.7 – Concepts saillants du corpus CARE

spécificités du projet : multi-expertises appliquées sur des ressources communes et relatives à un domaine partagé et multi-langues de part la dimension européenne du projet.
Le projet CARE présente toutes les caractéristiques des Systèmes d'Information Scientifique : collaboration, flexibilité, complexité et qualité des informations à gérer. En effet, ces informations sont disciplinaires, hétérogènes, floues, incertaines, incomplètes, contradictoires et régulièrement remises en cause. Le caractère hétérogène de l'information est dû à la multiplicité des sources (manuscrits, iconographies, cartographies, campagnes de fouilles contemporaines, etc.). L'imprécision de l'information est due au caractère vague ou approximatif des sources documentaires ou des techniques d'analyse utilisées. L'incertitude est liée à la validité d'une information. L'incomplétude est due aux absences d'informations ou à des informations lacunaires. De plus, les informations peuvent être interprétées par exemple le détail d'un chapiteau peut suffire pour le restituer dans sa globalité car il est identifiable comme un type décrit dans un traité d'architecture. D'une manière générale, les hypothèses deviennent multiples et parfois très éloignées les unes des autres.

Dans un premier temps, nous avons réalisé une modélisation conceptuelle afin de comprendre les attentes des archéologues. Quatre réunions ont eu lieu entre octobre et décembre 2008, époque où les archéologues n'avaient pas encore défini leur fiche de dépouillement. Nous présentons dans les sections suivantes, les concepts saillants du domaine, la modélisation conceptuelle que nous avons effectuée avec UML et la logique de description puis les modèles exécutables correspondants.

2.4.2 Mise en exergue des concepts saillants

Le corpus CARE est centré sur la notion d'édifices religieux et son objectif est de connaître les évolutions architecturales d'un édifice ou d'un groupe d'édifices. En effet les édifices n'ont en général pas une forme constante au cours du temps, il faut donc différencier les phases de changement de celles où l'édifice reste invariable. La phase de changement désigne des modifications qui peuvent avoir des causes humaine ou naturelle, par exemple un tremblement de terre peut être à l'origine de la destruction d'un édifice. Les évolutions d'un édifice sont nombreuses et de natures différentes : certaines concernent la totalité de l'édifice (construction, destruction, reconstruction, etc.), d'autres concernent seulement certains éléments constitutifs (déplacement, variation, dégradation), ce sont des changements morphologiques. Mais d'autres changements qui ne comportent pas nécessairement de changements morphologiques peuvent caractériser l'édifice, ce sont des changements dans l'usage religieux de l'édifice (fonction cimetériale, abbatiale, etc.). La figure 2.7 présente les différents cas de figure d'évolution d'un édifice qui peuvent

correspondre à des propagations de changements morphologique ou d'usage.

En résumé, ces constations font apparaître trois dimensions liées à l'évolution des édifices : 1) l'usage religieux ; 2) l'espace avec l'édifice et ses différents éléments ; 3) le temps représenté par des états sur des intervalles temporels.
Soient $\mathcal{U}$ l'ensemble des usages, $\mathcal{E}$ l'ensemble des espaces et $\mathcal{T}$ l'ensemble des entités temporelles. Un édifice $\mathcal{B}$ est un sous-ensemble ordonné du produit cartésien des trois ensembles $\mathcal{U}$, $\mathcal{E}$, $\mathcal{T}$. Un édifice à un instant t_0 ou durant une durée $[t_1, t_2]$ est un élément de l'ensemble des parties de $\mathcal{U} \times \mathcal{E} \times \mathcal{T}$ ce qui s'écrit :

$$Etat_{t_0} \in \mathcal{P}(\mathcal{U} \times \mathcal{E} \times \mathcal{T})_{t_0} \; ou \; Etat_{[t_1,t_2]} \in \mathcal{P}(\mathcal{U} \times \mathcal{E} \times \mathcal{T})_{[t_1,t_2]}$$

Les évolutions étudiées sont déterminées par les changements dans chacun de ces ensembles, ce qui se traduit par la création d'un nouvel état dans la fiche de dépouillement. Cette approche permet de :
– restituer tous les états possibles d'un édifice ou d'un de ses éléments constitutifs ;
– analyser et observer tous les changements d'état possibles c'est-à-dire de procéder à une différence entre deux états ;
– comprendre les évolutions c'est-à-dire le processus de changement d'état. Le principe d'indépendance entre l'usage, l'espace et le temps permet de les regrouper en produit deux à deux afin d'observer quels sont les facteurs qui exercent une influence sur le changement et d'estimer le rôle ou la prépondérance de l'un par rapport à l'autre. Le tableau 2.1 présente l'interaction des dimensions entre elles.

Dimension	Signification
$\mathcal{E} \times \mathcal{U}$	Un espace donné (cloître, baptistère, cimetière) détermine l'usage
$\mathcal{U} \times \mathcal{T}$	Un usage donné appartient à une seule période chronologique
$\mathcal{E} \times \mathcal{T}$	Un changement de plan de l'édifice ; Une redistribution spatiale telle qu'une réorganisation des bâtiments d'un couvent
$\mathcal{U} \times \mathcal{E} \times \mathcal{T}$	Une étude de l'édifice sur une longue période

Table 2.1 – Interaction des trois dimensions

Comme nous l'avons vu les édifices constituent un terrain d'analyse très complexe, une modélisation spatio-temporelle spécifique aux besoins des domaines architectural et patrimonial culturel est nécessaire.

2.4.3 Modélisation avec UML

À partir de ces concepts saillants, nous isolons deux préoccupations essentielles : 1) restituer les états passés ; 2) comprendre les transformations morphologiques et d'usage subies par les édifices. Cette première approche nous a permis de délimiter les frontières du corpus, de mettre en évidence les aspects qui peuvent varier et de faire apparaître certaines notions clés. Nous avons appliqué les directives de Linster [Lin92] décrites plus haut au projet CARE.

Cadre pour la sémantique

La figure 2.8 structure, à l'aide d'un diagramme de classes conceptuel (couleur orange) respectant les règles d'Evermann, les connaissances du projet CARE en deux grands groupes :

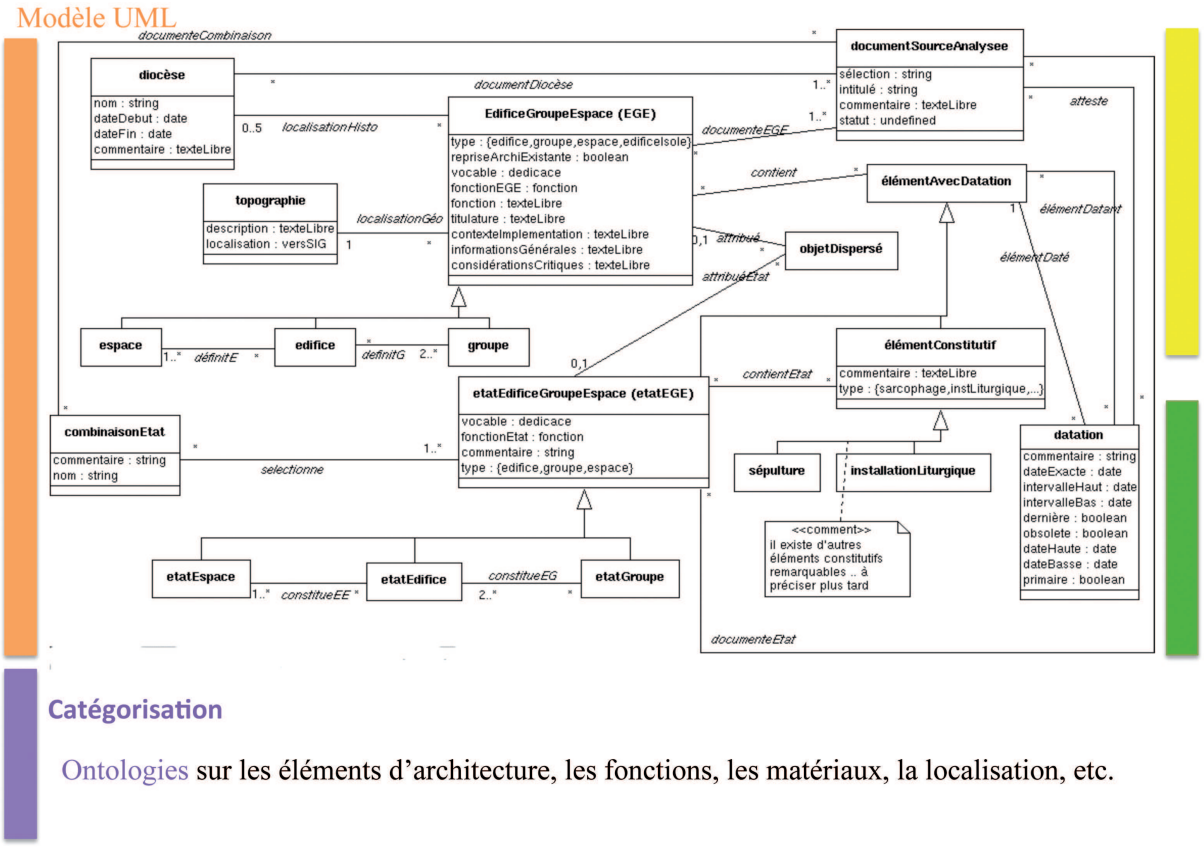


FIGURE 2.8 – Vision globale du modèle conceptuel UML et des ontologies correspondant au projet CARE

– la partie haute (couleur jaune à droite) décrit les éléments non datés. Parmi ces éléments, la classe `EGE` décrit de façon synthétique les composants spatiaux où aucune variation temporelle n'est indiquée. Les composants spatiaux peuvent être des groupes d'édifices composés d'édifices eux-mêmes composés d'espaces (nef, abside, etc.). Cette partie de la description correspond aux classes `EGE`, `groupe`, `edifice` et `espace` ;
– la partie intermédiaire (couleur vert à droite) décrit les états des composants spatiaux avec la classe `etatEGE`. Un état a obligatoirement une datation qui fait référence à d'autres éléments datants. Une datation peut être attestée par des documents ou des techniques de datation (au carbone-14 par exemple). Les composants spatio-temporels, qui représentent les variations temporelles de composants spatiaux sont composés de la même façon que les composants spatiaux : les états de groupe sont composés d'états d'édifices eux-mêmes composés d'états d'espace. Cette partie de la description correspond aux classes `etatEGE`, `etatGroupe`, `etatEspace`, `etatEdifice` et `etatEspace` ;

La partie basse de la figure comprend des ontologies (couleur violet) qui constituent le vocabulaire métier.

Pour une fiche, à partir de la classe `EGE` les informations suivantes sont accessibles :
– une localisation soit d'un point de vue religieux (qui correspond au diocèse avec leurs dates de début et de fin) soit d'un point de vue géographique (qui correspond à l'adresse, la latitude et la longitude) ;
– les états de l'édifice ou du groupe d'édifices. Dans chaque état, le plan du monument avec les

Dans une classe élémentConstitutif sans datation *Les datations correspondent à l'extrait de texte en bleu, les éléments à l'extrait de texte en vert.* 4. INSTALLATIONS LITURGIQUES À la limite du collatéral nord et de la nef centrale, une grosse pierre présentant une mortaise témoigne d'un tracé de clôture. Dans une classe élémentConstitutif avec datation 5. SÉPULTURES 5.1 Emplacement et relation avec l'édifice : Les sépultures sont présentes aussi bien à l'est et à l'ouest marquant une pérennité de l'usage funéraire du lieu. [...] trois groupes d'inhumations ont été identifiés entre Antiquité tardive (seconde moitié du IV^{e} s.) et l'époque moderne (XVII^{e} s.). Plusieurs sarcophages et fragments de sarcophages du IV^{e} siècle ont été retrouvés en relation avec l'édifice. [. . .] 6. OBJETS DISPÉRSES NON RATTACHABLES À L'ARCHITECTURE DE L'ÉGLISE Des fragments de marbre ont été découverts hors contexte composés de : plaques, bases, moulures antiques. [...] Un fragment de pilier a été découvert réemployé. [...]

Table 2.2 – Extrait de la fiche de dépouillement Saint-Pierre-Estrier à Autun

concepts d'éléments architecturaux, de fonction sont connus à partir des éléments de datation relative telles que les techniques de construction, les matériaux de construction, les sépultures, etc.

Afin de valider cette modélisation, nous avons catégorisé plusieurs fiches, c'est-à-dire sur chacune des fiches nous avons « projeté » la structure du modèle conceptuel UML. Par exemple, à partir d'un extrait de la fiche de l'église Saint-Pierre-Estrier à Autun (texte encadré en 2.2), la figure 2.9 montre la partie du diagramme de classes correspondante : l'attribut `commentaire` des différentes classes est catégorisé avec le numéro de la section de la fiche qui lui correspond, l'attribut `type` de la classe `élémentConstitutif` avec par exemple sépulture ou sarcophage, etc. L'intégralité de la fiche et sa catégorisation sont données en section 2 de l'annexe A.

Cadre pour implémenter le modèle UML

Le modèle conceptuel UML proposé est difficilement exécutable car il demande de recourir à du SQL récursif. Nous nous sommes heurtés à la grande variabilité dans la description des édifices, chaque élément d'un édifice doit être représenté même si cet élément n'a été recensé que dans un seul édifice. Ce traitement de la variabilité introduit dans le diagramme de classes des classes avec très peu d'instances, de multiples relations de généralisation/spécialisation (par exemple la classe `chapiteau` sera spécialisée avec les classes `chapiteau cubique`, `chapiteau à godron`, `chapiteau végétal`, `chapiteau historié`, etc.) et des associations réflexives. De plus, le projet impliquant des chercheurs de plusieurs disciplines, différents points de vue doivent pouvoir être modélisés sur une même classe introduisant de nouvelles relations de généralisation/spécialisation [8].
Notre approche demandant une représentation spatio-temporelle d'un édifice, quatre aspects doivent être considérés. Premièrement, il faut décrire les éléments constitutifs de l'édifice par

8. La classification multiple permet de modéliser plusieurs décompositions sémantiques correspondant à des points de vue différents sur une même classe. Pour exprimer la classification multiple en UML, la généralisation avec un discriminant est utilisée.

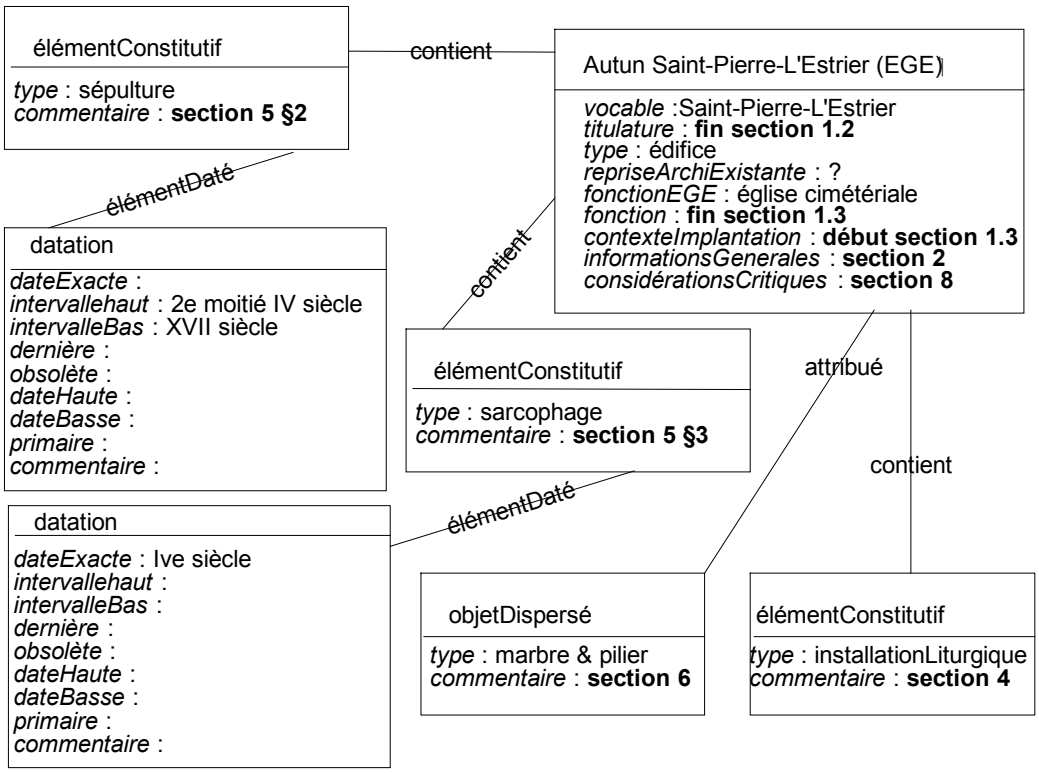


FIGURE 2.9 – Extrait de la catégorisation sur la fiche Saint-Pierre-Estrier à Autun

rapport à leur position dans l'espace. Deuxièmement, il est nécessaire de décrire les caractéristiques physiques d'un édifice par des mesures, des matériaux, etc. Troisièmement, il faut ajouter une dimension temporelle pour prendre en compte les différentes évolutions. La représentation spatio-temporelle doit pouvoir modéliser des transitions à la fois instantanées et graduelles. Enfin, le dernier aspect à prendre en compte est l'aspect documentaire car la reconstitution de tous les états s'appuie aussi sur les sources disponibles. La figure 2.10 schématise cette représentation spatio-temporelle particulière, le stéréotype $<< Variation >>$ indique que le package `Modèle Architectural` contient des éléments qui varient, ce stéréotype appartient à un profil défini pour les lignes de produits qui présentent la même caractéristique de variabilité [ZHJ04]. À partir du diagramme de classes conceptuel de la figure 2.8, nous avons construit le diagramme de classes UML donné en annexe A - figure A.10. Pour élaborer ce diagramme, nous avons :

– explicité les concepts de groupe d'édifices, d'édifice, d'espace et d'élément constitutif ce qui nous a permis de supprimer les associations récursives et les relations d'héritage ;
– utilisé *executable* UML [MB02] permettant ainsi une traduction automatique vers le modèle relationnel.

Ce modèle exécutable comporte plus d'une trentaine de classes, plus de cinquante associations pour une modélisation partielle qui s'est révélée difficile à stabiliser après trois mois de travail avec les archéologues. Chaque édifice ayant des spécificités qui lui sont propres, nous sommes amenés à introduire dans le diagramme de classes des éléments qui seront très peu utilisés. De plus ce projet s'accompagnant de nouvelles fouilles (sur le terrain ou bibliographique) le diagramme de classes est amené à évoluer. Nous avons alors décidé de ne garder dans notre diagramme de classes UML que les concepts et relations correspondants aux concepts saillants du domaine (c'est-à-dire aux données référentielles) et compléter cette modélisation avec une

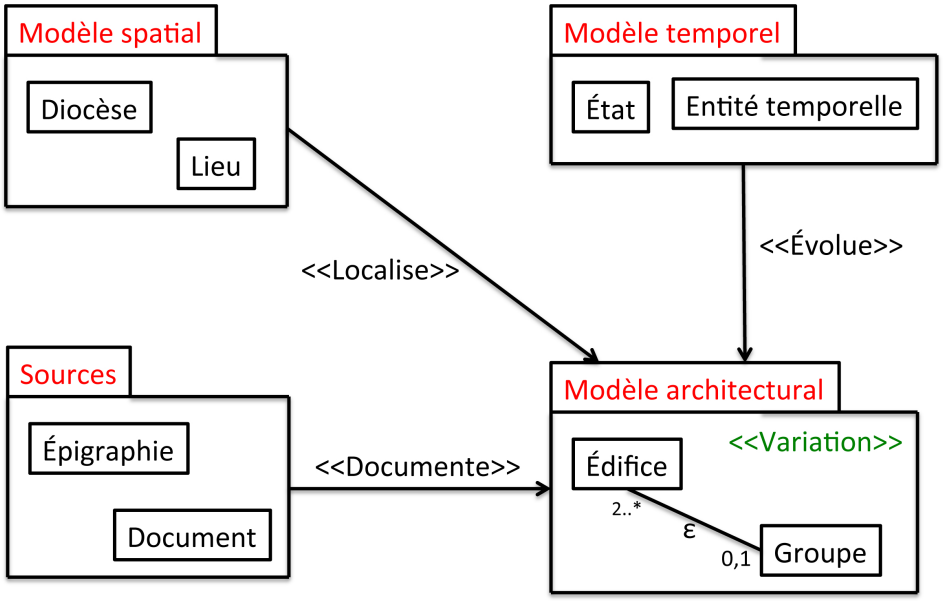


Figure 2.10 – Représentation de la spatio-temporalité et de la variabilité des éléments du projet CARE

modélisation basée sur la Logique de Description donc sur les ontologies.

2.4.4 Modélisation avec la Logique de Description

Comme Richards [Ric09], nous estimons nécessaire l'utilisation de standards en archéologie. Notre objectif est donc de construire une ontologie exhaustive en adéquation avec les faits qui ont une importance déterminante. Nous avons des objets complexes dont nous voulons suivre l'évolution. Ce qui implique au niveau modèle de pouvoir détailler chaque composant de l'édifice et chaque élément factuel ou non ayant provoqué une évolution ou permettant une datation.

Cadre pour la sémantique

Dans le projet CARE, l'ontologie spécialisée dans les édifices religieux en France et leur évolution sert de cadre. Les experts du domaine (archéologues, historiens, etc.) ont contribué directement à la construction de l'ontologie en fournissant des spécifications fonctionnelles (c'est-à-dire métier). Nous avons traduit ces spécifications dans un tableur pour dégrossir le terrain (voir annexe B). Cela fait nous avons pu travailler à une représentation formelle des connaissances des éléments d'architecture en :

1. classifiant les éléments architecturaux d'un édifice : éléments maçonnés (châteaux, fûts, etc.), charpentes, sols, etc. ;
2. définissant les relations entre les éléments. Goulette [Gou99] a énuméré les différentes relations entre éléments : relation partie-tout, relations spatiales, relations de composition. Nous avons aussi dans notre analyse pris en compte les relations qu'un élément d'architecture peut avoir avec une technique de construction, un élément stylistique, etc. ;
3. établissant une correspondance entre les éléments d'architecture et le domaine religieux. Nous avons pour cela utilisé la signification donnée par un élément comme une installation liturgique ou par un espace. Par exemple la nef est l'endroit où sont rassemblés les fidèles, le chœur est le lieu où se trouve l'autel et où se déroule la liturgie ;

4. tenant compte de la dimension temporelle.

Cette méthodologie nous a permis de dégager un ensemble de concepts connectés à des termes architecturaux organisés par des relations méronomiques (décomposition morphologique) et représentationnelles (dimensions, matériaux, relations spatiales).

L'ontologie obtenue s'appuie à la fois sur :
– une ontologie de domaine dans la gestion du patrimoine. Au prix d'un délai de prise en main qui dépend de la taille des ontologies de domaine, leur usage évite de reconstruire totalement la représentation de la connaissance d'un domaine. Cela permet aussi d'élargir le socle consensuel de l'ontologie obtenue, aux communautés qui se reconnaissent dans les ontologies de domaine utilisées à sa construction. Cela facilite par voie de conséquence l'interopérabilité qui est une caractéristique du projet CARE qui sera exploitée lors de l'analyse du corpus par des outils externes comme un SIG. Nous avons retenu CIDOC-CRM [CC02] présenté en annexe A ;
– des ontologies développées localement par les membres du projet (iconographe, historien de l'art) pour répondre à un besoin non satisfait.

Ainsi, l'ontologie prend en compte l'ensemble des données archéologiques : bâtiments, éléments architecturaux, données intangibles telles que des mesures, des relations spatiales, des associations entre des éléments et des données d'interprétation. CIDOC-CRM apporte la modélisation temporelle et a été spécialisée par l'apport des branches suivantes : concepts religieux, concepts architecturaux et relations spatiales.

Nous avons ensuite utilisé la logique de description [BB03] pour traduire les concepts de CARE dans une ontologie formelle dont une partie est présentée dans l'encadré 2.3 :

Cette modélisation conceptuelle peut être rapprochée de la modélisation conceptuelle réalisée avec le diagramme de classes UML : nous retrouvons des éléments non datés (ici la structure) et des éléments datés avec l'introduction de la notion d'état. Dans les deux cas, il s'agit d'une structure formelle de la réalité. Il apparait donc que le raisonnement en monde ouvert est particulièrement bien adapté à la recherche notamment en archéologie où n'importe quelle proposition est acceptée tant que elle n'a pas été contredite.
La figure 2.11 illustre le haut du tableau 2.3 c'est-à-dire les éléments généraux d'un bâtiment. L'ontologie est présentée en détail dans la section 5 de l'annexe A.

```
ObjetFabriqueE22 sameas Batiment

Batiment ⊑ (and (the Nom String)
                (the APourRegion Region)
                (the APourDepartement Departement)
                (the APour Commune Commune)
                (the Adresse String)
                (the Altitude real)
                (the Latitude real)
                (the Longitude Real)
                (all APourFonction Fonction)
                (all APourSource DocumentE31)
                (the ApourDiocese Diocese (max 2))
                (the ApourTitulature))

Groupe ⊑ (and Batiment
          (At least 2 APourComposant Edifice) )
Edifice ⊑ (and Batiment
          (At least 1 Etat))

Structure ⊑ (and QuelqueChoseDeMaterielEtFabriqueE24
              (all APourRel RelationSpatiale)
              (the APourForme FormeGeometrique max 1)
              (the APourDimension DimensionE54 max 1))

EspaceLogique ⊑ Structure
EspacePhysique ⊑ Structure

Etat ⊑ (and
(the EstAssocie TrancheChronologiqueE52 )
        (the APourCause EvenementE5)
        (all PorteSur QuelqueChoseDeMaterielEtFrabriqueE24)
        (the Connu TechniqueDatation)
        (the Concerne Batiment))
```

Table 2.3 – Extrait de l'ontologie formelle CARE (modèle conceptuel) - Les concepts se terminant par EXX sont issus de CIDOC-CRM

Cadre pour implémenter l'ontologie

Enfin, il est nécessaire de traduire cette ontologie dans un langage formel et opérationnel de représentation de connaissances. Dans notre cas il s'agit d'OWL en utilisant l'éditeur Protégé [9]. Nous voyons l'ontologie dans l'éditeur Protégé comme un modèle exécutable. Une capture d'écran en figure 2.12 montre un extrait des classes, des individus et des relations créés, l'intégralité de l'ontologie est donnée en section 5 de l'annexe A.

9. Protégé, site Web : http://protege.stanford.edu/

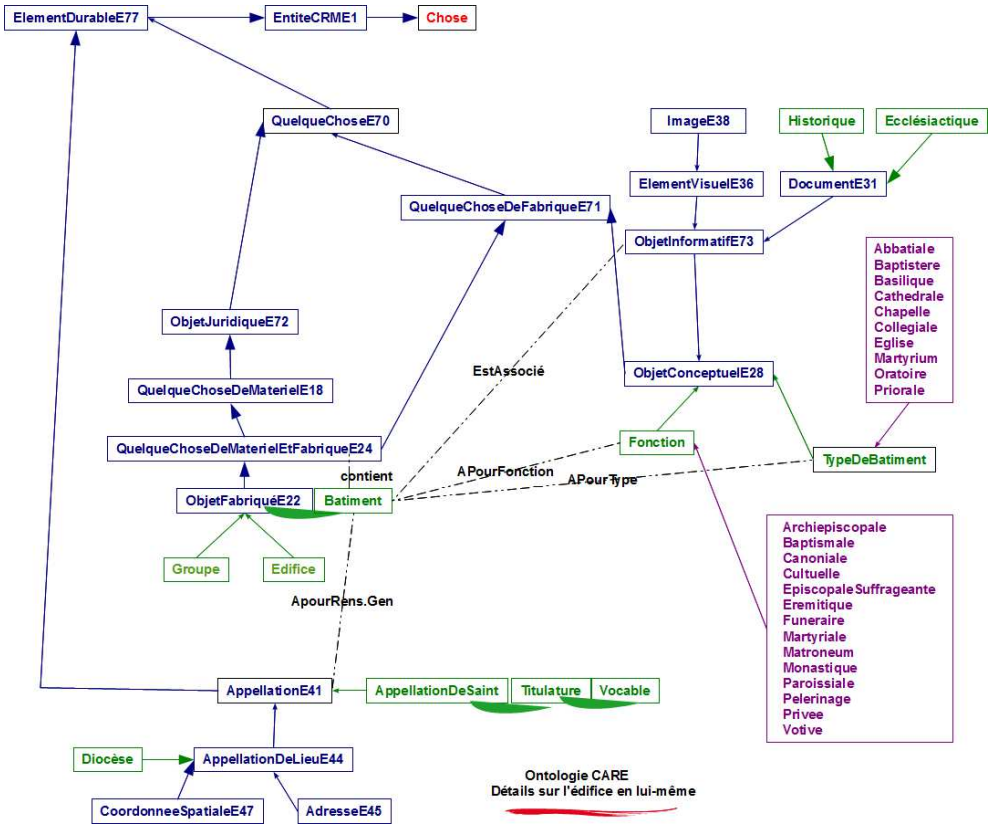


FIGURE 2.11 – Partie d'ontologie relative aux éléments généraux d'un bâtiment

2.4.5 Modèles exécutables dans CARE : structuration de la connaissance

Tout comme Lock [Loc09] qui pense que l'interprétation des données ne doit pas être déterminée par la technologie utilisée mais bien par la discipline étudiée, Bachimont [Bac00] montre qu'instrumenter un travail n'est jamais une opération neutre, n'importe quel outil détermine par sa structure des usages possibles (ce qui n'empêche pas des usages déviants). La question de l'adéquation de l'outil au travail est donc primordiale. Le contexte du projet CARE demandant une interface Web avec une composante fortement collaborative nous a amené à choisir un wiki comme modèle d'interaction. Ce dernier présente comme avantage de respecter la façon de travailler des archéologues qui est centrée sur des descriptions textuelles.

Chaque document se traduit sous la forme d'un article dans le wiki. L'organisation des documents du projet CARE est modélisée sous la forme d'un formulaire complexe avec onglets et zones variables reflétant l'organisation du document papier. Cette organisation s'établit autour d'abstractions représentant les différentes parties du document qui ont été mises en évidence par la modélisation conceptuelle UML. Ces abstractions correspondent à la notion de données référentielles.
La grande majorité des wikis permet d'organiser les articles en leur assignant une ou plusieurs catégories. L'ensemble des catégories est défini et maintenu manuellement. Les limites de cette approche deviennent visibles lorsque les utilisateurs du wiki veulent rechercher des informations très précises sur un sujet particulier ou des informations réparties dans plusieurs articles. Les wikis proposent seulement un moteur de recherche textuel. Pour extraire des informations quantitatives, effectuer des comparaisons, des vérifications ou des analyses spatiales, les utilisateurs du projet CARE ont besoin des capacités d'un véritable langage de requête comparable

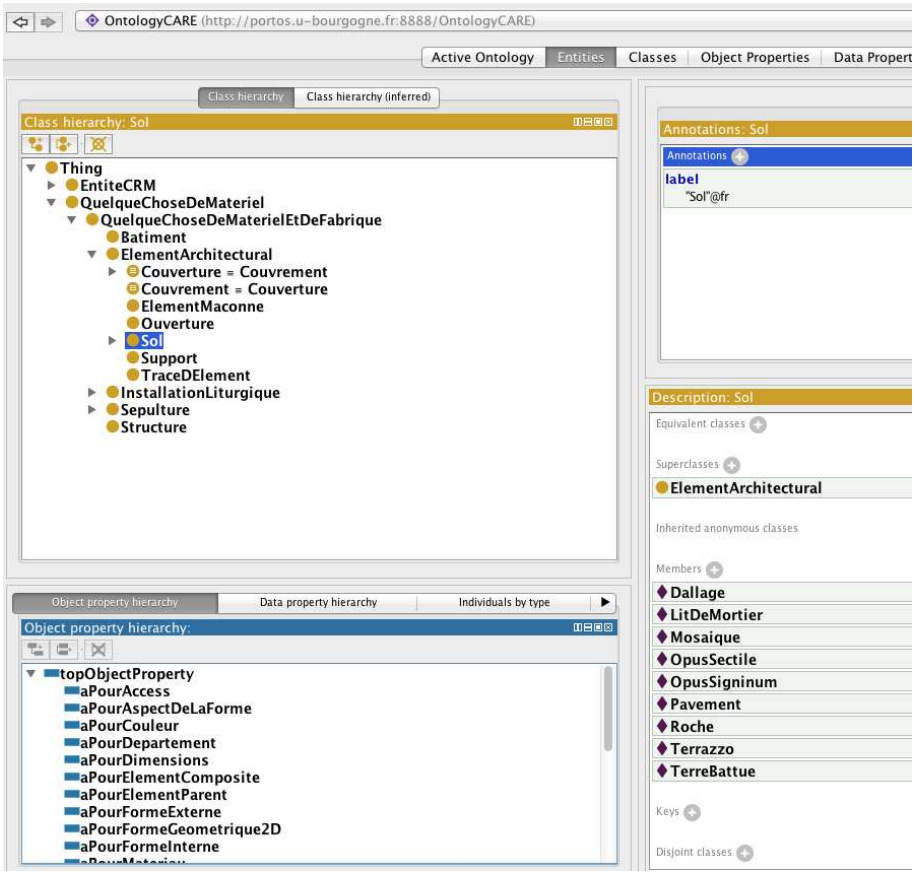


FIGURE 2.12 – Extrait de l'ontologie CARE dans l'éditeur Protégé (modèle exécutable)

à SQL. Par conséquent, le sens du document c'est-à-dire les informations qu'un utilisateur est susceptible de demander lors d'une recherche doit être rendu explicite par une structure sémantique. Pour cela, nous utilisons une ontologie et un mécanisme d'annotations pour représenter les connaissances associées aux articles, l'ensemble constitue une base de données annotées. La figure 2.13 synthétise cette démarche.

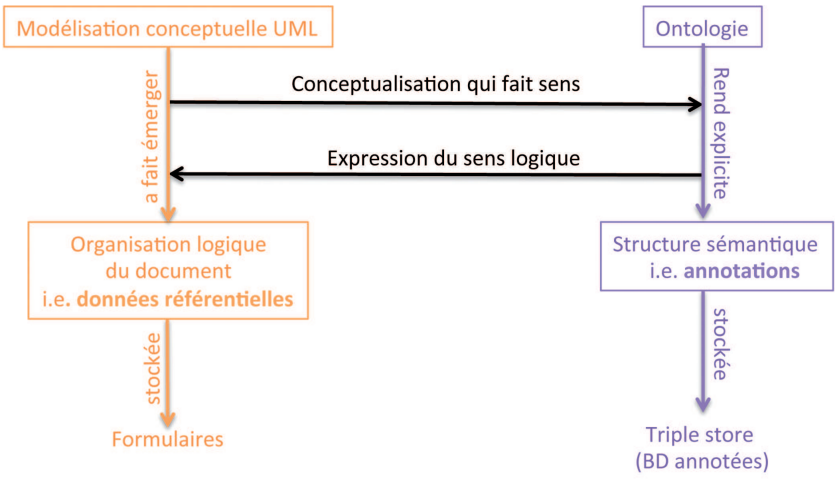


FIGURE 2.13 – Structuration de la connaissance dans CARE

La section suivante présente le projet *eClims* qui a donné lieu à la thèse de Pierre Naubourg (soutenue en 2011) que j'ai co-encadrée.

2.5 Application au projet *eClims* : gestion des données cliniques d'une plate-forme de protéomique clinique

Un partenariat, développé depuis 2005, avec des chercheurs de la plate-forme de protéomique CLIPP [10] nous a conduit à nous intéresser à la protéomique clinique. Ce sous-domaine de la biologie a pour objectif de détecter, identifier, caractériser des protéines impliquées dans l'apparition de pathologies humaines ou la résistance à des traitements. Les caractéristiques étudiées sont la quantité de protéines présentes dans l'organisme, leurs structures, leurs localisations, leurs interactions, leurs fonctions. L'un des principaux enjeux de la protéomique clinique est l'identification de bio-marqueurs permettant la détection précoce, le diagnostic ou le suivi de pathologies ainsi que le contrôle de la réponse des patients aux traitements reçus.

2.5.1 Caractéristiques des données biologiques utilisées en protéomique clinique

Les techniques informatiques mises en œuvre dans le domaine biomédical et en particulier dans le domaine de la protéomique clinique doivent faire face aux enjeux suivants :
- une croissance exponentielle du volume des données à gérer : un volume de l'ordre de 2 petabytes de données était prévu pour l'année 2010. De plus il existe à ce jour plus de 1 330 bases de données biomédicales d'audience internationale [GFS12] et plus de 60 ontologies dans le seul consortium OBO [Smi05, SAR+07] ;
- une représentation des données complexes [CC03] : en effet, les données sont disparates car elles sont issues de nombreux processus et représentent souvent différents aspects du même phénomène en utilisant différents modèles, elles sont hétérogènes car elles proviennent de sources multiples ;
- les données peuvent présenter des problèmes de qualité car elles sont souvent incertaines, incomplètes, incohérentes et inconsistantes [Wil98].

En outre, la connaissance évolue rapidement sous l'influence des technologies dites à haut débit et des techniques d'analyse de données. L'évolution de la connaissance se traduit par :
- des évolutions coûteuses des schémas de bases de données ;
- la publication d'un grand nombre d'articles et de bases de données de référence qui doivent être associés aux données acquises ou produites.

Ces enjeux ne sont pas forcément en adéquation avec les modèles utilisés dans les Bases de Données tant au niveau du volume à gérer qu'au niveau du modèle de données qui doit supporter à la fois la disparité des données et leur évolution.
La protéomique clinique demande de plus l'association entre des données expérimentales provenant de différentes sources et des données clinico-biologiques obtenues de différents partenaires.

2.5.2 Modèle de traitement des études de protéomique clinique

L'organisation la plus répandue au sein des systèmes informatiques gérant des plate-formes de protéomique clinique se présente sous la forme d'une hiérarchie de trois éléments : un programme est défini par un thème de recherche relatif à une pathologie, il est divisé en plusieurs projets qui structurent les différentes hypothèses faites par les experts en fonction des ressources disponibles. Un projet est à son tour décomposé en plusieurs études qui réduisent le champ d'interrogation

10. CLIPP : CLinical Innovation in Proteomics Platform, http://www.clipproteomic.fr

des hypothèses en essayant de répondre à une question précise. Chacune des études donne lieu à la création d'une conclusion permettant ou non de répondre à l'hypothèse du projet dont elle dépend.

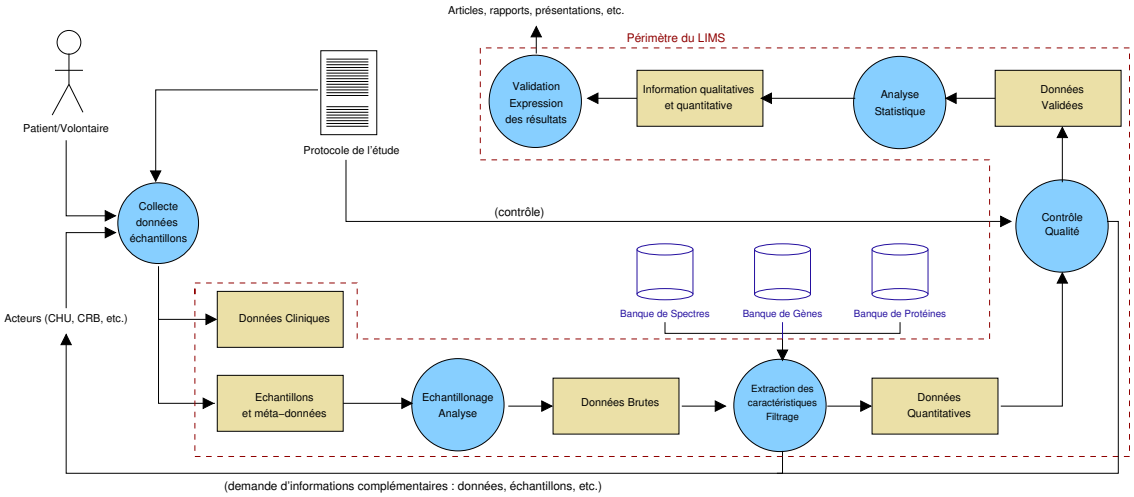


Figure 2.14 – *Workflow* simplifié d'une étude de protéomique clinique

Nous représentons en figure 2.14 un *workflow* simplifié d'une étude de protéomique clinique qui porte sur un ensemble de patients pour lesquels des données cliniques sont collectées auprès de partenaires. Ces données recouvrent des données sur les patients (sexe, date de naissance, variables biologiques, variables cliniques, etc.) et des données sur les échantillons biologiques comme le type de prélèvement (fluide, tissu, etc.), les paramètres d'échantillonnage, les protocoles de prélèvement et les conditions de stockage. Cette collecte des données se fait dans le cadre de la loi « Informatique et libertés » du 6 janvier 1978. Les partenaires d'une plate-forme protéomique peuvent être des Centres Hospitaliers Universitaires (CHU), des centres de lutte contre le cancer, des centres de prélèvements, des Centres de Ressources Biologiques (CRB), des laboratoires pharmaceutiques. La préparation pré-analytique des échantillons est basée sur des protocoles et recommandations normés. Les expériences protéomiques sont fondées sur la spectrométrie de masse qui permet la détection et la quantification relative à grande échelle des peptides contenus dans les échantillons prélevés sur les patients. À partir de la sélection de marqueurs potentiels, il est ensuite possible d'identifier les protéines pertinentes. Cette technologie produit de grandes quantités de données brutes qui sont des spectres de masse exprimés sous forme de tableaux de nombres. Les données brutes subissent une succession de pré-traitements pour retirer les bruits techniques et extraire ainsi l'information biologique convenable (élimination du bruit aléatoire de mesure et du bruit de fond, normalisation, détection dans le spectre des pics qui correspondent à des peptides). Un contrôle de qualité des données quantitatives ainsi obtenues est alors effectué pour produire des données validées. Les données validées obtenues font l'objet d'analyses statistiques pour identifier parmi les peptides détectés les bio-marqueurs potentiels : analyses descriptives non supervisées (classification ascendante hiérarchique par exemple), analyses supervisées à visée diagnostique ou pronostique. Les résultats obtenus sont alors expertisés en reprenant les paramètres de l'étude, puis ils sont publiés sous la forme d'articles, rapports, présentations, etc. Après validation des résultats obtenus, des informations complémentaires sur les échantillons analysés peuvent être demandées aux partenaires (clinicien par exemple). Mischak et al. dans [MAB+07] présentent plus en détail les étapes d'une étude en protéomique clinique et l'information nécessaire à chacune des étapes.

2.5.3 Modélisation avec UML

La plate-forme CLIPP a souhaité informatiser son Système d'Information de laboratoire par l'utilisation d'un LIMS (*Laboratory Information Management System*) [CHLBS98]. Pour cela, CLIPP a pris contact avec la plate-forme EDyP [11]. EDyp a développé en interne un LIMS dédié à la protéomique baptisé ePims (*experiment Proteomic Information Management System*) [DBB09]. Les recherches menées par EDyP n'étant pas liées à des méthodologies de protéomique clinique, une confrontation des besoins de CLIPP aux services proposés par ePims a été réalisée par la société ASA [12], responsable de l'exploitation d'ePims et nous-même. Nous avons identifié trois fonctionnalités majeures absentes au sein de ePims et qu'il fallait développer :

1. la gestion des données cliniques en amont des expériences. Ces données concernent tous les évènements conditionnant le contexte général du prélèvement (quantité, état, provenance, etc.), les conditions d'arrivée sur la plate-forme (conditions de stockage, durée de transport, etc.) ou encore les antécédents des patients donneurs (leurs pathologies, leurs traitements, etc.) ;
2. le contrôle de qualité des données issues des expériences ou des échanges avec les partenaires ;
3. la gestion des études statistiques.

Nous avons défini trois composants pour gérer les données relatives aux recherches en protéomique clinique, en amont des expériences, pendant les manipulations et en aval pour l'exploitation des données expérimentales obtenues :

1. un composant clinique permettant la gestion de l'ensemble des données disponibles en amont des expériences réalisées ;
2. un composant laboratoire fournissant les outils nécessaires à la configuration des expériences et au suivi des données brutes issues des spectromètres (ePims pour CLIPP) ;
3. un composant statistique permettant de gérer les études statistiques, réalisées en aval des expériences, et nécessaires à la réalisation des conclusions des études protéomiques.

Cadre pour la sémantique

Nous avons pris en charge le composant clinique, baptisé *eClims* (*experiments Clinical Information Management System*), qui sert d'interface entre les données provenant des différents partenaires (CRB, CHU, etc.) et le Système d'Information de CLIPP. Ce composant apporte à la plate-forme CLIPP un outil de suivi et de contrôle de la qualité des données cliniques associées aux échantillons reçus. Il est détaillé dans le chapitre 5.

La figure 2.15 présente les grandes lignes de l'organisation des données des deux premiers composants (couleur orange). Ce modèle montre que le composant clinique traite essentiellement des groupes de prélèvements issus d'un ensemble de patients sélectionnés d'après les paramètres de l'étude. La classe `Information` permet de gérer l'ensemble des données associées à l'étude, au patient et au prélèvement. Pour un patient, outre les informations habituelles (sexe, date de naissance, état) il faut connaître ses pathologies avec leur date de diagnostic, les traitements mais aussi des informations spécifiques à l'étude (par exemple son origine ethnique peut être importante dans certaine maladie). Pour un prélèvement, les informations de traçabilité sont indispensables : lieu de prélèvement, conditions de transport et de stockage (c'est-à-dire les transports entre les différentes entités, les températures de congélation, etc.). Ces prélèvements

11. EDyp : laboratoire d'Étude de la Dynamique des protéines du Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) de Grenoble, site Web : http://www-dsv.cea.fr/ledyp
12. ASA : Advanced Solutions Accelerator, site Web : http://www.advancedsolutionsaccelerator.com/

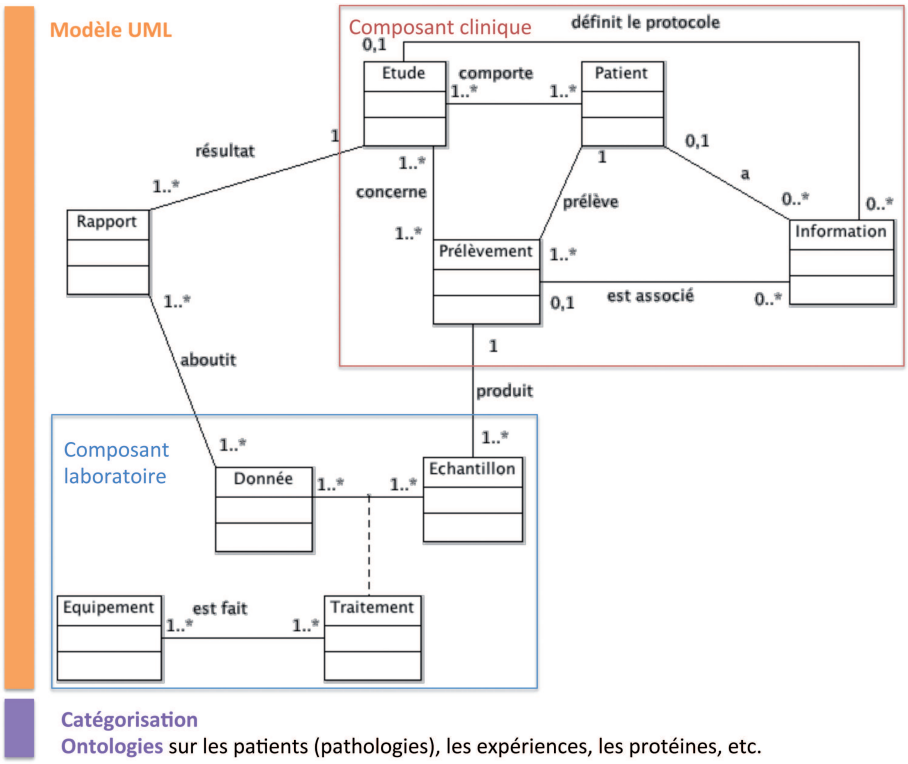


Figure 2.15 – Vision globale du modèle conceptuel UML et des ontologies d'une étude de protéomique clinique

sont échantillonnés pour donner les échantillons qui seront analysés durant les expérimentations. Ce sont les échantillons qui font l'interface entre le composant clinique et le composant laboratoire. Après expérimentations, les données brutes subissent divers traitements pour aboutir à des rapports (intermédiaires, statistiques, finaux). La partie basse de cette figure comprend des ontologies et ressources consensuelles (couleur violet) qui constituent le vocabulaire métier.

Cadre pour implémenter le modèle UML

Nous nous sommes heurtés à la grande variabilité des études protéomiques. En effet, les études en protéomique clinique répondent à des questions très variées ; par exemple, une étude sur la détection de marqueurs prédictifs de la stéatose [13], une étude sur la recherche de marqueurs salivaires reliés aux préférences gustatives chez les nourrissons. De même, il serait souvent très intéressant dans le cadre de certaines pathologies de croiser les données protéomiques de différentes études pour définir des dénominateurs communs à certaines "sous catégories de patients" ; par exemple, une étude sur la détection de marqueurs prédictifs de l'apparition d'hépatocarcinome [14] cellulaire chez les patients cirrhotiques *versus* une étude sur la détection de marqueurs prédictifs de résistance thérapeutique chez les patients présentant un hépatocarcinome cellulaire. Cela implique que la classe `Information` du modèle conceptuel doit être capable de modéliser des données variées multi-modèles issues de multiples sources.
Nous avons alors décidé de ne garder dans notre diagramme de classes UML que les concepts et relations correspondants aux données référentielles et compléter cette modélisation avec des ontologies.

13. La stéatose est l'accumulation d'une graisse, appelée triglycéride, dans la cellule hépatique elle-même.
14. L'hépatocarcinome est un cancer du foie.

2.5.4 Modélisation avec la Logique de Description

Un point important dans ces études est la contextualisation des données. Ainsi deux résultats identiques peuvent selon le patient ou l'échantillon conduire à des conclusions différentes. Chaque étape du *workflow* doit donc prendre en compte la connaissance consensuelle s'y rapportant pour réaliser cette contextualisation. Dans de nombreux domaines de la biologie, des initiatives ont été entreprises dans ce sens. Elles sont essentiellement réalisées par une communauté d'experts d'un domaine.

Cadre pour la sémantique

Il existe de nombreuses ressources consensuelles en protéomique clinique. Nous pouvons citer par exemple :
- le consortium Gene Ontology [15] regroupe un ensemble de bases de données et de communautés de recherche dont l'objectif est de structurer la description des gènes et des produits géniques dans le cadre d'une ontologie commune à toutes les espèces ;
- la Classification Internationale des Maladies (CIM) [16] définit une nomenclature précise des pathologies grâce à un codage des maladies, des traumatismes et d'une manière générale de l'ensemble des motifs de recours aux services de santé ;
- la nomenclature TNM [17] est un système international permettant de définir les stades de développement des tumeurs ;
- le thésaurus MeSH [18] est un outil réalisé par la National Library of Medicine. Il est utilisé pour l'indexation et la recherche d'informations médicales ;
- le dictionnaire ADICAP [19] est une codification des lésions élaborée par l'Association pour le Développement de l'Informatique en Cytologie et en Anatomie Pathologique ;
- les conseils aux tumorothèques [20] définissent des cadres législatifs relatifs à la conservation des échantillons tumoraux et des données personnelles sur les patients ;
- les modèles de données des expériences de FuGE [21] servent de cadre à la définition de standards pour des expériences en biologie et PSI [22][THa06] adapte les propositions de FuGE au domaine protéomique. PSI est organisé en six groupes de travail s'attachant à représenter différents points de vue de la protéomique. MIAPE [23] propose un cadre pour les méta-données en protéomique [TPL+07] ;
- des bases de données sur les protéines. La base de données PDB [BWF+00] répond aux besoins des chercheurs voulant identifier et visualiser une représentation en trois dimensions de

15. site Web : http://geneontology.org/
16. En anglais, International Classification of Diseases (ICD) est proposée par l'Organisation Mondiale de la Santé (OMS), site Web : http://www.who.int/classifications/icd
17. TNM : *Tumor, Node, Metastasis*, site Web : http://www.cancerstaging.org/mission/whatis.html#. Dans le système TNM, T1 et T2 indiquent une tumeur de stade précoce, T3 et T4 une tumeur de stade plus avancé. Le N du système révèle que la tumeur a atteint les glanglions lymphatiques et M que le cancer s'est propagé à d'autres organes (Métastases).
18. MeSH : *Medical Subject Headings*, site Web : http://www.ncbi.nlm.nih.gov/mesh. Les descripteurs MeSH sont organisés en 16 catégories : la catégorie A pour les termes anatomiques, la catégorie B pour les organismes, la catégorie C pour les maladies, etc. Chaque catégorie est subdivisée en sous-catégories comportant des descripteurs structurés hiérarchiquement, des plus généraux aux plus spécifiques.
19. site Web : http://www.adicap.asso.fr/spip.php?rubrique1
20. Une brochure détaillant les conseils aux tumorothèques est accessible à http://cpp.med.univ-tours.fr/tiki/tiki-download_file.php?fileId=9
21. FuGE : *Functional Genomics Experiment*, site Web : http://fuge.sourceforge.net/ [JMA+07]
22. PSI : *Proteomics Standards Initiative*, réalisé par le consortium HUPO (*Human Proteome Organisation*), sites Web : http://psidev.info et http://hupo.org
23. MIAPE : *Minimum Information About a Proteomics Experiment*, site Web : www.psidev.info/miape/

la protéine. La base de données PQS [24] stocke des données issues de différentes sources et les réorganise entres elles selon des caractéristiques structurelles. La base de données SwissProt [25] contient des données moins à jour que dans PDB mais des annotations ont été ajoutées (par exemple, la description des fonctions des protéines, ses modifications possibles, etc). Une partie de cette base de données est spécialisée dans l'espèce humaine avec la base HPI (*Human Proteome Initiative*). La base de données eProtS [26] est une base de données spécialisée dans les interactions entre les protéines. La base de données CATH [PBB+03] propose une classification des protéines en fonction de leur structure. Ces différentes banques / bases de données sur les protéines se complètent et répondent à différents besoins des chercheurs sur les protéines.

Cadre pour implémenter l'ontologie

Les branches de l'ontologie d'application que nous avons mise en place correspondent aux différentes ressources consensuelles propres à la partie clinique. Nous avons choisi la Classification Internationale des Maladies (CIM), la nomenclature TNM, la branche anatomie de la classification MeSH et des recommandations de l'Institut National du Cancer (INCa) aux tumorothèques [27]. Cette recommandation couvre les concepts communs des données cliniques (les concepts patient, échantillon, etc.) utilisés dans notre domaine. Nous avons mis en place des relations entre les différentes branches de l'ontologie. Ces relations, dont un exemple est représenté en pointillé sur la figure 2.16, rendent compte de la connaissance métier, par exemple, en spécifiant quels sont les organes qui sont touchés par une pathologie. Pour cela nous définissons une relation générique *organeTouché* qui relie le concept *Anatomie* de la branche MeSH et le concept *Pathologie* de la branche CIM, permettant ainsi de spécifier les organes touchés par une pathologie. Les experts doivent ensuite «spécialiser» la connaissance en précisant les organes touchés par une pathologie donnée : le *Foie* est un organe susceptible d'être touché par la pathologie *C78.7* qui correspond à une tumeur maligne secondaire du foie.

2.5.5 Modèles exécutables dans *eClims* : espaces technologiques utilisés

Une des principales fonctionnalités du composant *eClims* est d'importer les données cliniques, fournies sous forme de fichier CSV ou issues directement de tableur. Notre approche consiste à utiliser les atouts propres aux ontologies et aux modèles afin de garantir la qualité des données lors du processus d'importation. Pour cela :

- nous utilisons une ontologie d'application comme médiatrice entre les modèles des partenaires et *eClims via* des mappings (pour plus de détails voir Naubourg et al.[NSLY11] et le chapitre 5) ;
- nous avons défini grâce à un diagramme de classes UML exécutable le modèle de données utilisé au sein de *eClims* afin de garantir la structure attendue des données importées.

Afin d'identifier les données référentielles au sein de notre composant, nous avons comparé les schémas des différents systèmes (du LIMS et de plusieurs partenaires) afin de trouver les données référentielles. Cette recherche a mis en exergue les concepts de `Patient`, `Prélèvement`, `Échantillon` et `Diagnostic`. Ces concepts sont effectivement partagés par tous les partenaires, stables car utilisés par toutes les études et fréquemment consultés. Ces quatre concepts forment la base de nos données référentielles auxquelles nous avons rajouté leurs attributs (`NumPatient`, `Date_Naissance`, `Volume`, etc.) et des ressources consensuelles portant sur le patient, sur la

24. PQS : *Protein Quaternary Structure*, site Web : http://www.ebi.ac.uk/pdbe/pqs/
25. site Web : http://www.expasy.ch/sprot/
26. eProtS : *Encyclopedia of Protein Structures*, site Web : http://eprots.pdbj.org/
27. Les tumorothèques sont des banques de tissus tumoraux cryo-préservés.

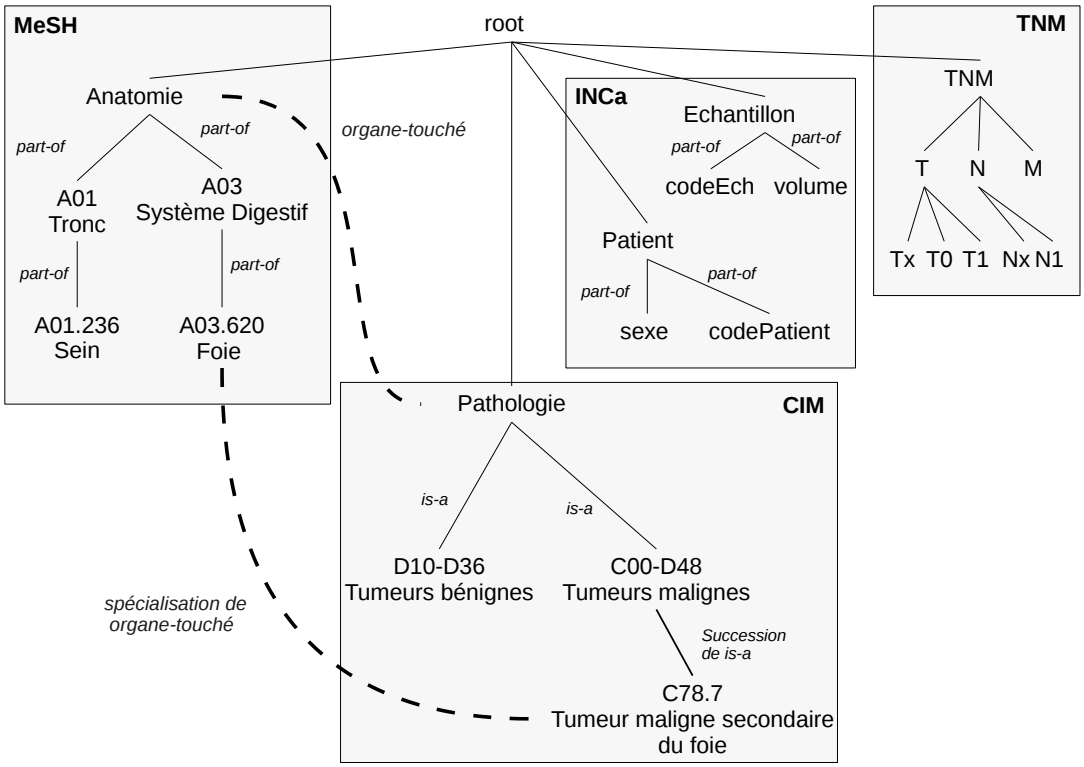


FIGURE 2.16 – Ontologie du composant *eClims* (extrait)

maladie et sur l'ensemble de techniques et protocoles à respecter afin de préserver la qualité des échantillons biologiques. Les éléments soulignés en violet dans le diagramme de la figure 2.17 sont issus du couplage entre l'ontologie et le diagramme de classes. Par exemple, ce diagramme reprend la structure `Chapitre - Section - Elements` de la CIM sous la forme de classes et de relations d'héritage. Ensuite, les classes correspondant aux ressources consensuelles sont instanciées [28] avec les individus de l'ontologie d'application construite (voir figure 2.18).

Ce diagramme de classes (modèle exécutable) a été transformé de manière semi-automatique en schéma de base de données relationnelles et en une collection de classes Java qui effectuent le mapping objet-relationnel. La communication bi-directionnelle entre la base de données et les objets Java peut être réalisée grâce au framework Hibernate [29]. Dans notre approche, les ontologies sont utilisées de deux manières : 1) comme médiatrices entre les systèmes partenaires et *eClims* et 2) comme support de la connaissance. Nous aboutissons à une architecture de modèles construites dans différents espaces technologiques [30] (voir figure 2.19). Chaque espace technologique propose ses propres mécanismes afin de répondre aux différentes attentes des concepteurs d'application. Leurs capacités d'expression peuvent soit se compléter soit au contraire complètement se contredire. Un nouvel espace technologique apparaît le plus souvent lorsqu'une question au sein des autres espaces est restée sans réponse ou que la solution est trop complexe à mettre en œuvre.
Nous avons associé au diagramme de classes des contraintes formulées en OCL. Par exemple, la contrainte OCL énonçant que *le stade TNM associé au diagnostic d'une pathologie doit*

28. La catégorisation est assurée par le mécanisme d'instanciation des classes en Orienté Objet.
29. site Web : http ://www.hibernate.org
30. Les espaces technologiques correspondent à un « *working context with a set of associated concepts, body of knowledge, tools, required skills, and possibilities.* »[KBA02]

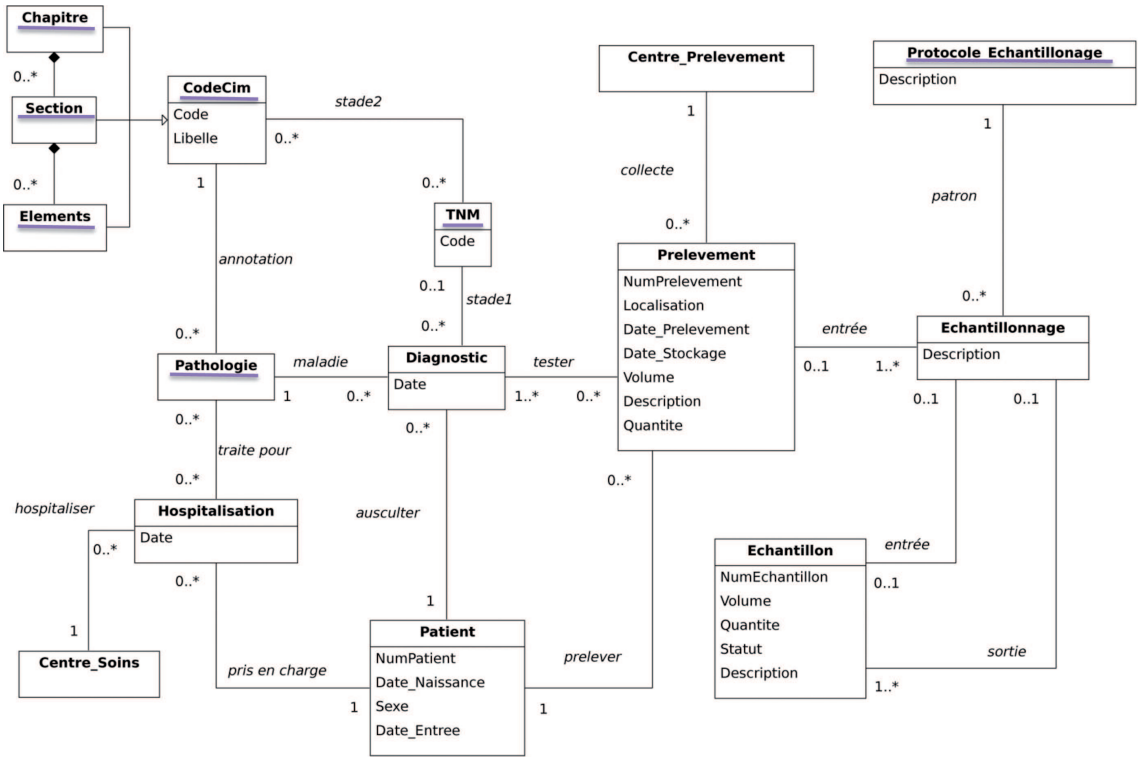


FIGURE 2.17 – Diagramme de classes exécutable du composant *eClims* (données référentielles)

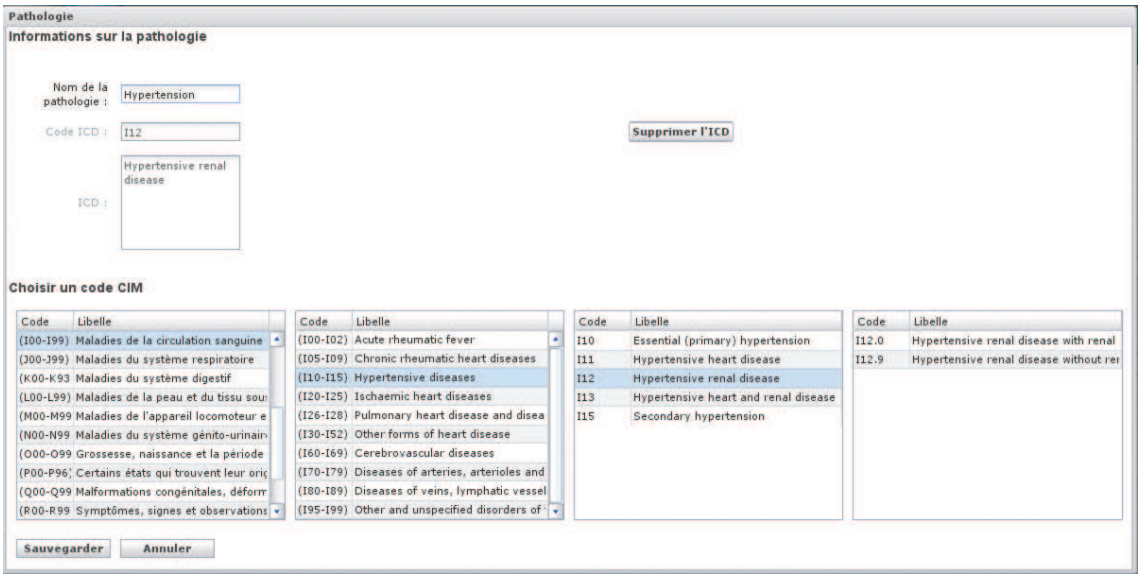


FIGURE 2.18 – Couplage diagramme de classes / ontologie (exemple de la CIM)

correspondre aux stades TNM possibles de la pathologie s'écrit :

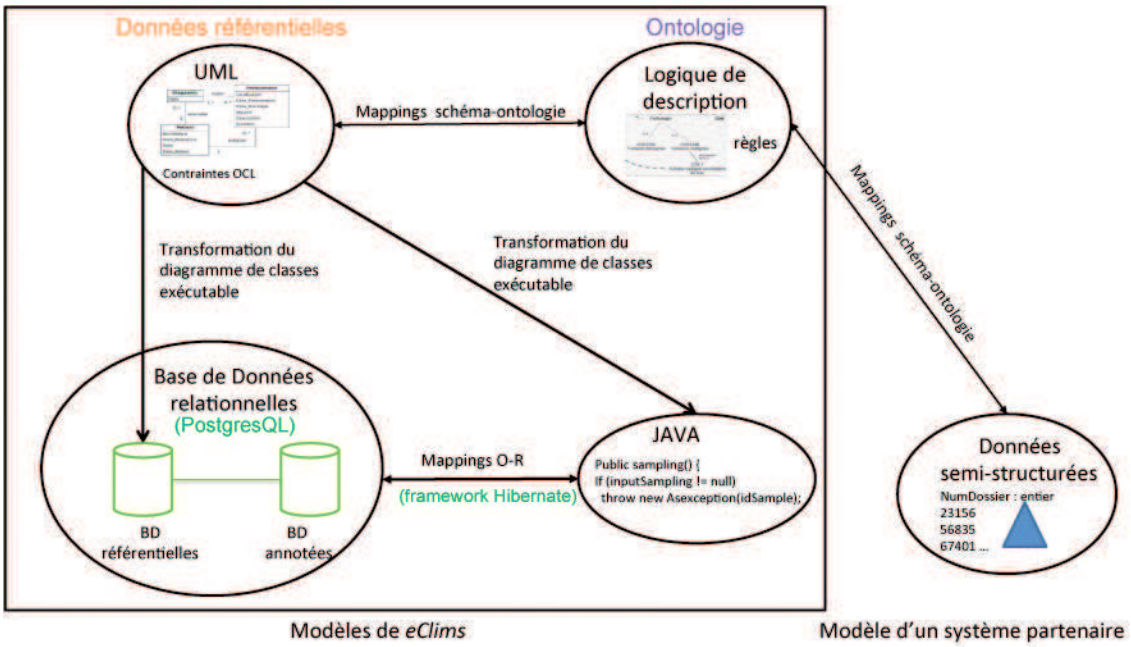


FIGURE 2.19 – Espaces technologiques mis en œuvre dans le composant *eClims*

```
Context Diagnostic
inv stades :
self.stade1 -> notEmpty() implies
        (self.maladie.annotation.stade2-> includes(self.stade1))
```

Afin de traiter le caractère dynamique de la connaissance de notre contexte nous avons choisi de mettre en place un système de gestion évolutif de la logique métier. Ce système repose sur la définition de règles s'appuyant sur les concepts et les relations entre concepts de notre ontologie. De récents travaux ont montré l'intérêt des règles au sein du Web Sémantique [HPS04, KRH08]. Ces travaux ont abouti à la définition d'un langage d'écriture des règles nommé *Semantic Web Rule Language*[31] (SWRL) combinant les langages OWL-DL[32] et RuleML[33]. Malheureusement, les règles écrites en SWRL peuvent se révéler indécidables et rendre l'ontologie inconsistante (section 6 de l'article de Horrocks et al. [HPSBT05]). Afin de ne traiter que des règles décidables, nous avons choisi de travailler en respectant les recommandations des règles DL-Safe [MSS05]. Ces règles sont décidables si elles travaillent sur des classes nommées de l'ontologie et sur un ensemble d'individus connus.
Les règles que nous avons retenues permettent de définir des connaissances qui ne sont pas directement modélisables dans l'ontologie. Seuls des experts du domaine peuvent définir quelles sont les règles devant être prises en compte afin de s'approcher au plus près de la réalité du travail des plate-formes de protéomique.

31. site Web : www.w3.org/Submission/SWRL
32. *Ontology Web Language - Description Logics* est une version intermédiaire de OWL permettant de restreindre certains constructeurs. *OWL 2 Web Ontology Language Document Overview*, site Web : www.w3.org/TR/owl2-overview
33. Rule Markup Language est un langage de balisage basé sur XML qui permet le stockage, l'échange, la récupération et la vérification de règles, site Web : www.ruleml.org

Par exemple, une règle métier énonçant *un prélèvement est valide si la pathologie pour lequel il est étudié et l'organe dont il provient sont mutuellement pertinents* , sera définie ainsi :

```
Prelevement(p), Provient(p,o), Organe(o), OrganeTouché(o, m),
Pathologie(m) => PrelevementValide(p)
```

Les contraintes OCL et les règles sont les garants de la qualité des données cliniques.

2.6 Synthèse

Notre synthèse est en deux points : une comparaison entre les modèles UML et les ontologies puis un résumé de notre approche de représentation des connaissances qui couple modèle UML et ontologie.

2.6.1 Comparaison modèle UML et ontologie

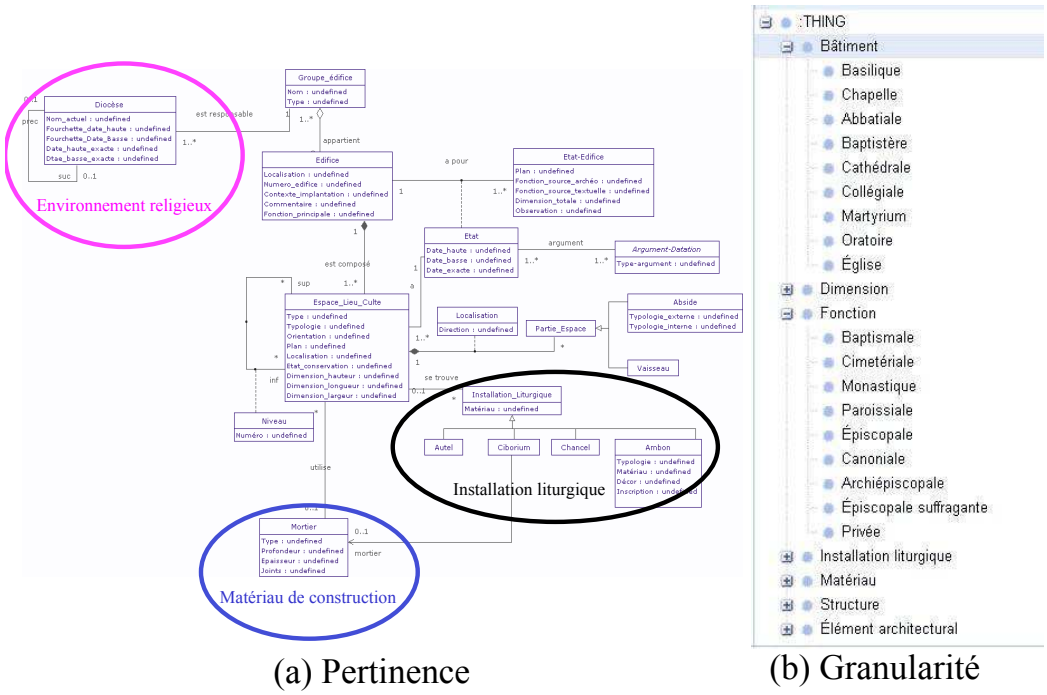
Nous pensons que les modèles conceptuels UML sont construits pour un Système d'Information spécifique et qu'ils ont comme objectif de définir, contraindre et limiter ce qui va être enregistré et manipulé dans le Système d'Information. Tandis qu'une ontologie explique un domaine pour révéler ce qui est un tout cohérent, et dont le but est d'être partagé par plusieurs applications. Les ontologies et les modèles sont similaires pour la représentation des connaissances d'un domaine dans la mesure où ils proposent des modes de description à base de concepts et de relations entre ces concepts. Selon Spear [Spe06], la description d'un domaine suppose un choix précis dans les limites des descriptions. Ces limites peuvent être appréhendées selon deux dimensions :

– la dimension horizontale ou pertinence a pour objectif de déterminer l'étendue de l'information qui sera incluse dans la description. Par exemple, si l'on représente le domaine de l'archéologie, la pertinence recouvre le choix d'inclure ou non des éléments aussi variés que les sources documentaires, la géologie, les techniques de construction ;
– la dimension verticale ou granularité a pour objectif de déterminer le niveau de détail de la représentation des connaissances. Par exemple, si l'on représente le domaine de l'archéologie, la granularité recouvre le choix de représenter un édifice de la structure des murs jusqu'aux éléments du décor, aux sols.

Il est difficile d'inclure dans un même modèle UML une description générale de certains éléments et des détails sur d'autres (dimension verticale), sauf à prendre le risque de construire un modèle difficile à comprendre et à maintenir. En revanche, un modèle peut utiliser différentes sources pour représenter la connaissance, et ainsi ajuster l'étendue (dimension horizontale) de la connaissance qu'il recouvre. La figure 2.20(a) présente un modèle UML combinant des connaissances issues des installations liturgiques avec des connaissances sur l'environnement historique et religieux, sur les matériaux de construction utilisés, etc. Ces connaissances sont au même niveau de détail c'est-à-dire une description à l'aide d'une classe éventuellement spécialisée pour le cœur de la connaissance à représenter. Elle illustre le fait qu'un modèle UML privilégie la dimension horizontale.

Au contraire, les langages de description d'ontologies offrent un grand degré de liberté pour gérer la granularité de la connaissance (dimension verticale) (voir figure 2.20(b)) car ils se focalisent sur la relation `is-a` et contrôlent strictement les autres relations utilisées.

En résumé, ces deux approches sont complémentaires :
- dans la modélisation : la granularité est apportée par les ontologies et l'étendue du domaine est apportée par les modèles UML ;



(a) Pertinence (b) Granularité

FIGURE 2.20 – Illustration des dimensions horizontale et verticale

- au niveau des capacités de raisonnement sous-jacentes : la classification des concepts et leurs propriétés peuvent être inférées dans les ontologies et vérifiées dans les modèles UML grâce aux contraintes OCL par exemple ;
- au niveau de la logique : l'hypothèse du monde ouvert des logiques de description utilisées dans le cadre du Web Sémantique en association avec les ontologies et l'hypothèse du monde clos associée aux modèles centrés données (exprimés par exemple par un diagramme de classes UML) est utilisée dans les SGBD. Motik et al. [MHS07] étudient plusieurs approches pour réconcilier ces deux hypothèses.
- au niveau du contrôle de la qualité et de la cohérence des données : les ontologies proposent le langage de règles SWRL, UML le langage OCL.

2.6.2 Résumé de notre approche

Nous considérons que la connaissance peut se réduire à un couple modèles-ontologies (voir figure 2.21). Le modèle UML permet une description « gros grain » du domaine (dimension horizontale) et délimite les données référentielles c'est-à-dire les données stables, le plus souvent utilisées et rarement modifiées. L'ontologie apporte la dimension verticale qui est utilisée à la fois

– pour la catégorisation du modèle : les instances du modèle sont liées aux individus de la description ontologique, l'ontologie définit les domaines de valeurs de certains attributs ; les concepts et les relations de la description ontologique peuvent être liées aux classes et relations du modèle ;
– et pour l'extensibilité de la structure des données à l'aide d'annotations dont la sémantique est garantie par l'ontologie.

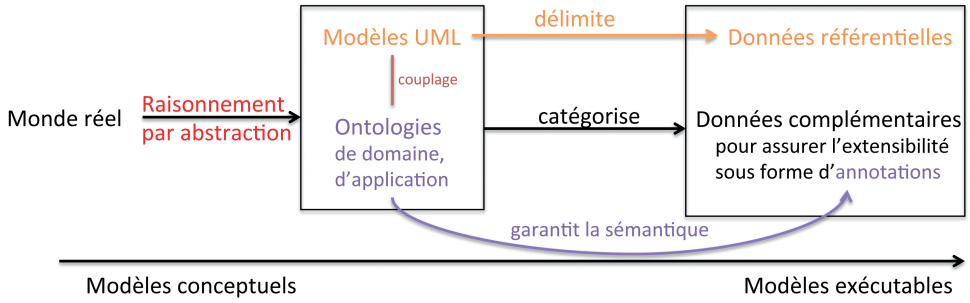


Figure 2.21 – Approche mise en place

La connaissance dans des domaines complexes fortement évolutifs ne pouvant pas être stabilisée dans les phases initiales de modélisation, la combinaison de deux langages de description du domaine (modèles UML et ontologies) permet de travailler avec une fiabilité raisonnable. Les données complémentaires (qui ne sont pas prises en compte dans la modélisation UML sur laquelle s'appuie la couche de persistance) sont traitées sous la forme d'annotations garanties par une ontologie.

Cette approche permet de modéliser de façon conceptuelle l'application, de gérer le raffinement fonctionnel jusqu'au modèle exécutable et de prendre en compte les préférences de l'utilisateur en vue de choisir une plate-forme répondant à leurs contraintes. La plate-forme doit garantir les contraintes (sécurité, traçabilité, anonymisation, etc.) qui ont été définies lors de la phase d'analyse. Les chapitres 4 et 5 ont comme plate-formes applicatives un wiki sémantique et un LIMS.

Publications associées

[1] Elie Abi-Lahoud, Marinette Savonnet, Marie-Noëlle Terrasse, Marco Viviani, Kokou Yétongnon, A Community-based Approach for Service-based Application Composition in an Ecosystem, *Web-Based Information Technologies & Distributed Systems, WITDS, Atlantis, Editors : A. Gabillon, M. Sheng et W.Mansoor*, 2009.

[2] Marinette Savonnet, Jean-Claude Simon, Marie-Noëlle Terrasse, Éric Leclercq, A top-down approach based on business patterns for web information systems design, *WISM'09 (CaiSE workshop)*, the Netherlands, 2009.

[3] Marie-Noëlle Terrasse, Marinette Savonnet, George Becker, A UML-based Metamodeling Architecture for Database Design, Proceedings of the International Symposium on Database Engineering and Applications (IDEAS'01), pp. 231-236, Munich, Germany, 2001

[4] Marie-Noëlle Terrasse, Marinette Savonnet, Éric Leclercq, Thierry Grison, George Becker, Points de vue croisés sur les notions de modèle et métamodèle, Premières Journées sur l'Ingénierie Dirigée par les Modèles (IDM'05), Paris, France, 29 Juin 2005

[5] Marie-Noëlle Terrasse, Marinette Savonnet, Éric Leclercq, Thierry Grison, George Becker, Do We Need Metamodels AND Ontologies for Engineering Platforms ?, First International Workshop on Global Integrated Model Management (G@mma), pp. 21-27, Shanghai, China, 2006

Chapitre 3

Annotations sémantiques

« Assis sur un fauteuil à vis, l'échine courbée en avant, [...] lit, annote, compile, rédige et enserre dans son cerveau la quintessence de quelques mille volumes qui garnissent sa chambre tout alentour. »
Nouvelles Genevoises de Rodolphe Töpffer http://gallica.bnf.fr/ark:/12148/btv1b86002289

Sommaire

Bringay et al. [BBC04] définissent une annotation comme « *une note particulière attachée à une cible. La cible peut être une collection de documents, un document, un segment de document (paragraphe, groupe de mots, mot, image ou partie d'image, etc.), une autre annotation.* [Une annotation peut donc être définie à différents niveaux de granularité.] *À une annotation correspond un contenu, matérialisé par une inscription, qui est une trace de la représentation mentale que l'annotation fait de la cible. Le contenu de l'annotation pourra être interprété à son tour par un autre lecteur . . .* ». Le processus d'annotation peut être automatique, manuel ou assisté par ordinateur. Une annotation peut être objective (auteur, date, etc.) ou subjective c'est-à-dire exprimer un point de vue sur la cible. Dans le cadre du Web Sémantique, une annotation, lorsqu'elle s'intéresse au contenu d'un document, est généralement appelée annotation sémantique. Les annotations sémantiques ont donc pour objectif d'exprimer le sens du contenu d'une ressource afin d'en améliorer sa compréhension, sa recherche, son traitement automatique et donc sa réutilisation par les utilisateurs finaux. Florence Amardheil dans [Ama07] définit l'annotation sémantique comme « *une représentation formelle d'un contenu, exprimée à l'aide de concepts, relations et instances décrits dans une ontologie, et reliée à la ressource documentaire source* ». C'est cette définition de l'annotation que nous adoptons pour la suite.

3.1 Modèles d'annotation existants

Les annotations sont utilisées aujourd'hui dans de nombreuses applications comme les blogs, les réseaux sociaux, les wikis sémantiques et les bases de données annotées. Suivant le domaine, l'annotation peut prendre plusieurs formes, de la forme la plus simple c'est-à-dire le tag à la forme la plus complexe le triplet ⟨`sujet, prédicat, objet`⟩ dont la nature des éléments est contrôlée par une ontologie et un ensemble de règles. Même si les formes syntaxiques des annotations sont variées, elles se réfèrent à deux bases formelles : les graphes conceptuels [1] proposés par Sowa [Sow84b] et la logique des prédicats du premier ordre.
Des travaux de recherche récents, débutés en 2007, concernent l'extension du modèle relationnel pour prendre en compte des annotations. Les formalismes des $\mathcal{K}$-relations et des semi-anneaux proposent une uniformisation des propositions spécifiques d'annotations dans les SGBD (bases de données incomplètes, probabilistes, temporelles, etc.). Green et al. [GKT07] ont étendu le modèle relationnel en considérant des $\mathcal{K}$-relations où les tuples sont annotés par une valeur d'un ensemble $\mathcal{K}$ muni de deux lois $\oplus$ et $\otimes$ avec deux éléments distincts 0 et 1. Dans le modèle relationnel, les tuples sont des fonctions $t : U \rightarrow \mathbb{D}$ avec U un ensemble fini d'attributs et $\mathbb{D}$ un domaine de valeurs, on note l'ensemble de tels tuples U-Tup. Une $\mathcal{K}$-relation r de schéma U est une fonction U-Tup $\rightarrow K$ dont le support $supp(R) = \{t | R(t) \neq 0\}$ est fini. Par exemple, le modèle relationnel standard correspond à $\mathcal{K}_{\mathbb{B}}$-relation avec $\mathcal{K}_{\mathbb{B}} = \langle \mathbb{B}, \vee, \wedge, false, true \rangle$ avec $\mathbb{B} = \{\text{true, false}\}$. Karvounarakis et Green [KG12] proposent un état de l'art sur les extensions de l'algèbre relationnelle en montrant que la structure $\langle K, \oplus, \otimes, 0, 1 \rangle$ doit être un semi-anneau commutatif pour conserver les propriétés de l'algèbre. Cependant ces travaux n'abordent pas la sémantique des annotations, c'est-à-dire leur lien avec la connaissance du domaine. De ce fait, il est très difficile de pouvoir vérifier la cohérence, la consistance des annotations par rapport à la connaissance et par rapport aux autres annotations (alors que les SGBD relationnels supportent ces propriétés).
La notion de graphe ou d'arbre a été reprise pour définir des modèles implémentés d'annotations. Bodganschi et al. [BS09] définissent une annotation comme un graphe orienté acyclique coloré qui est ensuite représenté en XML puis stocké dans la base de données MonetDB et interrogé en utilisant XQuery. Egyed-Szigmond et al. [EZPMP00] utilisent la bibliothèque LEDA de structure de graphes et XML Document Object Model (DOM) pour exprimer des annotations sur des documents multimédia. Bhatnagar et al. [BJR07] proposent un modèle d'annotation exprimé directement en XML. Chaque annotation est divisée en différentes sections XML : une section d'identification qui indique quel objet est annoté, une section niveau indique si l'objet est une base de données, une relation, un attribut, une ligne ou la valeur d'un attribut, une section donne la valeur de l'annotation et finalement une dernière section permet de faire référence à d'autres annotations. Les documents XML obtenus sont soit stockés dans un système de fichier soit dans une base de données relationnelles.
Depuis les années 2000, le langage RDF s'est peu à peu imposé comme un outil de représentation des annotations. Les annotations sont vues comme des triplets ⟨`sujet, prédicat, objet`⟩ où un sujet est lié à un objet par une relation définie par le prédicat. Du fait du caractère générique de RDF, plusieurs extensions, de la forme ⟨`sujet, prédicat, objet`⟩ `: valeur`, ont été définies de manière plus ou moins formelle. Ces extensions traitent par exemple de données temporelles [GHV07] où `valeur` définit l'intervalle de temps durant lequel le triplet est valide, d'imprécision [Str09] où `valeur` est un réel pris dans [-1,1] avec -1 indiquant un faible niveau de croyance et 1 un fort niveau de croyance ou de provenance [DSSS09] où `valeur` est un littéral indiquant la provenance du triplet. Lopes et al. [LPSZ10] ont proposé le langage AnQL qui

1. Le chapitre 4 de l'ouvrage de Chein et Mugnier [CM09] présente l'état des connaissances sur les graphes conceptuels et leur cadre théorique.

inclut des caractéristiques de SPARQL 1.1 pour interroger du RDF annoté. Zimmerman et al. se basant sur les travaux de Udrea et al. [URS10] proposent de traiter les extensions de RDF *via* des annotations, c'est-à-dire avec un niveau supplémentaire de méta-données exprimées sur la composante prédicat afin de contourner la réification [ZLPS12]. RDF permet ainsi de définir de nombreuses annotations sur une même donnée (sujet). Cependant, comme les éléments présents dans une annotation RDF ne sont pas contrôlés, il est difficile de s'assurer de la qualité d'une annotation. RDFS permet de restreindre le vocabulaire utilisé et donc de pouvoir effectuer des déductions à partir de triplets existants. Quelques approches sont proposées pour RDF [ACP10, LMS08] ou XML [BT10] afin de prendre en compte la notion de contraintes d'intégrité. Un travail similaire [BCG+10] a été réalisé entre RDF et les graphes conceptuels.

En conclusion, les récents travaux autour des extensions de RDF cherchent à inclure dans RDF des éléments de la sémantique du domaine alors que nous pensons que celle-ci doit être découplée de la connaissance sur les données. Les approches de type *annotated RDF* qui fixent les annotations complémentaires sur le prédicat sont ambiguës (même si elles ont une base formelle claire). Dans leur utilisation, elles peuvent signifier un ajout de connaissance sur le prédicat lui-même ou sur l'ensemble de l'annotation et donc mélanger le niveau connaissance du domaine avec le niveau connaissance sur les données.

3.2 Stockage des annotations

Les paragraphes suivants présentent deux modes de stockage des annotations, l'un spécifique dans les Bases de Données relationnelles, l'autre plus général avec les bases de données NoSQL.

3.2.1 Bases de Données annotées

Mon travail dans le domaine des Sciences du Vivant m'a amené à étudier les bases de données nettoyées (*curated databases*) et les bases de données annotées (*annotated database*) qui ont été proposées par les bio-informaticiens pour répondre à des exigences de qualité.
Le terme de Base de Données nettoyées est associé à des Bases de Données qui sont mises à jour avec un investissement humain considérable [Bun09]. Par exemple, il y a environ 150 experts (les curateurs) qui travaillent à temps-plein sur la base de données UniProt [2]. La plupart de ces Bases de Données jouent le rôle de publications scientifiques de référence. La majorité contient un schéma très simple qui représente les données « core », et progressivement la structure du schéma évolue en fonction des découvertes réalisées.

Dans une Base de Données annotées, les annotations sont utilisées pour permettre une meilleure compréhension des données [3]. Le SGBD doit offrir des mécanismes pour créer, stocker et interroger des annotations portant sur les données de la base. Le modèle d'annotation doit être extensible et neutre pour la Base de Données et indépendant des plate-formes cibles. La possibilité d'annoter à différentes granularités de façon transparente (comment et où stocker les annotations pour réaliser cette transparence) est une caractéristique essentielle. Plusieurs mécanismes de stockage des annotations au sein d'un SGBD relationnel ont été développés, la plupart sont des variantes des trois mécanismes suivants :

1. le premier mécanisme stocke une référence à l'annotation au sein même du tuple à l'aide de colonnes supplémentaires (une par champ) dans la table. Par exemple, le système

2. www.uniprot.org/

3. Les annotations indiquent comment la donnée a été obtenue, pourquoi certaines valeurs ont été ajoutées ou modifiées, quelles expériences ou analyses ont été exécutées pour les obtenir, etc. Li et al. [LLC08] présentent une comparaison de divers systèmes d'annotations.

DBNotes (DataBase anNOTation ManagEment System) [4] [CTV05, BCTV05] utilise la forme la plus simple de stockage puisque à chaque attribut de chaque table est associé un autre attribut qui contiendra l' (les) annotation(s), une relation R(A,B) devient alors R'(A, A_a, B, B_a) ;

2. le second mécanisme stocke dans une seule table particulière toutes les annotations de la base avec une référence (logique ou physique) à la donnée annotée ;
3. le dernier mécanisme crée, pour chaque table dont les données sont annotées, une table d'annotations. Le système bdbms (*biological database management system*) [EAE+09, EOA+08, EOA07] permet d'associer plusieurs relations d'annotation à une relation. Chaque annotation a un identifiant unique, sa valeur et l'ensemble des cellules concernées par l'annotation. Les cellules concernées sont identifiées de la façon suivante : dans la table contenant les données annotées, les attributs sont associés à un numéro correspondant à l'ordre physique de création de l'attribut dans la table et les lignes sont associées à un numéro basé sur leur ordre d'insertion dans la table. Ces numéros permettent de déterminer la première et la dernière cellule touchées par l'annotation. La table d'annotation aura autant de lignes que de rectangles couverts par la région annotée. Les avantages de ce type de stockage sont de deux ordres : 1) la structure des tables contenant les données annotées n'est pas affectée par les annotations ; 2) les annotations sont définies selon la granularité désirée. Néanmoins ces avantages sont contrebalancés par les inconvénients liés à la manipulation des données. En effet, l'ajout ou la suppression d'une donnée au sein d'une table ont des répercussions sur les annotations présentes car elle oblige à recalculer les plages de cellules couvertes par les annotations. Par exemple, la suppression d'un tuple d'une relation oblige ce système soit à conserver virtuellement ce tuple (afin de conserver la numérotation horizontale des cellules) soit à décaler toutes les plages de cellules.

Des extensions au langage SQL ont été apportées afin de pouvoir consulter les annotations : 1) l'approche connue sous le nom d'*annotation propagation* consiste à effectuer les requêtes uniquement sur les données puis à ajouter les annotations au résultat ; 2) l'approche connue sous le nom d'*annotation querying* consiste à effectuer les requêtes à la fois sur les données et sur les annotations.

Nous remarquons que les outils développés dans le cadre du Web Sémantique sont très peu utilisés. De plus, les mécanismes de contrôle d'intégrité référentielle sont peu abordés et plus généralement les contraintes ne sont pas traitées dans le SGBD. On a alors recours à des contraintes vérifiées dans les applications ce qui freine l'extensibilité et l'évolution des applications développées avec ces systèmes.

3.2.2 Approches *NoSQL*

Les Bases de Données Relationnelles ne sont pas adaptées aux environnements distribués traitant des volumes importants de données hétérogènes. Depuis quelques années, de nouvelles approches de stockage et de manipulation de données qui ne suivent pas les principes des Bases de Données relationnelles sont apparues. Ces approches regroupées sous le terme générique de *NoSQL* pour *Not Only SQL* sont portés par des acteurs du Web comme Google, Facebook, Twitter. On classifie généralement les bases *NoSQL* en quatre catégories :

– les bases clé-valeur (Redis [5], Voldemort [6], Riak [7]) ont un modèle très simple, chaque objet est identifié par une clé unique, la structure de l'objet est libre (XML, JSON, etc.). Le langage de

4. http://users.soe.ucsc.edu/~wctan/Projects/dbnotes/index.html
5. Redis : http://redis.io/
6. Voldemort : http://project-voldemort.com/
7. Riak : http://wiki.basho.com/

requêtes est aussi très rudimentaire, il ne permet généralement que l'utilisation des opérations CRUD (Create, Read, Update et Delete). Ces bases sont utilisées comme dépôt de données et peuvent fonctionner facilement en cluster ;
– les bases orientées colonnes (Cassandra [8], HBase [9], Google Big Table, MonetDB [10]) sont une extension des bases clé-valeur. La colonne est l'entité de base représentant un attribut d'une donnée, chaque colonne est définie par un couple clé-valeur. Cassandra et HBase sont utilisées dans des projets très connus (Google, twitter ou digg.com) et supportent le passage à l'échelle, MonetDB est utilisée pour plusieurs applications de Bases de Données scientifiques ;
– les bases documentaires (MongoDB [11], CouchDB [12]) sont constituées d'un ensemble de documents hétérogènes. Il n'est pas nécessaire de définir au préalable les attributs utilisés dans un document, ces bases sont qualifiées de *schemaless* ;
– les bases orientées graphe (Neo4j [13], HyperGraphDB [14], OrientDB [15]) sont constituées d'objets appelés nœuds, il est possible de décrire des arcs (relations entre les objets) qui sont orientés et qui disposent, tout comme les objets, de propriétés.

Les bases de données clé-valeur et les bases de données orientées colonnes ont un schéma de données très simple et peuvent emprunter les modes de gestion des contraintes du modèle relationnel.
Les bases de données orientées graphe permettent de prendre en compte de gros volume d'annotations fortement liées les unes aux autres [HWL12]. Dans le cadre de Bases de Données scientifiques les données dépendent de la connaissance en cours au moment de leur production et nécessitent des requêtes complexes pour explorer les relations entre données, d'après Vicknair et al. [VMZ+10] les bases de données orientées graphe ont de meilleures performances que les bases de données relationnelles pour les opérations récursives.
Les annotations sont souvent représentées sous la forme de graphe ou d'arbre, ce qui rend les bases de données orientées graphe bien adaptées à leur stockage. Les annotations sont aussi représentées sous XML ce qui rend les bases clé-valeur et orientées colonnes comme une solution possible à leur stockage.

Il convient de remarquer que les modes de stockage ne sont pas exclusifs, une base de données scientifiques peut utiliser un SGBD relationnel, une base de données orientée colonne pour stocker les jeux de données et une base orientée graphe pour les annotations [Gho10].

3.3 Définition et sémantique de notre modèle d'annotation

Nous proposons un modèle d'annotation couplé à une ontologie et des règles qui permettent de découpler la connaissance du domaine de la connaissance sur les données.
Dans la suite, nous présentons les définitions de notre modèle pour lui donner une sémantique en faisant abstraction de l'environnement technologique dans lequel il sera utilisé (RDF, triple-store ad-hoc, Bases de Données, etc.) puis nous présentons sa syntaxe abstraite.

Notre modèle repose sur la définition d'une annotation sous la forme d'un triplet liant le sujet s (donnée que l'on souhaite annoter) et un objet o (contenu de l'information que l'on souhaite ajouter) au moyen d'un prédicat p qui identifie la sémantique de l'annotation. Cette syntaxe

8. Cassandra : http://cassandra.apache.org/
9. HBase : http://project-voldemort.com/
10. MonetDB : http://www.monetdb.org/Home
11. MongoDB : http://www.mongodb.org/
12. CouchDB : http://couchdb.apache.org/
13. Neo4j : http://neo4j.org/
14. HyperGraphDB : http://www.kobrix.com/hgdb.jsp
15. OrientDB : http://www.orientdb.org/index.htm

générale $\langle s, p, o\rangle$ peut être traduite en RDF. En fonction du domaine, nous incluons une forme de contrainte avec l'ontologie.

Soient $\mathcal{O} = \langle \mathcal{C}, \mathcal{P}, \mathcal{I}\rangle$ une ontologie définie par un ensemble de concepts, de propriétés et d'instances, $\mathcal{A}$ l'ensemble des annotations, U et L deux ensembles désignant respectivement les URI et les littéraux. Tout élément des ensembles $\mathcal{C}, \mathcal{P}, \mathcal{I}, \mathcal{A}$ peut être désigné par un élément de U. Soit $\mu : U \rightarrow \mathcal{C} \cup \mathcal{P} \cup \mathcal{I} \cup \mathcal{A}$ la fonction d'association d'un URI à un élément d'un des ensembles $\mathcal{C}, \mathcal{P}, \mathcal{I}, \mathcal{A}$. De plus, nous définissons U_C comme la restriction de la fonction inverse de μ à $\mathcal{C}$: $U_{\mathcal{C}} = \mu^{-1}_{|\mathcal{C}}$. De la même manière, on définit $U_{\mathcal{P}}$, $U_{\mathcal{I}}$ ainsi que les combinaisons deux à deux des ensembles.
Une annotation est un triplet $a = \langle s, p, o\rangle \in U \times U_{CP} \times (U \cup L \cup null)$.
Notre modèle d'annotation nous permet de définir trois structures de base pour les annotations :

1. une annotation simple $\langle s, p, o\rangle$ permet d'annoter un sujet en lui associant un couple (propriété, objet) avec s et p toujours non *null*. Si o est *null* et si p fait référence à un concept, l'annotation spécifie le type du sujet. Cela peut être assimilé à une contrainte de type restriction sur le domaine d'un attribut. Si o n'est pas *null*, o est soit un littéral soit un individu du concept spécifié par p. Cela peut être assimilé à une contrainte de Base de Données qui vérifie que la valeur d'un attribut est prise dans une liste. Ces annotations sont liées au niveau ABox de l'ontologie ;
2. une annotation complexe ou n-aire (notée A-cplx dans le tableau 3.1) permet de mettre en relation un même sujet avec deux ou plusieurs couples (propriété, objet), o pouvant être un littéral, un individu, ou une référence à un autre sujet. Tous les prédicats utilisés doivent être différents ;
3. une annotation réflexive permet d'expliquer ou de préciser pourquoi/comment l'objet et le prédicat sont liés au sujet. L'annotation réflexive peut elle-même être simple ou complexe, l'objet devient alors un sujet. Elle possède différents niveaux (obtenus en utilisant le système de parenthèses) : une annotation de niveau i explique la relation entre l'objet, le prédicat et le sujet de l'annotation de niveau i-1. Si toutes les annotations du niveau i sont contenues dans une liste alors elles partagent toutes le même sujet o.

Par définition, nous ne pouvons pas utiliser le prédicat comme sujet. Dans notre modèle, le prédicat qui est un terme d'une ontologie ne peut pas faire l'objet d'annotation car la connaissance sur ses éléments est exprimée dans l'ontologie exclusivement.

Il est facile de définir un isomorphisme entre la représentation des annotations sous forme de graphe et leur représentation sous la forme de chaînes de caractères. En effet, les nœuds du graphe correspondent aux ressources impliquées dans l'annotation (sujet ou objet) et les arcs correspondent aux relations entre ces ressources (prédicat). Par exemple, soit l'arbre d'annotation donné en figure 3.1, il peut être écrit sous la forme d'une chaîne de caractères (voir tableau 3.1) :
– au premier niveau : $((s, p_1, o_1), (s, p_5, o_5), (s, p_6, o_6))$
– au second niveau : $((s, p_1, o_1)((o_1, p_2, o_2), (o_1, p_3, o_3), (o_1, p_7, o_7)), (s, p_5, o_5), (s, p_6, o_6))$
– au dernier niveau :
$((s, p_1, o_1)((o_1, p_2, o_2), (o_1, p_3, o_3)((o_3, p_4, o_4)), (o_1, p_7, o_7)((o_7, p_8, o_8), (o_7, p_9, o_9))), (s, p_5, o_5), (s, p_6, o_6))$
Sa représentation sous une forme plus compacte est donnée par :
$s(p_1, o_1((p_2, o_2), (p_3, o_3)(p_4, o_4), (p_7, o_7)((p_8, o_8), (p_9, o_9)), (p_5, o_5), (p_6, o_6))$

Une annotation exprimée sur un sujet contribue à définir son contexte selon deux dimensions : 1) la dimension référentielle inclut la structure de l'ontologie et les règles portant sur les concepts utilisés ; 2) la dimension assertionnelle est l'ensemble de toutes les annotations portant sur le même sujet. Une annotation est valide seulement si elle est consistante vis-à-vis des deux dimensions.

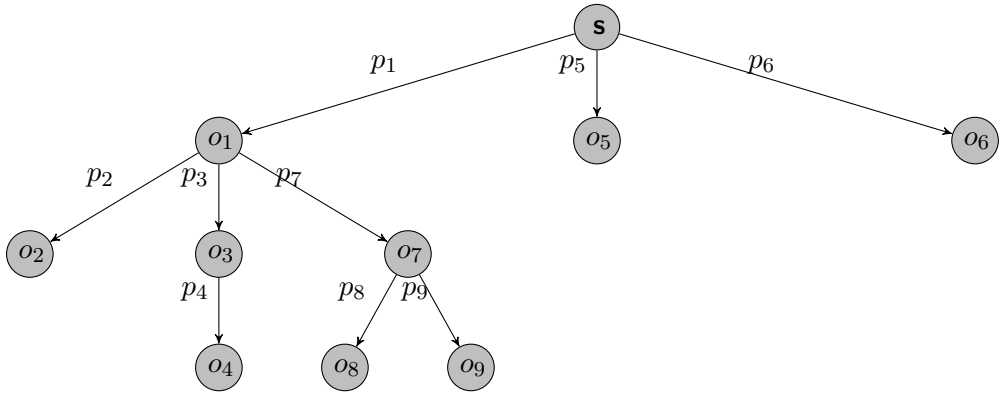


FIGURE 3.1 – Exemple d'un arbre d'annotation

⟨A⟩→⟨A-simple⟩| ⟨A-cplx⟩| ⟨A-reflexive⟩
⟨A-simple⟩→(⟨s⟩, ⟨p⟩, ⟨o⟩)
⟨A-cplx⟩→(⟨A-simple⟩, ⟨A-liste⟩)
⟨A-liste⟩ → ⟨A-simple⟩ | ⟨A-simple⟩, ⟨A-liste⟩ | ⟨A-reflexive⟩
⟨A-reflexive⟩→(⟨s⟩, ⟨p⟩, ⟨o⟩⟨A⟩) | (⟨s⟩, ⟨p⟩, ⟨o⟩⟨A⟩⟨A-ref-liste⟩)
⟨A-ref-liste⟩→ ⟨A-simple⟩|⟨A-simple⟩⟨A-ref-liste⟩
⟨s⟩→ **URI** | **URL**
⟨p⟩→ **concept de l'ontologie** | **propriété de l'ontologie**
⟨o⟩→ **individu de l'ontologie** | **littéral** | **URI** | **URL** | **null**

TABLE 3.1 – Syntaxe abstraite de notre modèle d'annotation

3.4 Sémantique du processus d'annotation : une analogie avec la sémantique des langages de programmation

L'ontologie et la syntaxe abstraite jouent le rôle de DSL (*Domain Specific Language*) puisqu'elles permettent d'exprimer quelles sont les annotations syntaxiquement correctes. Afin de contrôler la consistance des annotations (deux annotations ne peuvent pas être contradictoires) nous développons un ensemble de mécanismes basés sur les travaux dans la sémantique des langages de programmation afin de nous assurer que le processus d'annotation est conforme à la syntaxe abstraite et à la connaissance du domaine (voir figure 3.2).

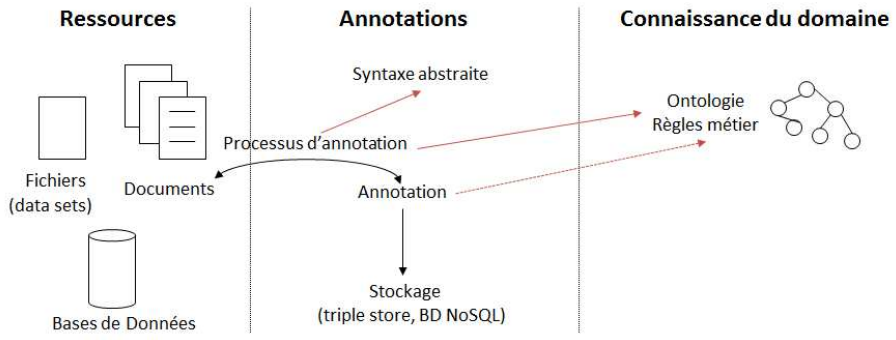


FIGURE 3.2 – Aperçu de la proposition

Processus d'annotation et sémantique axiomatique La sémantique axiomatique voit un programme comme une transformation de propriétés logiques, la signification d'un programme est donnée par un ensemble de prédicats qui sont vérifiés par l'état de la machine (la mémoire) à condition qu'un autre ensemble de prédicats ait été vérifié avant l'exécution. Les propriétés de sémantique axiomatique s'expriment, en général, sous la forme d'expressions de la logique de Hoare :

$$\{P\}I\{Q\}$$

où P et Q sont des propriétés exprimées dans la logique des prédicats, P est censé être vérifié par la mémoire avant exécution des instructions I, et Q doit être vérifié après exécution de I sur une machine vérifiant P.

Cette sémantique considère le processus d'annotation comme une transformation des propriétés attachées à un sujet. La cohérence du processus est donnée par le fait que les règles validées avant la nouvelle annotation doivent rester valides après son ajout.
Ce type de règles sont des règles globales connues qui doivent toujours rester vraies, elles nécessitent donc un mode de raisonnement avec l'hypothèse du monde clos. Par exemple, la règle « un édifice dédié à un Saint doit avoir été construit après la sanctification du personnage » est une règle globale qui doit être vérifiée quelles que soient les annotations, elle peut s'exprimer en logique du premier ordre par :
`isConsecrated(?b,?p)` $\leftarrow$ `hasConstructionDate(?b,?d1)` $\wedge$
`hasDateDead(?p,?d2)` $\wedge$ `d1` $\geq$ `d2`.

Processus d'annotation et sémantique dénotationnelle Dans la sémantique dénotationnelle, le programme est vu comme une fonction mathématique qui représente l'effet du programme sur un état de la mémoire.

$$S \rightarrow S' = [|A|](S) \textit{ où } [|A|](S) \textit{ est l'évaluation de } A \textit{ dans l'état mémoire } S$$

La sémantique dénotationnelle du processus d'annotation exprime la correspondance entre la structure de l'annotation et la sémantique du domaine au moyen de fonctions de l'ensemble des termes utilisés vers les concepts de l'ontologie. Par exemple, cette sémantique est utilisée pour vérifier si une propriété de l'ontologie peut être associée avec le sujet d'une annotation en utilisant les mécanismes de raisonnement et d'inférence afférents aux logiques de description avec l'hypothèse du monde ouvert. Elle a été mise en œuvre avec l'assistant d'annotation développé pour *WikiBridge* (voir section suivante).

Processus d'annotation et sémantique opérationnelle La sémantique opérationnelle voit un programme comme un système de transition d'états.

La sémantique opérationnelle des annotations assimile le processus à un changement d'état vu non pas sous l'angle des propriétés comme dans la sémantique axiomatique mais sous la forme des états accessibles à partir d'un état donné. Les évolutions valides sont déterminées par les changements de valeur dans une des composantes de l'un des triplets. Les techniques de *model checking* basées sur des automates finis semblent être utiles dans ce cas.
Cette sémantique permet de restituer tous les états possibles, d'analyser et observer tous les changements d'état possibles c'est-à-dire de procéder à une différence entre deux états, de caractériser les évolutions c'est-à-dire le processus de changement d'état.
Dans le projet CARE, nous utilisons trois concepts essentiels pour modéliser l'évolution des édifices : l'usage religieux ou fonction, les espaces, le temps. L'édifice est représenté par un ensemble d'annotations qui le situe dans un espace à trois dimensions. Soit $\mathcal{U}$ l'ensemble des usages, $\mathcal{E}$ l'ensemble des entités spatiales et $\mathcal{T}$ l'ensemble des entités temporelles. Un édifice a

est sous-ensemble du produit cartésien des trois ensembles $\mathcal{U} \times \mathcal{E} \times \mathcal{T}$, soit $a = \{(u, e, t), f \in \mathcal{U}, e \in \mathcal{E}, t \in \mathcal{T}, \}$. Le principe d'autonomie de la fonction, de l'espace et du temps permet d'observer les facteurs influant sur le changement, de restreindre l'étude à des produits deux à deux des trois ensembles et d'estimer ainsi le rôle ou la prépondérance de l'un par rapport à l'autre.

3.5 Mise en œuvre des annotations dans deux domaines

Deux applications que nous présentons dans la suite, ont été développées pour valider le modèle et la méthodologie des annotations : *WikiBridge* dans le domaine de l'archéologie avec comme plate-forme un wiki et *eClims* dans le domaine de la protéomique clinique avec comme plate-forme un LIMS.
La première application est orientée documents et met en évidence l'utilisation de formulaires pour représenter les données référentielles et l'utilisation de l'ontologie afin de proposer aux utilisateurs un assistant d'annotation qui contrôle la structure de l'annotation définie. La seconde application repose sur un stockage d'annotations associé à une Base de Données relationnelles (pour les données référentielles) sous la forme de Base de Données orientées colonnes qui met en évidence les capacités d'extension fournies par les annotations. Dans les deux cas, notre objectif est de dépasser le modèle des Bases de Données classiques en gérant non seulement la variabilité des données mais aussi l'évolution de la connaissance qui entraînent toutes deux une évolution du schéma de la Base de Données.

3.5.1 Annotations dans *WikiBridge*

WikiBridge est un wiki sémantique présenté en détail dans le chapitre 4 : les pages du wiki (c'est-à-dire les documents) sont structurées à l'aide de formulaires qui mettent en évidence les grandes entités modélisées auxquelles se superposent des annotations. Le mécanisme d'annotation permet de traiter la variabilité des acteurs et des structures de données en fournissant une association entre les concepts de l'ontologie et une portion du document. Ce mécanisme permet d'éviter la modification du schéma général du formulaire afin de prendre en compte toutes les spécificités liées aux différents acteurs.
Nous offrons dans *WikiBridge* des annotations automatiques à partir des sections du formulaire renseignées (faites à gros grain) et des annotations manuelles réalisées par les experts du domaine. Les méta-données (par qui et quand) sont implémentées en natif par le wiki.

Modèle d'annotation dans *WikiBridge*

Dans *WikiBridge*, le mécanisme d'annotation permet d'annoter n'importe quel élément (portion de texte, image, lien, etc.) du document en lui associant des propriétés et des valeurs sélectionnées dans les termes d'une ontologie. Ce processus est sensible au contexte car les termes sont sélectionnés dans l'ontologie par rapport au champ actif du formulaire.
La figure 3.3 montre une annotation complexe qui associe deux prédicats (`Accès` et `Dimensions`) à un sujet dont le type est spécifié par le prédicat `InstallationLiturgique`. Le prédicat `Dimensions` comporte une liste de prédicats simples spécifiant la `Hauteur`, la `Largeur` et la `Longueur`. Cette annotation se traduit par un graphe de triplets présenté en figure 3.4.
L'annotation engendrée par la grammaire (donnée dans le tableau 3.1) est la suivante :

```
( (http://care.u-bourgogne.fr/care/index.php/Saint-Clement#Autel123,
              InstallationLiturgique, Autel),
```

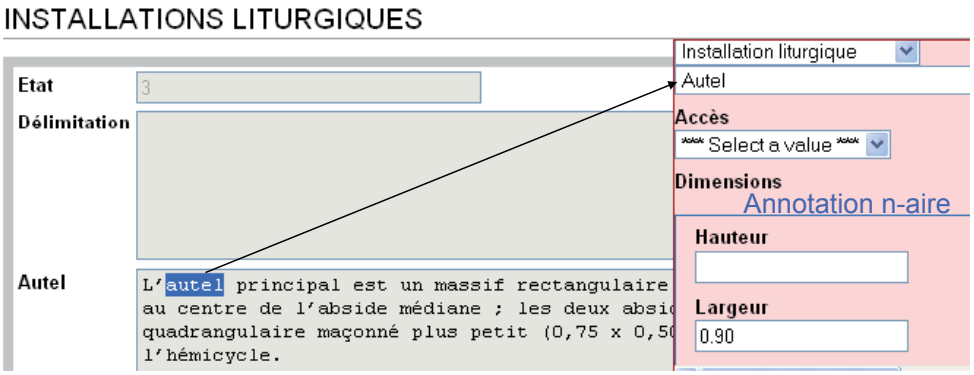


Figure 3.3 – Exemple d'annotation complexe dans *WikiBridge*

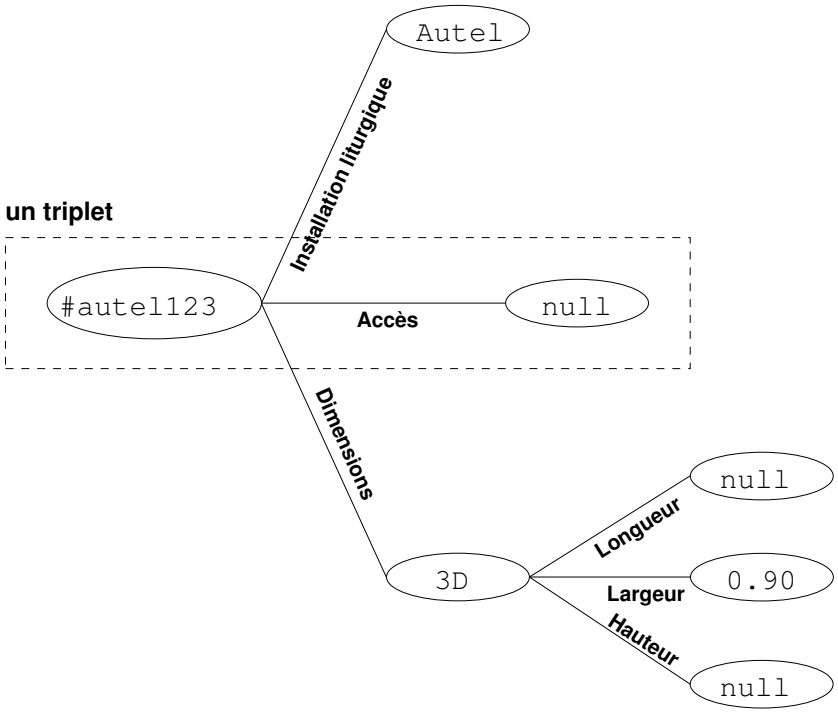


Figure 3.4 – Annotation de la figure 3.3 représentée sous la forme d'un graphe de triplets

```
  (Autel, Accès, null)
  (Autel, Dimensions, 2D
   ((2D, Hauteur, null),
    (2D, Largeur, 0.90))
  )
)
```

Validité des annotations dans *WikiBridge*

Bien qu'il soit toujours possible d'écrire une annotation dans un document en utilisant la syntaxe du wiki, l'assistant d'annotation permet un contrôle de la syntaxe des annotations, en utilisant l'ontologie, selon trois niveaux :

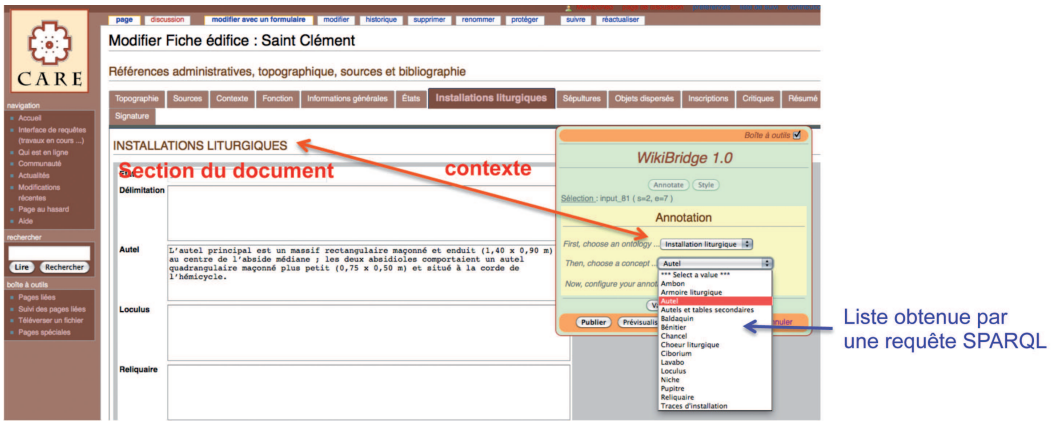


FIGURE 3.5 – Assistant d'annotation dans *WikiBridge*

1. le typage : un sujet possède un type défini par des annotations de la forme ⟨s, p, null⟩. Par exemple, dans la figure 3.5, le champ actif du formulaire se situe au niveau de la section `INSTALLATIONS LITURGIQUES` du document, alors automatiquement le sujet est typé par le concept `InstallationLiturgique` et les termes sont sélectionnés par rapport à ce concept ;
2. les domaines de valeurs de la composante *o* de l'annotation en utilisant les possibilités fournies par la ABox : les domaines sont contrôlés par l'ensemble des individus associé au concept/prédicat déterminé par le prédicat, l'ensemble des individus possibles est inféré ;
3. la cohérence des prédicats en utilisant les possibilités fournies par la TBox. Il est impossible de créer une annotation complexe sur un sujet en utilisant des prédicats qui ne sont pas applicables à son type. Par exemple, dans la figure 3.3, le sujet `autel123` est une installation liturgique et ce type d'installation a pour prédicat `Accès` et `Dimensions`.

Cependant au delà de la syntaxe des annotations, il existe des règles de cohérence plus globales dépendantes fortement du domaine. Par exemple il n'existe pas en France du IV^{e} au X^{e} siècle, d'église avec des murs construits en terre crue alors que ce type de construction peut exister en Irlande. Toutes les contraintes ne peuvent donc pas être vérifiées avec la structure de l'ontologie et le mécanisme d'inférence des logiques de description.

Caractéristiques du modèle d'annotation de *WikiBridge*

Oren et al. dans [OMS+06] ont analysé les annotations dans différents domaines et ont proposé la classification suivante :

- la dimension **association** est la façon dont l'annotation est associée avec la ressource annotée : est-ce que l'annotation est stockée avec (*embedded*) la ressource annotée ou est-ce que l'annotation fait référence de façon externe à la ressource ;
- la dimension **granularité** indique si l'annotation concerne un document, une section à l'intérieur du document, une phrase, un mot ;
- la dimension **terminologie** indique si une annotation a sa propre terminologie ou si l'annotation utilise des termes d'une ou plusieurs ontologies. Ils distinguent trois catégories d'annotations : 1) les annotations informelles ; 2) les annotations formelles qui ont des composants définis formellement et sont ainsi lisibles par des machines ; 3) les annotations ontologiques dont les composants sont définis formellement et utilisent des termes d'une ontologie formellement définie. Cette dernière catégorie permet d'assurer l'échange et la compréhension des annotations par les différents composants et acteurs d'une application ;

– la dimension **type d'objet** indique si l'objet est un littéral ou un objet structurel incluant un lien hypertexte vers une autre ressource ou un terme ontologique ;
– la dimension **contexte** représente quand et qui a fait l'annotation ainsi que l'étendue spatiale et temporelle de cette dernière ;
– la dimension **flexibilité** indique si les annotations peuvent être détruites et s'il existe un composant pour les visualiser ;
– la dimension **sortie** indique si l'annotation est écrite ou non dans un standard du W3C.

Jacques Virbel de l'IRIT ajoute la dimension **usage** qui décrit l'objectif visé par l'annotation à savoir : contextualiser (borner le sens) ; architecturer (typer des fragments) ; corréler (relier deux fragments) ; programmer (projeter une action : traduire, relire, analyser, etc.) ; reformuler ; hiérarchiser (attribuer un niveau d'importance à un fragment de texte) ; commenter (critiquer, associer une nouvelle idée) ; documenter (ajouter un fragment pour en comprendre un autre).

Les tableaux suivants reprennent les critères d'Oren et les objectifs de Virbel et indiquent comment ils sont mis en œuvre dans *WikiBridge*. *WikiBridge* stocke à la fois les annotations dans le document et dans un triple store. Le mécanisme d'annotation permet d'annoter n'importe quel élément du document (paragraphe, mot, image, lien, etc.) en sélectionnant des termes dans une ontologie, il respecte les standards du W3C. Le contexte est réalisé nativement par le wiki. La dimension usage est assurée par l'expressivité de notre modèle d'annotation.

Critères de Oren et al.	*WikiBridge*
association	page courante & Triple store
granularité	document & fragment du document
terminologie	ontologie
type d'objet	littéral & terme d'ontologie & URI & URL
contexte	nativement par le wiki
flexibilité	oui
sortie	respect des standards du W3C

Table 3.2 – Positionnement de *WikiBridge* par rapport aux critères de Oren et al.

Usages de Virbel	*WikiBridge*
contextualiser	ontologie
typer des fragments	possibilité *o* null
corréler	annotation réflexive
analyser/interpréter	annotation automatique et posée par un expert
reformuler	ontologie vue comme un vocabulaire commun
hiérarchiser	formulaire fourni par le wiki
commenter	natif dans le wiki par les pages de discussion
documenter	natif dans le wiki (ajout de documents multimédia, liens, etc.)

Table 3.3 – Positionnement de *WikiBridge* par rapport aux objectifs de Virbel

3.5.2 Annotations dans *eClims*

eClims est un composant clinique d'un LIMS détaillé dans le chapitre 5. Il a été développé pour la plate-forme de protéomique CLIPP qui réalise des études de protéomique clinique. Par exemple, la plate-forme CLIPP mène deux études complètement différentes, l'une sur les leucémies, l'autre sur la recherche de marqueurs salivaires liés aux préférences gustatives chez les nourrissons. Dans les deux études pré-citées, les données communes sont l'âge et le sexe. Les données complémentaires sont pour la première étude le diamètre de la rate, la pâleur du patient, le nombre de plaquettes et pour la seconde l'âge d'apparition successif de chaque dent et le type d'allaitement (au sein, lait maternisé ou mixte). Nous nous trouvons face à une très grande variabilité de la structure des données quand les SGBD relationnels demandent de déclarer au préalable l'ensemble des attributs représentant un objet.
Nous proposons de traiter le besoin d'extensibilité de la structure de données par un mécanisme d'annotation qui permet d'ajouter dynamiquement, sans modifier les applications, des données. Les données sont divisées en deux catégories : les données référentielles communes à toutes les études sont stockées dans un SGBDR auxquelles viennent s'ajouter des données complémentaires traitées sous forme d'annotations.

Modèle d'annotation dans *eClims*

Le modèle d'annotation dans *eClims* reprend la forme de base du triplet $\langle s, p, o \rangle$.
Définition 1. Un sujet d'annotation s est un couple $(Conteneur, IdSujet)$ où $Conteneur$ est le nom du conteneur de l'objet identifié par une valeur $IdSujet$.
Dans le cas de l'utilisation d'un SGBDR, le conteneur est composé d'un nom de la table, préfixé par la chaîne de connexion permettant de se connecter à la base de données au moyen d'un protocole réseau (ODBC, JDBC par exemple). L'identifiant $idSujet$ peut être le ROWID ou un identifiant technique ou encore un identifiant logique comme une clé primaire.

Définition 2. Un prédicat d'annotation p est constitué d'une composante désignant un concept ou une propriété de l'ontologie o.
D'un point de vue opérationnel, il s'agira par exemple d'une URI faisant référence à un terme d'un fichier OWL ou une référence identifiant un tuple dans une base de données.

Définition 3. Un objet d'annotation o est constitué soit :
– d'une seule composante ($Sujet$) définissant un sujet d'annotation ;
– de deux composantes ($Valeur, Type_o$) où $Valeur$ désigne la valeur de l'objet et $Type_o$ désigne le type de la $Valeur$ issu de l'ontologie o.

Grâce au triplet $\langle s, p, o \rangle$, six modes d'annotations sont alors possibles en fonction de la nature des sujets et des objets :

1. le sujet est une donnée commune à toutes les études et l'objet est une valeur. Par exemple, l'annotation A_1 indique que le patient P1 est fumeur (voir figure 3.6) :
   $A_1 = (PatientId\#P1, Fumeur, < oui, booléen >)$;
   Nous faisons l'hypothèse que l'attribut PatientId est la clé primaire de la table Patient.
2. le sujet est une annotation et l'objet est une valeur. Ce mode sert à compléter une annotation déjà réalisée. Par exemple, l'annotation suivante exprime que la quantité de cigarettes fumées par la patient P1 est de 4 :
   $A_2 = (A_1, quantité, < 4, entier >)$;
3. le sujet et l'objet sont des données communes à toutes les études. Par exemple lors de la réalisation d'une étude sur l'hérédité d'une maladie, il est possible de créer une annotation entre deux patients *via* un prédicat déterminant leur relation de filiation
   $A_3 = (PatientId\#P1, père, PatientId\#P5)$;

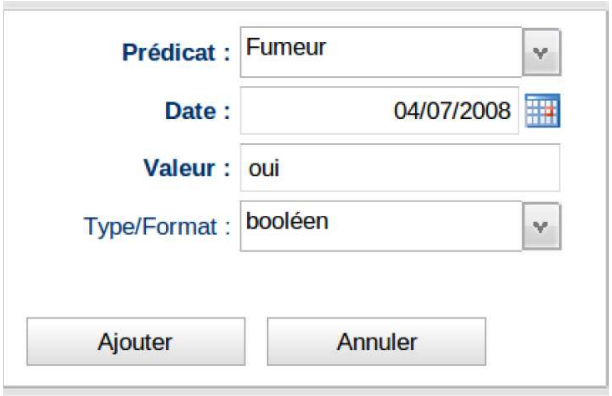


FIGURE 3.6 – Fenêtre de gestion des annotations dans *eClims*

4. le sujet est une donnée commune et l'objet est une annotation. Par exemple, les annotations suivantes expriment le fait que le patient P1 fait l'objet d'un prélèvement décrit par l'annotation A_4 :

   $A_4 = (Prélèvement Id\#Pr45, OrganeTouché, < foie, chaîne >)$ et

   $A_5 = (PatientId\#P1, Donne, A_4)$

5. le sujet est une annotation et l'objet est une donnée commune :

   $A_6 = (A_4, Provient, PatientId\#P1)$

6. le sujet et l'objet sont des annotations. Ce mode permet de relier deux annotations pour signifier l'existence d'une relation entre les deux annotations ou bien pour partager des annotations complexes. L'annotation suivante exprime le fait que les annotations A_5 et A_6 sont équivalentes : $A_7 = (A_5, Equivaut, A_6)$

Implémentation des annotations dans *eClims*

Nous avons mis en œuvre le modèle **E**ntity-**A**ttribute-**V**alue qui permet de séparer un objet et ses champs. Le modèle EAV correspond à une relation avec trois colonnes où la première colonne correspond à l'identifiant du sujet, la seconde à un prédicat et la troisième à la valeur prise par l'objet. Les annotations implémentées dans *eClims* sont seulement liées aux patients, prélèvements et échantillons, nous avons ajoutées trois tables dans la base de données de *eClims* pour stocker les triplets.

Caractéristiques du modèle d'annotation de *eClims*

Les tableaux suivants positionnent *eClims* par rapport aux critères d'Oren et aux objectifs de Virbel. Dans *eClims*, chaque relation stockant une donnée référentielle est associée à une relation d'annotation. Le mécanisme d'annotation permet de poser des annotations au niveau du tuple ou de la valeur de l'attribut en sélectionnant des termes dans une ontologie pour les prédicats. L'interrogation est basée sur l'utilisation des mécanismes présents dans le framework Hibernate, et notamment les requêtes basées sur les critères. La dimension usage est assurée par l'expressivité de notre modèle d'annotation.

Critères de Oren et al.	*eClims*
association	relations spécifiques : une par donnée référentielle
granularité	tuple & valeur d'attribut
terminologie	ontologie
type d'objet	littéral & terme d'ontologie & URI & URL
contexte	possible par la pose de méta-données sous la forme d'annotation
flexibilité	oui
sortie	utilisation des mécanismes du framework Hibernate (*Criteria Queries*)

TABLE 3.4 – Positionnement de *WikiBridge* par rapport aux critères de Oren et al.

Objectifs de Virbel	*eClims*
contextualiser	ontologie
typer des fragments	possibilité *o* null
corréler	annotation de type n^o 3 ou 6
analyser/interpréter	annotation
reformuler	ontologie vue comme un vocabulaire commun
hiérarchiser	deux types de données (référentielle et complémentaire)
commenter	oui
documenter	oui

TABLE 3.5 – Positionnement de *eClims* par rapport aux objectifs de Virbel

3.6 Conclusion & discussion

Les Systèmes d'Information Scientifique produisent chaque jour d'importants volumes de données très variables pour des équipes de recherche multi-disciplinaires. Il en résulte une variabilité des partenaires impliqués dans le Système d'Information Scientifique et une très grande variabilité des structures de données qui va au delà de la simple hétérogénéité. Les Bases de Données classiques (SGBDR) ne permettent pas de répondre aux exigences d'extensibilité des Bases de Données scientifiques au niveau de la structure des données (schéma) et des applications.
Nous proposons une approche qui apporte aux Systèmes d'Information Scientifique l'extensibilité de la structure et la prise en compte de la multi-disciplinarité au moyen d'un unique paradigme : l'annotation sémantique. L'annotation est une structure simple et quasi universelle qui permet de développer des composants génériques pour traiter l'extensibilité nécessaire. Deux plate-formes, dans deux domaines d'application, ont été développées pour tester notre proposition. Le chapitre 4 présente *WikiBridge* un wiki sémantique développé pour le domaine des Sciences Humaines et Sociales dans le cadre de l'élaboration d'un corpus en archéologie. Cette plate-forme met en avant les aspects sémantiques *via* un assistant d'annotation s'appuyant sur une ontologie d'application. L'outil utilisé est un wiki, les langages sont OWL, RDF, SPARQL, PHP. Le chapitre 5 présente *eClims* un composant de LIMS dans le domaine des Sciences du Vivant dans le cadre d'un projet pour la protéomique clinique. Cette plate-forme prend en charge les aspects extensibilité de la structure des données et qualité des données à travers un outil d'importation. Les outils mis en jeu sont un SGBDR, un LIMS, les langages Java et OWL sont employés.
Bien que la variabilité entre les partenaires et la variabilité des structures de données tendent à réduire la qualité, les fonctionnalités des Systèmes d'Information Scientifique doivent maintenir et contrôler cette qualité. Notre modèle d'annotation est basé sur une ontologie et des règles qui permettent de découpler la connaissance du domaine de la connaissance sur les données. Nous proposons une analogie entre la sémantique des langages de programmation et le processus

d'annotation afin de contrôler que les annotations posées sont conformes à la syntaxe abstraite et à la connaissance du domaine.

Notre proposition repose sur un stockage des données multi-paradigmes dont le dénominateur commun est la logique du premier ordre (FOL). Les données référentielles sont stockées dans un SGBDR dont la sémantique est donnée par un sous-ensemble de la logique du premier ordre. Les annotations sont des triplets RDF $\langle s, p, o \rangle$ qui sont aussi un sous-ensemble de la logique du premier ordre avec des prédicats binaires et des variables existentielles, $p(s, o)$. Les ontologies sont exprimées en OWL et leur sémantique est la Logique de Description. Les Logiques de Description constituent une famille de formalismes dotées d'une sémantique qui fait appel à des fragments décidables de la logique du premier ordre. La logique du premier ordre unifie donc les données référentielles et certains types d'annotations. Pour unifier contraintes et règles logiques, nous avons utilisé SWRL dans le projet *eClims*. En effet, SWRL est un langage de règles combinant OWL et le langage RuleML. RuleML est un langage de règles dont le noyau est basé sur Datalog [16], on peut grossièrement dire que SWRL est approximativement l'union de la logique de Horn et d'OWL. Datalog pourrait alors servir de base pour garantir la qualité des données stockées dans l'architecture SemLab, il offre un support de relation par les prédicats extensionnels, un support de graphe et peut traiter de grand volume de faits et de règles.
Les annotations contribuent à traiter la problématique de variabilité, une suite à nos travaux peut porter sur la stabilité des annotations. En effet, les annotations récurrentes peuvent être vues soit comme une nouvelle connaissance du domaine ou comme de nouvelles données référentielles et ainsi pouvoir être propagées soit dans l'ontologie soit dans le schéma des données. Seuls les experts du domaine peuvent prendre une telle décision.

Publications associées

[1] Éric Leclercq et Marinette Savonnet, Enhancing Scientific Information Systems with Semantic Annotations, *Symposium On Applied Computing (SAC)*, pp. 322-327, 2013.

16. Datalog connait un regain d'intérêt concrétisé par le workshop Datalog Reloaded en 2010. La relation sémantique entre Datalog et les Logiques de Description a d'ailleurs été étudiée [KRS10, Vol04, DLNS98].

Chapitre 4

La plate-forme *WikiBridge* : un exemple d'application pour des corpus dans le domaine des Sciences Humaines et Sociales

[Wiki] « *The simplest online database that could possibly work.* »
Ward Cunningham, 2001

« *A semantic wiki is a system that allows collaborative authoring, editing and linking of pages, but also the authoring and adding semantics to the data on the wiki itself.* »
Kousetti, 2008

Sommaire

L'objectif de ce chapitre est de présenter la plate-forme que nous avons développée pour l'ANR CARE (*Corpus Architecturae Religiosae Europeae - IV-X saec.* - ANR-07-Corp-011) qui travaille sur des corpus dans un environnement documentaire complexe. Les corpus sont constitués de façon collaborative : il s'agit de connaissances que l'on veut agréger, produire, partager, échanger et pouvoir faire évoluer. Les corpus doivent pouvoir être consultables sur Internet, être exploités par des utilisateurs experts du domaine mais aussi par des utilisateurs novices. La plate-forme doit avoir une maintenance réduite c'est-à-dire qu'elle ne doit pas nécessiter en permanence un informaticien pour maintenir l'application.

4.1 Un wiki sémantique comme modèle d'interaction pour la conception d'un corpus numérique

Nous sommes d'accord avec Evans lorsqu'il déclare dans [Eva04] que « *Si la conception du logiciel ... ne correspond pas au modèle du domaine, ce modèle n'a que très peu de valeur, et la correction du logiciel est suspecte* ». L'application proposée doit permettre de travailler avec un contenu informel comme des articles écrits en langage naturel ou des images, complété par des structures plus formelles, permettant de transiter graduellement d'une description informelle vers une description plus formelle de cette connaissance. Dans le même temps, l'essor du Web a entraîné la multiplication des pratiques collectives centrées sur la création de corpus numériques allant de la mise à disposition des documents à l'utilisation des annotations accompagnant la rédaction collective du corpus. Notre choix s'est porté vers une plate-forme Web offrant non seulement un environnement ouvert et permettant le partage, l'échange, la collaboration entre utilisateurs tout en supportant une évolution du système. Cette plate-forme doit fournir 1) une présentation électronique (dite donnée primaire) qui doit être le reflet exact de la version imprimée pour des fins de citation et 2) des outils d'annotation sémantique. L'annotation sémantique ajoute des niveaux supplémentaires d'interprétation grâce à la formulation et la vérification d'hypothèses, elle offre une meilleure qualité dans le processus de recherche qu'un moteur de recherche plein-texte, et les résultats peuvent être affichés selon les profils des utilisateurs. En outre, elle permet l'interopérabilité entre les corpus grâce aux annotations s'appuyant sur une ontologie commune. Dans la plate-forme Web, ces fonctionnalités doivent être fournies par la même interface que celle de saisie des données primaires afin d'éviter à l'utilisateur de passer d'une interface à l'autre. Parallèlement, ces annotations doivent être maintenues sur des couches nettement séparées afin d'assurer l'intégrité des données primaires.
Ce contexte centré sur des descriptions textuelles et demandant une interface Web avec une composante fortement collaborative, nous a amené à choisir un **wiki comme modèle d'interaction**.

4.1.1 Vers une convergence entre Web Sémantique et Web collaboratif

Le Web Sémantique a pour objectif de rendre interprétable le Web, non seulement par des humains mais aussi par des machines. Berners-Lee et al. [BLHL01] donnent la définition suivante : « *The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.* » Il s'agit bien d'une évolution reposant sur la définition d'annotations sémantiques modélisant de façon formelle les données présentes dans les pages Web. Ces annotations demandent d'une part un modèle RDF et d'autre part des ontologies pour définir de façon formelle mais aussi interprétable et interopérable la sémantique des données. RDF (*Resource Description Framework*) permet de représenter des ressources sur le Web de manière uniforme

par la notion de triplet (sujet, prédicat, objet). Les ontologies sont écrites avec OWL (*Web Ontology Language*) qui ajoute de nouveaux constructeurs et axiomes permettant d'accroître l'expressivité des ontologies avec une sémantique plus poussée que RDFS. Enfin, SPARQL (*SPARQL Protocol and RDF Query Language*) propose à la fois un langage et un protocole pour interroger des données modélisées en RDF.

Parallèlement à cette évolution du Web, est apparu ces dernières années un nouveau concept plus social que technique communément appelée Web 2.0. Le Web 2.0 est vu comme un Web collaboratif ou social où les internautes ont un rôle de « *consommacteurs* » et « *consommauteurs* ». Cette définition est complétée par une dimension technique avec les feuilles de style, la syndication de contenu, Ajax et par la primauté des données liées à l'application considérée. Les plate-formes Web 2.0 partagent un ensemble de principes : 1) l'utilisateur est au centre de la plate-forme en terme de publication ; 2) le passage du statut de consommateur à celui d'auteur doit se faire simplement. Les interfaces doivent être intuitives et ne pas nécessiter de pré-requis techniques ; 3) la composante sociale doit être présente. Les plates-formes doivent être en mesure de stimuler les synergies entre internautes.

De nombreux auteurs [AKTV08, Gan06, Gru08, HM08] pensent que ces deux voies du Web ne sont pas contradictoires mais bien au contraire qu'elles peuvent et doivent profiter des apports et des travaux de chaque communauté, pour converger vers un Web optimisé pour les humains et les machines. Pour corroborer ce point, nous pouvons citer l'interview [1] donnée par Berners-Lee lors de la 4ème conférence internationale Semantic Web : « *I think we could ... have both Semantic Web technology supporting online communities, but at the same time also online communities can also support Semantic Web data by being the sources of people voluntarily connecting things together.* ». Le Web Sémantique peut réutiliser les paradigmes suivants du Web collaboratif : 1) un grand nombre d'utilisateurs producteurs de données et une collaboration entre utilisateurs à des fins de création collective et consensuelle ; 2) des outils simples pour la production à grande échelle de données formalisées selon les principes du Web Sémantique et pour la visualisation des annotations. De la même façon, si les outils du Web 2.0 proposent des méthodes efficaces pour les utilisateurs, les outils du Web Sémantique offrent un apport essentiel pour la structuration et l'échange de données. C'est par cette complémentarité que pourront se former des espaces informationnels par dessus la frontière séparant les deux domaines (voir figure 4.1).

Nous proposons dans la suite du chapitre de décrire *WikiBridge* un serveur de wiki sémantique permettant de créer un tel éco-système informationnel.

4.1.2 Wiki sémantique ou *"Semantic Web in the small"*

La combinaison des Wikis et des technologies du Web Sémantique est considérée par de nombreux membres des deux communautés comme une approche pour créer collaborativement et accéder à des informations sur le Web. Le succès des Wikis est principalement dû à leur simplicité et à leur convivialité. Cependant, les wikis ont encore besoin de moyens pour organiser et interroger les contenus créés. Bien que l'importance de ce problème soit largement reconnue, les moteurs de Wikis actuels se contentent de limiter l'organisation des pages Wiki grâce à un ensemble de catégories définies et maintenues manuellement. Les limites de cette approche apparaissent lorsque les utilisateurs du Wiki recherchent des informations sur un sujet précis ou des renseignements répartis sur plusieurs pages. Le Web Sémantique fournit l'infrastructure technologique pour remédier à cette situation. RDF peut être utilisé pour améliorer la sémantique des pages Wiki et les liens entre elles, tandis que les ontologies et les services de raisonnement associés sont des extensions efficaces pour leurs capacités de recherche d'information.

1. http://esw.w3.org/topic/IswcPodcast

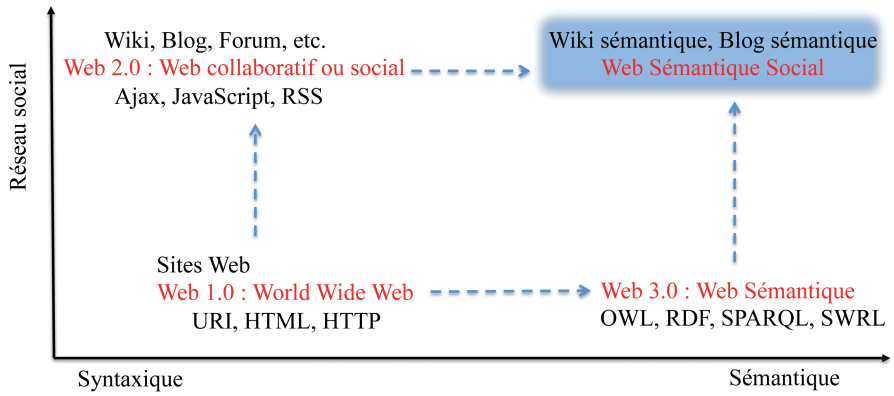


Figure 4.1 – Convergence entre le Web Sémantique et le Web collaboratif (d'après [BBP+08])

On peut distinguer deux méthodes dans la conception de serveurs de wikis sémantiques. La première dite *wikis for ontologies* considère les pages du wiki comme des concepts et les liens typés comme des propriétés. Elle concerne le plus grand nombre de wikis. La seconde dite *ontologies for wikis* utilise une ontologie pré-existante importée dans le wiki pour la mise en place d'annotations. Cette méthode fournit généralement des listes de sélection ou utilise l'auto-complétion pour baser les annotations sur l'ontologie. Ces moteurs de Wiki sont le plus souvent destinés à des domaines spécifiques comme en atteste l'état de l'art des différents moteurs de wikis sémantiques réalisé par Meilender et al. [MJLP11].

4.2 Aperçu de *WikiBridge*

Nous avons développé une plate-forme numérique, *WikiBridge* (http://care.tge-adonis.fr), s'appuyant sur les technologies du Web 2.0 et du Web Sémantique. D'un point de vue technique, la plate-forme est déployée pour sa partie utilisateur sous la forme d'un wiki sémantique de type *ontologies for wikis* et pour sa partie infrastructure de gestion de données et de connaissances sous la forme de base de données annotées couplée avec un triple-store. Nous avons validé nos concepts dans le domaine de l'archéologie avec le projet CARE (*Corpus Architecturae Religiosae Europeae - IV-X saec.*) accepté par l'ANR pour 2008-2011 (ANR-07-Corp-011) [CS12]. CARE a pour objectif de recenser les édifices religieux de France et leur évolution entre le IV^{e} et le début du XI^{e} siècle. Certaines caractéristiques de ce domaine viennent complexifier le problème : 1) la complexité des données (hétérogènes, incomplètes, incertaines, inconsistantes, spatio-temporelles) ; 2) la barrière de la connaissance du domaine nécessaire pour utiliser la plate-forme ; 3) l'évolution de la connaissance ; 4) les compétences des utilisateurs. L'objectif était d'offrir une plate-forme qui intègre un continuum des informations par stockage dans un même endroit d'informations aussi variées qu'une description d'objets, une bibliographie, des textes, des photographies qui sont généralement stockés dans une base de données, un système documentaire, un système de fichiers, etc.

4.2.1 Architecture de *WikiBridge*

La figure 4.2 présente les composants de *WikiBridge* et leurs interactions. Le système de gestion des articles incluant la saisie ainsi que la mise en forme est implanté en utilisant MediaWiki que nous avons étendu avec les composants sémantiques suivants :

– une interface d'acquisition de données basée sur des formulaires sémantiques permettant une annotation en mode automatique et en mode manuel ;
– un assistant pour la construction des annotations ;
– un moteur de requêtes SPARQL ;
– un mécanisme de contrôle des annotations incluant la prise en compte du contexte et des contraintes sémantiques.

Le support des annotations de *WikiBridge* permet d'annoter n'importe quel élément au moyen d'une ontologie construite collaborativement avec les experts du domaine. L'ontologie OWL est importée dans *WikiBridge* et stockée dans une Base de Données.

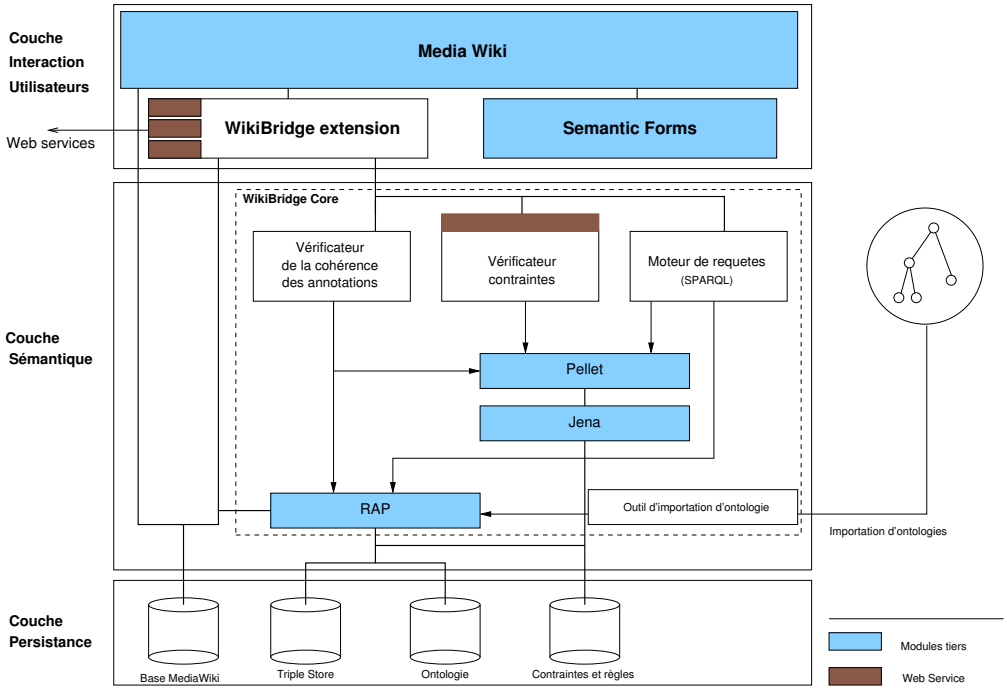


FIGURE 4.2 – Architecture de *WikiBridge*

Nous détaillons dans le paragraphe suivant les exigences qui ont servi au développement de *WikiBridge*.

4.2.2 Exigences auxquelles répond *WikiBridge*

La conception de *WikiBridge* a répondu aux quatre exigences suivantes : gestion de documents composites, interface utilisateur intuitive, conception collaborative du contenu, support de différentes catégories d'utilisateurs. Ces exigences ont été complétées par des exigences techniques identifiées par Uren et al. [UCI+06] : stockage de données de format hétérogène, respect des standards du Web Sémantique, stockage des annotations.
Nous avons déterminé deux axes :
– celui des aspects documentaires et coopératifs pouvant être couverts par un wiki ;
– celui des aspects sémantiques pouvant être couverts par la modélisation de la connaissance du domaine exploitée sous la forme d'annotations associées à des documents.

WikiBridge propose une **interface simple et riche** utilisant les technologies AJAX et un moteur WYSIWIG qui supporte la syntaxe du langage de mise en forme de MediaWiki.

WikiBridge supporte des **formats de données hétérogènes**. Plusieurs types de documents peuvent être ajoutés à un article. Après avoir été téléchargés sur la plate-forme, les documents

peuvent être visualisés dans l'article sous fome de vignettes. Les liens avec les documents annexes peuvent également être annotés.

La **conception collaborative du contenu** par les utilisateurs est possible au moyen d'une interface unique. En effet, l'environnement intégré du wiki permet aux utilisateurs d'annoter les données mais aussi de les créer, de les mettre en forme et de les partager.

Pour prendre en compte différents degrés d'utilisation de la plate-forme, c'est-à-dire à la fois des utilisateurs novices, des utilisateurs expérimentés et des chercheurs, certaines fonctionnalités doivent être désactivées. Par conséquent, un **mécanisme d'ACL** (*Access Control List*) est nécessaire pour définir les fonctionnalités accessibles par utilisateur et par groupe.

Afin d'être en mesure d'échanger des données avec d'autres applications (éditeurs d'ontologies, Services Web, autres wikis) un respect des standards du Web Sémantique est impératif. *WikiBridge* utilise exclusivement des technologies standards comme **OWL pour la description d'ontologies et RDF pour les annotations**.

Les **annotations sont stockées, séparément du document original, dans une base de données sous forme de triplets**.

Un **moteur de requête SPARQL** permet d'interroger les annotations et l'ontologie. De plus, les requêtes SPARQL peuvent être incluses dans des pages du wiki (*in-line queries*).

Nous considérons que les **capacités de raisonnement** sont une des fonctionnalités les plus importantes pour les plate-formes supportant des processus d'intelligence collective. Elles doivent permettre : 1) de faire émerger de nouvelles connaissances à partir de celles déjà acquises et des données du wiki ; 2) de contrôler la signification des annotations par rapport à leur contexte d'utilisation ; 3) d'améliorer la navigation et les recherches.

4.3 Couche d'interaction avec les utilisateurs

Les documents présents dans *WikiBridge* sont définis selon trois couches (voir figure 4.3) : leur contenu (agrégation de ressources textuelles et multimédia), leur structure, la connaissance qui leur est associée. La première couche est mise en place par le wiki. Les autres couches sont, dans la plupart des wikis, des fonctionnalités supplémentaires. La notion de *wiki template* permet de définir la structure du document. Les connaissances implicites exprimées dans les articles doivent être rendues explicites. Nous utilisons une ontologie et un mécanisme d'annotations pour représenter les connaissances associées aux articles. L'ensemble constitue une Base de Données annotées.

Couche Données

La couche d'interaction avec les utilisateurs est majoritairement couverte par MediaWiki en offrant un support aux couches Données et Structure. MediaWiki propose une syntaxe et un moteur WYSIWIG (voir élément (a) de la figure 4.4) pour gérer la mise en page. Le wiki permet *un premier enrichissement* du document papier grâce :

1. aux liens soit internes entre des articles du wiki soit externes vers d'autres ressources Web. Les liens externes permettent de compléter des parties du texte, par exemple on peut donner l'URL du site internet d'un musée dans lequel se trouve aujourd'hui un objet faisant partie de l'édifice décrit dans l'article. Les liens externes peuvent aussi offrir une aide lors de la saisie d'un article, par exemple le site de l'INSEE donnant le code et le libellé exacts des communes mais aussi la possibilité de trouver la latitude et la longitude en donnant une adresse (voir les éléments (b) et (c) de la figure 4.4) ;

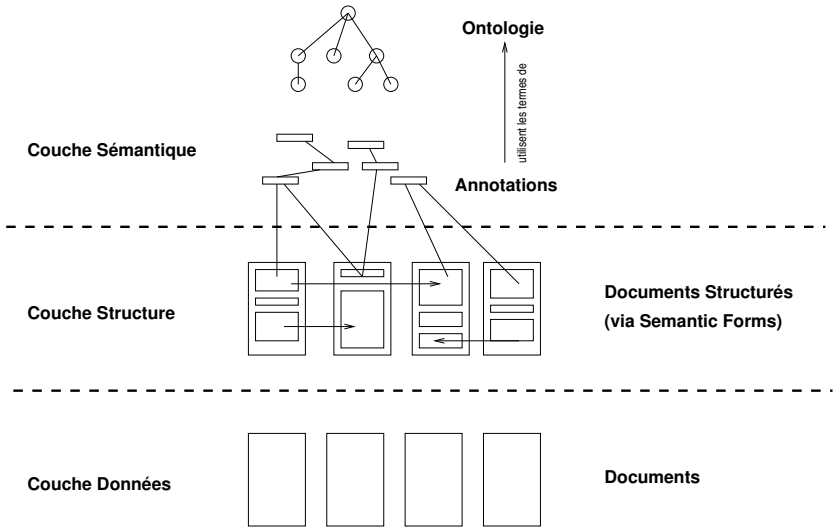


FIGURE 4.3 – Structuration des documents dans *WikiBridge*

2. aux supports multimédia comme des photographies, des plans, du son ou de la vidéo et
3. à des Services Web. Dans le cadre du projet CARE, nous avons utilisé des Services Web pour la géo-localisation comme Google Maps de MediaWiki [2]ou l'API Géoportail de l'IGN [3] (voir figure 4.5).

Couche Structure

Le concept de *wiki template* permet aux utilisateurs de déterminer la structure et l'apparence d'une page wiki [HLS05, IVZ08, IZ06]. Deux types de template ont été identifiés : 1) les templates fonctionnels correspondent à une page qui contient des espaces réservés complétés par des valeurs passées en paramètre, l'exemple type est l'infobox ; 2) les templates créationnels sont des pages utilisées comme point de départ pour la création, à la volée, de nouvelles pages sur la base de la structure du template. Un wiki utilisant des templates créationnels est appelé *"Lightly Constrained Wiki"*.
Nous utilisons à la fois une infobox pour synthétiser les données et des templates créationnels pour structurer les données. Les templates créationnels sont pris en charge par l'extension Semantic Forms [4] développée pour MediaWiki. Ils sont décrits en utilisant un langage spécifique, chaque partie du document est représentée par un modèle (voir figure 4.6.a). Les modèles peuvent être composés. L'annexe D présente le modèle correspondant à une fiche de dépouillement d'un édifice du projet CARE donnée en annexe C. La structure de la fiche nous a été fournie par les archéologues, elle respecte les normes retenues lors des diverses rencontres européennes [BJ12]. La figure 4.6.b montre le formulaire de saisie ainsi créé. Les champs peuvent prendre la forme de boîtes de sélection, de cases à cocher ou de texte libre, l'auto-complétion est disponible. De plus, Semantic Forms permet de remplir des champs en sélectionnant des valeurs dans des listes. Nous avons modifié ce mécanisme afin de construire, au moyen de requêtes SPARQL, les listes de valeurs à partir des termes de l'ontologie développée.
Les formulaires permettent à des utilisateurs non-experts du domaine de remplir grâce à un copier-coller les champs à partir des documents papier. Un aperçu de l'interface de saisie est

2. http://www.mediawiki.org/wiki/Extension:Google_Maps
3. http://depot.ign.fr/geoportail/api/doc/fr/developpeur/download.html
4. http://www.mediawiki.org/wiki/Extension:Semantic_Forms

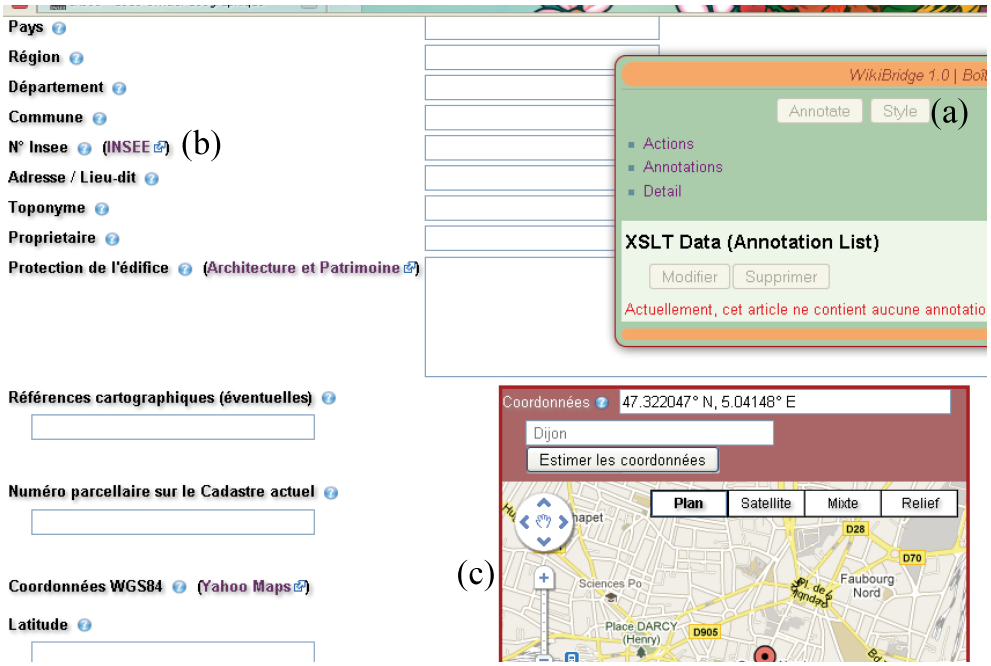


FIGURE 4.4 – Formulaire de saisie & aides possibles dans *WikiBridge*

donné en figure 4.7. La figure 4.8 donne un aperçu du rendu de la fiche : la partie gauche montre le template créationnel et l'infobox qui apparaît sous la forme d'un tableau, à droite un plan, une galerie de photographies. Les modules correspondant à la couche d'interaction utilisateur sont représentés en haut de la figure 4.2.

Ces deux couches sont l'expression dans une plate-forme numérique des méthodes de travail des archéologues et proposent de nombreuses fonctionnalités. Mais elles ne permettent un accès à l'information que par un moteur de recherche plein-texte. En effet, la structure de la fiche, qui est la seule entité modélisée par le template, ne correspond qu'à une modélisation "gros-grain" des concepts fondamentaux du corpus décrit dans le modèle conceptuel UML (voir figure 2.8 du chapitre 2). Nous aboutissons à la proposition d'une couche sémantique qui s'appuie sur 1) la structuration des concepts fondamentaux du corpus implémentés grâce au formulaire de saisie ; 2) la structuration de connaissances complémentaires implémentées par les annotations. Ces deux structurations s'appuient sur une ontologie spécialisée pour l'application. Cette couche sémantique nous offre la possibilité d'avoir un moteur de requêtes plus élaboré comparable au langage SQL.

4.4 Couche sémantique

Afin de contrôler la qualité des annotations, durant le processus d'annotation, nous proposons un ensemble de modules (boîtes blanches dans la figure 4.2) développé en utilisant des composants tiers : RAP, RDF API for PHP [5], Pellet [6] et Jena [7]. Ces composants mettent en œuvre la couche sémantique constituée d'une ontologie et d'annotations.

5. RDF API for PHP http://www4.wiwiss.fu-berlin.de/bizer/rdfapi/
6. http://pellet.owldl.com/
7. http://jena.sourceforge.net/

(a) Extension Google Maps (b) Extension Géoportail

FIGURE 4.5 – Services Web de géo-localisation dans *WikiBridge*

4.4.1 Ontologie d'application pour le corpus CARE

La haute technicité des descriptions métier fournies (nom et organisation des parties d'un édifice, techniques de construction, éléments de décoration, etc.) par les archéologues a rendu possible la réalisation de l'ontologie. Nous avons travaillé sur une représentation formelle des connaissances des éléments d'architecture religieuse en

1. classifiant les éléments architecturaux d'un édifice ;
2. définissant les relations entre les éléments : relation partie-tout, relations spatiales, relations de composition, etc. ;
3. établissant une correspondance entre les éléments d'architecture et le domaine religieux ;
4. enfin en tenant compte de la dimension temporelle (voir figure 4.9).

Par cette démarche nous avons dégagé un ensemble de concepts connectés à des termes architecturaux organisés par des relations méronomiques (décomposition morphologique) et représentationnelles (dimensions, matériaux, relations spatiales). Les concepts ont ensuite été rapprochés de l'ontologie de domaine CIDOC-CRM [8] [CC02] qui apporte une modélisation temporelle sous la forme d'états. L'ontologie développée a pris en compte l'ensemble des données archéologiques : des artefacts comme des bâtiments, des éléments architecturaux, des données intangibles telles que des mesures, des directions, des associations entre éléments et des données d'interprétation. Elle comporte 124 classes et 715 individus (en janvier 2012). Une description détaillée de l'ontologie CARE est donnée en section 5 de l'annexe A. L'éditeur Protégé a été utilisé pour produire cette ontologie en OWL, elle est ensuite importée dans *WikiBridge* (voir figure 4.10).

Prise en compte de l'évolution de l'ontologie

L'ontologie développée est amenée à évoluer sur deux axes : 1) l'ajout de concepts plus spécialisés ; 2) la prise en compte de domaines connexes à l'archéologie traditionnelle portant notamment sur l'étude de la caractérisation des matériaux, des techniques et débouchant sur le développement de nouvelles branches de l'ontologie. Existent de plus des évolutions mixtes induites par l'internationalisation du projet. La participation de la Grèce au projet CARE demandera la prise en compte des caractéristiques spécifiques à la religion orthodoxe qui seront développées par la création d'une nouvelle branche. La participation de l'Italie requiert une

8. http://www.cidoc-crm.org

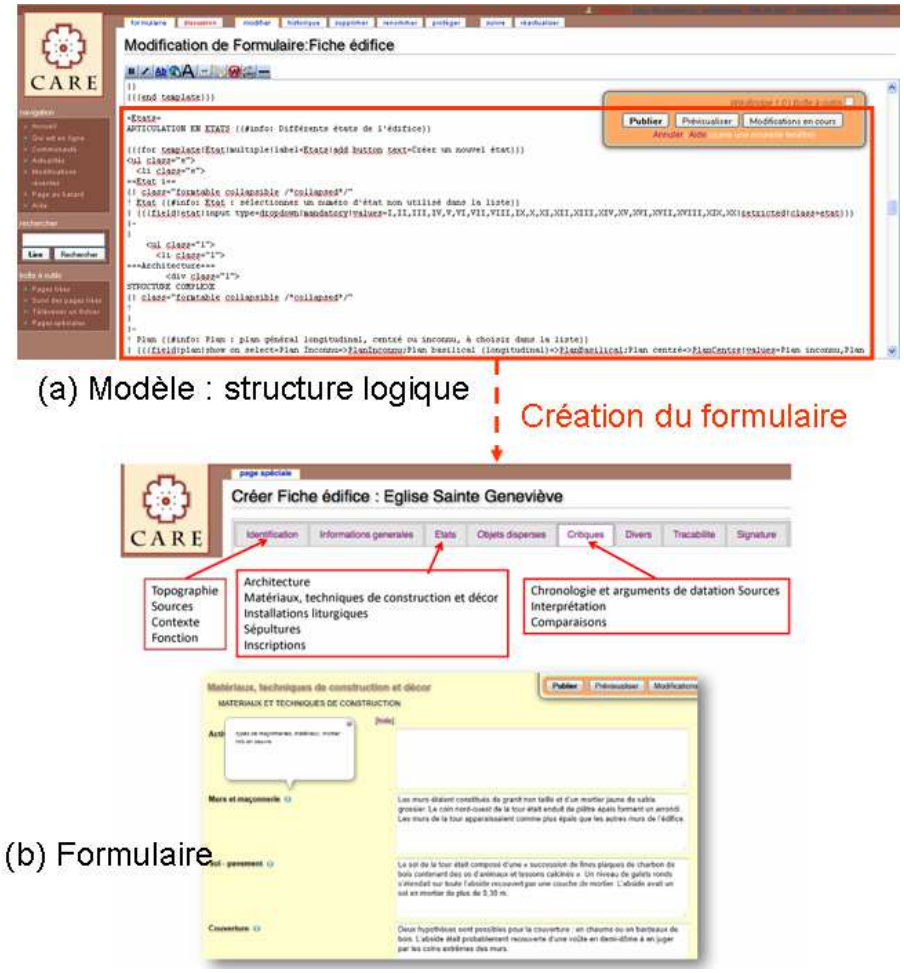


FIGURE 4.6 – Processus de création d'un formulaire dans *WikiBridge*

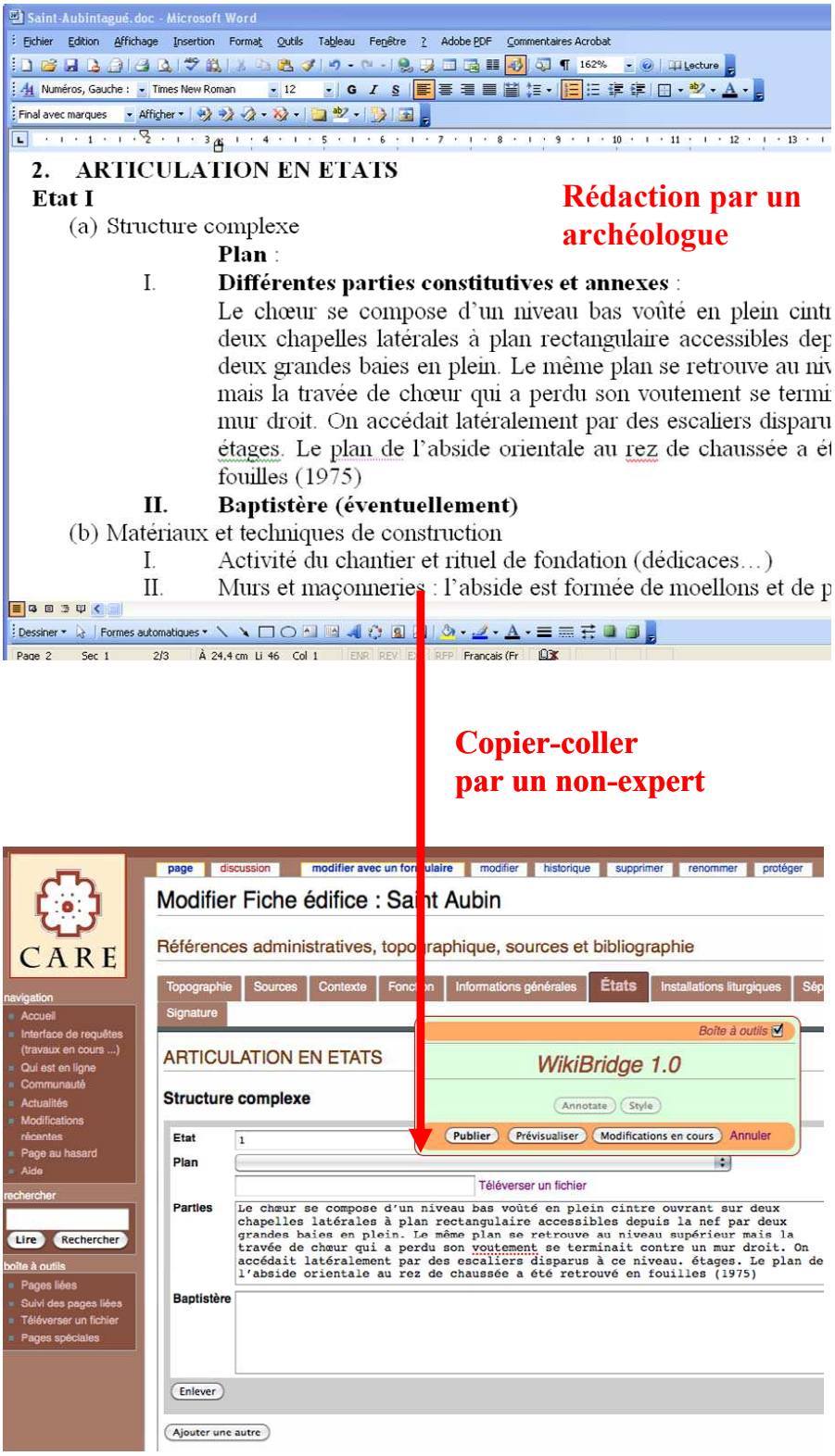


FIGURE 4.7 – Processus de saisie d'une fiche dans *WikiBridge*

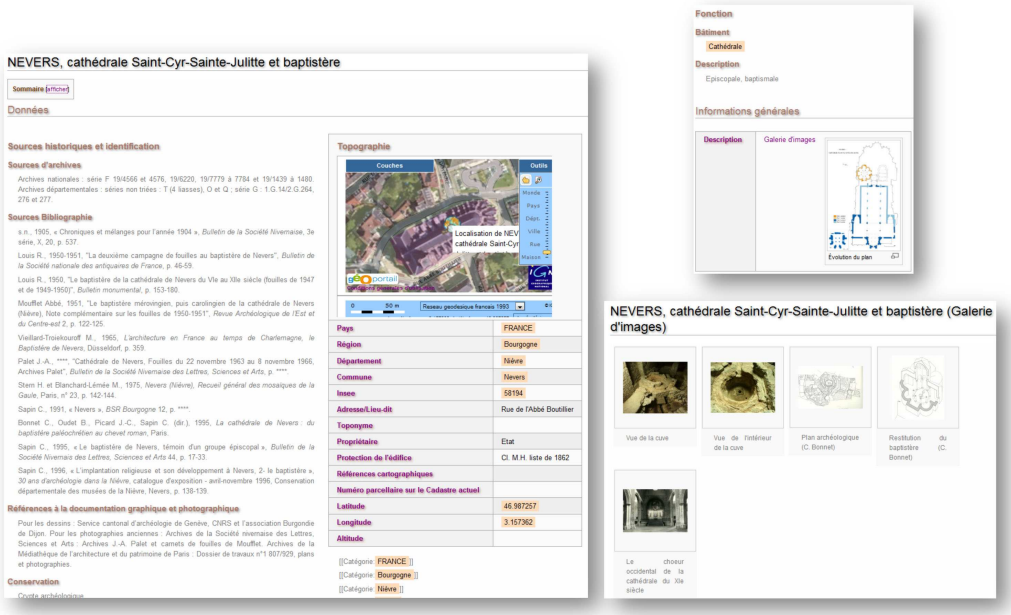


Figure 4.8 – Rendu d'une fiche du projet CARE

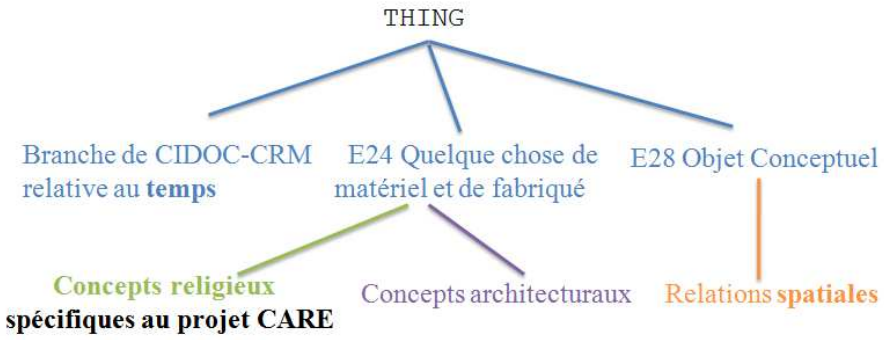


Figure 4.9 – Grandes branches de l'ontologie CARE

prise en compte du baptistère comme un édifice à part entière alors qu'en France ce dernier n'est qu'une partie d'un édifice.
Les conclusions du colloque "Evolution d'ontologie" de 2010 (http://www.irit.fr/evolution2010) ont souligné que l'évolution d'ontologie se heurte aux difficultés suivantes : 1) cibler les connaissances à modifier ; 2) modifier la structure de l'ontologie ; 3) évaluer les impacts d'une modification sur la structure de l'ontologie et sur les applications. Dans le projet CARE, les deux premières difficultés n'apparaissent pas car il s'agit essentiellement de spécialisation de concepts ou d'ajout de nouvelles branches à l'ontologie. Le principal problème porte sur l'impact de l'évolution de l'ontologie sur les annotations existantes. Pour résoudre ce problème, les annotations sont estampillées avec le numéro de version de l'ontologie qui a servi à les créer. Ainsi, l'utilisateur est en mesure de détecter les annotations incohérentes avec l'évolution réalisée. Si cet ensemble est vide alors l'estampille de chaque annotation est remplacée par celle de la nouvelle version de l'ontologie. En revanche, les annotations incohérentes sont conservées avec leur estampille d'origine pour pouvoir être interrogées avec l'ontologie correspondante.

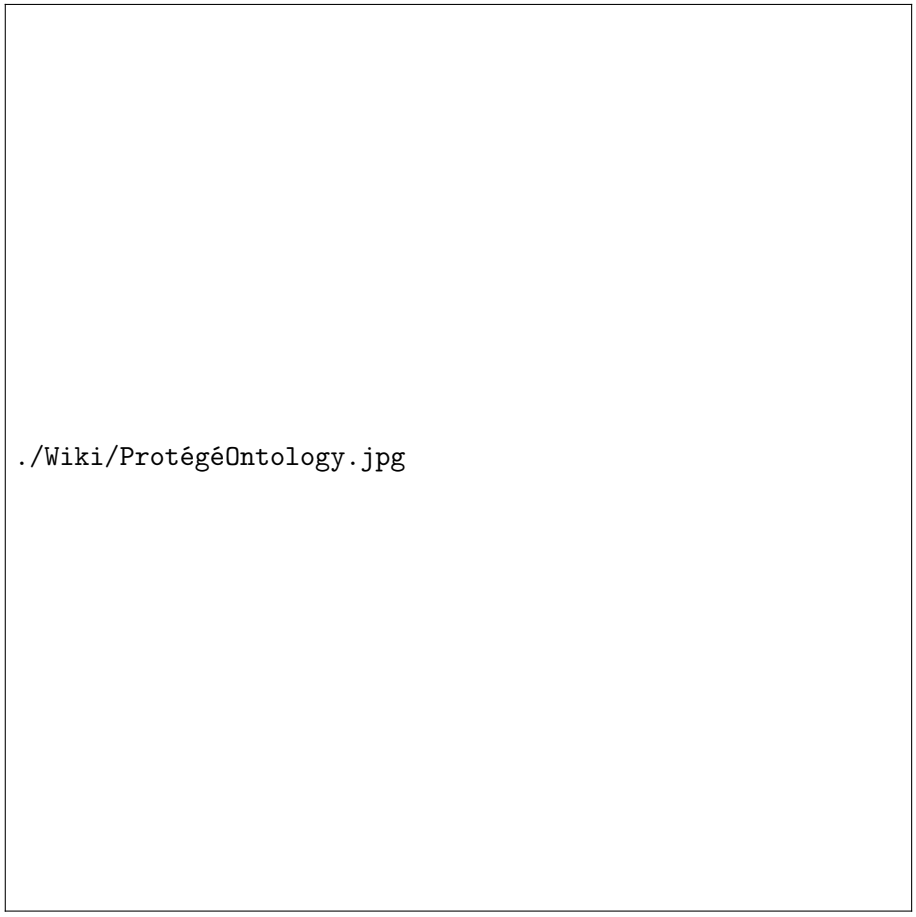


FIGURE 4.10 – Extrait de l'ontologie CARE dans l'éditeur Protégé

4.4.2 Annotations

L'annotation de ressources documentaires se pratique depuis longtemps dans le monde de la documentation. La Digital Library Federation (DLF) [9] a défini trois types d'annotations qui peuvent s'appliquer aux corpus :

1. l'annotation administrative indique les informations liées à la création et à la modification du document. Dublin Core est le standard pour ce type d'annotations avec des descripteurs comme l'auteur, le titre, la date de publication, la langue, etc. ;
2. l'annotation structurelle relie des parties de ressources documentaires entre elles afin de constituer une représentation logique d'un document ;
3. l'annotation descriptive décrit une ressource documentaire par rapport à son contenu en mettant en avant les concepts, les relations entre les concepts et leurs instances mentionnés dans la ressource. Dans le Web Sémantique l'annotation descriptive est connue sous le terme d'annotation sémantique.

L'annotation administrative est réalisée nativement par le wiki qui conserve par qui et quand une modification a été faite sur une fiche. L'annotation structurelle est réalisée par les templates créationnels, des annotations automatiques (simples) ont été mises en place si les données sont entrées *via* un formulaire.

9. DLF est une association composée des quinze plus grandes bibliothèques américaines http://www.diglib.org/dlfhomepage.thm

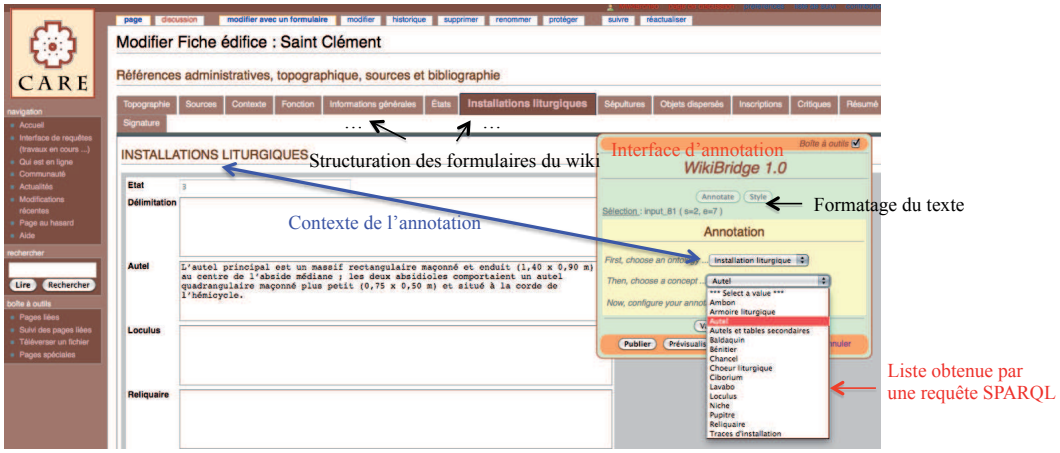


Figure 4.11 – Assistant d'annotation dans *WikiBridge*

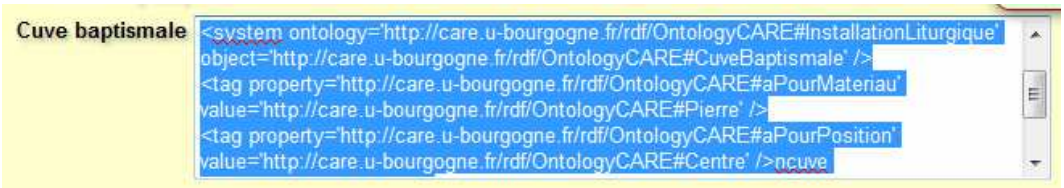


Figure 4.12 – Code source d'une annotation dans *WikiBridge*

L'assistant d'annotation permet aux utilisateurs de construire des annotations sémantiques. Le mécanisme d'annotation permet d'annoter n'importe quel élément (portion de texte, image, lien, etc.) en sélectionnant les termes de l'ontologie dans des listes et en leur associant des propriétés et des valeurs. Le processus d'annotation étant sensible au contexte, les termes sont sélectionnés dans l'ontologie par rapport aux champs actifs du formulaire. L'ontologie contraint les annotations sur les prédicats et objets utilisés. Les annotations ainsi construites sont vérifiées par rapport à l'ontologie lors de l'enregistrement ou de la modification des articles (voir figure 4.11). L'annotation générée est visible dans le texte source de l'article. Elle est modifiable par un expert. La figure 4.12 montre que, pour chaque article identifié par une URL, les annotations utilisent cette URL comme préfixe pour identifier les parties des articles ciblées (paragraphe, phrase, mot, etc.). 1 200 annotations ont été posées pour 150 fiches saisies (en janvier 2012).

Le processus de vérification de la cohérence des annotations comporte plusieurs composants spécifiques interagissant avec RAP, Jena et Pellet Les contraintes sémantiques exprimées en logique du premier ordre sont vérifiées en utilisant Pellet et un Service Web interconnectant RAP et Jena. De plus, des règles peuvent être ajoutées pour interroger l'ontologie et les annotations afin de tester de nouveaux faits et ainsi faire émerger de nouvelles connaissances. Afin d'implémenter les contraintes, deux solutions ont été testées : 1) transformation des contraintes dans un langage de programmation telle que PHP ; 2) utilisation d'un raisonneur et d'un ensemble de contraintes stockées dans un fichier. La seconde solution a été retenue car elle permet de définir et d'ajouter dynamiquement de nouvelles contraintes lorsque la connaissance évolue. Trois approches sont décrites dans [SSW08] : 1) sémantique basée sur la skolemisation, certaines contraintes sont taguées comme étant des contraintes d'intégrité ; 2) ruled-based semantics basé sur l'interaction avec une programmation logique qui fournit *negation as failure* sous l'hypothèse du monde clos and 3) query-based semantics qui lie des requêtes pour exprimer les contraintes.

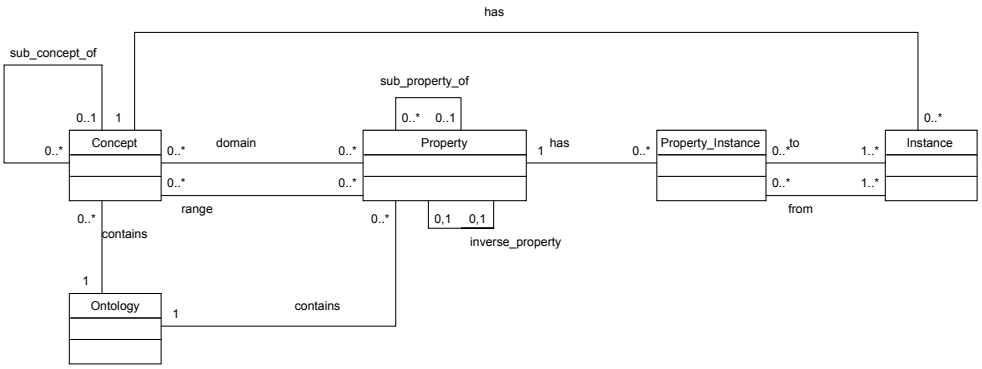


FIGURE 4.13 – Schéma de Base de Données pour l'ontologie

4.5 Couche de persistance

La couche de persistance inclut quatre types de stockage : le contenu des articles, l'ontologie, les annotations sémantiques, les contraintes et les règles.
Le stockage des articles est assuré par la base de données spécifique à MediaWiki.
La base de données annotées que nous avons mis en place dans *WikiBridge* complète la base des documents héritée de MediaWiki avec une structure pour l'ontologie et une structure pour les annotations. Les annotations sont stockées dans le triple store de RAP. Elles peuvent être interrogées au moyen de requêtes SPARQL incluses dans des articles (*in-line query*). Dans un premier temps, l'ontologie a été stockée dans un schéma spécifique donné en figure 4.13. Ceci rejoint les travaux de OntoDB [JHX+07] ou Sesame [BKvH02] qui proposent de stocker dans la même base de données, les données et les ontologies. Finalement, l'ontologie, importée à partir d'un fichier OWL, est stockée dans un schéma spécifique géré par RAP. L'ontologie peut également être interrogée en SPARQL et les résultats inclus dans des pages du wiki. Cette technique est utilisée pour construire les listes de termes, de propriétés et de valeurs proposés par l'assistant d'annotation.
Les contraintes exprimées sur les termes de l'ontologie et utilisées par Jena et Pellet sont stockées dans le format natif des outils, c'est-à-dire sous forme textuelle. Un marqueur permet de dissocier les règles appliquées à l'ontologie des contraintes utilisées pour vérifier la cohérence des annotations.

4.6 Couche d'accès à l'information

Pour extraire des informations quantitatives, effectuer des comparaisons, des vérifications ou des analyses spatiales, nous avons besoin d'un véritable langage de requête comparable à SQL. Nous utilisons les annotations pour réaliser la recherche d'information comme le propose Kiryakov et al. [KPT+04] : « *Semantic annotation ... aiming to enable new information access methods and to extend the existing ones.* ».

De plus, nous avons identifié différents types d'utilisateurs selon : 1) l'usage qu'ils veulent faire du corpus c'est-à-dire un simple lecteur, un investigateur ou une personne réalisant les annotations ; 2) le degré de connaissance du domaine qu'ils ont comme par exemple un néophyte, un amateur, un chercheur, etc. L'interface de requêtes doit donc permettre de répondre à la fois à des utilisateurs ayant une visée encyclopédique et à des utilisateurs ayant une visée d'expertise extrêmement ciblée.
Pour répondre aux attentes de ces différents types d'utilisateurs, nous offrons trois types de

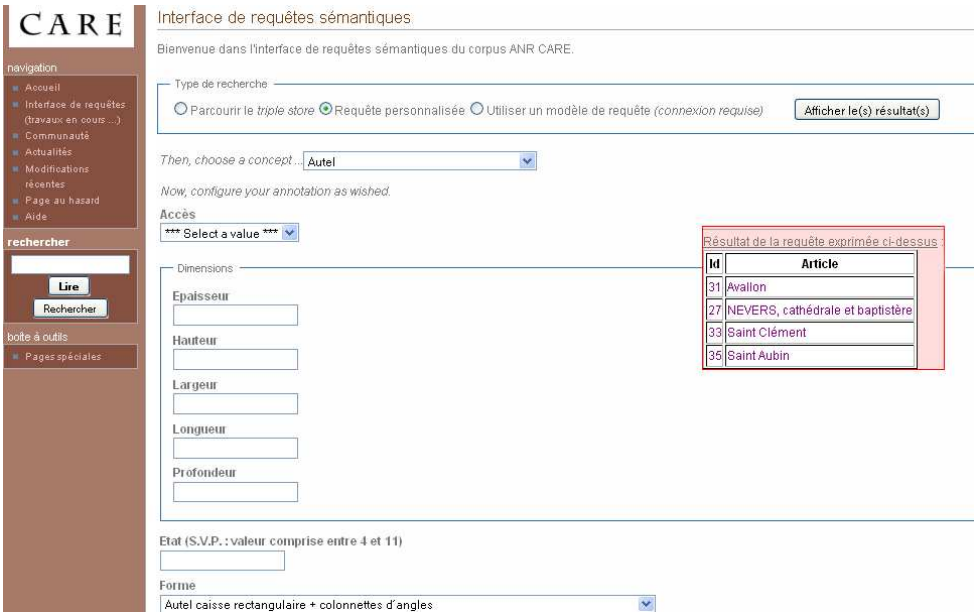


Figure 4.14 – Interface de requêtes dans *WikiBridge*

requêtes qui extraient des informations sans une connaissance *a priori* du schéma représentant les données :

1. la navigation par facette (*faceted browsing*) partitionne l'espace d'information en utilisant des caractéristiques importantes des éléments d'information que l'on appelle facettes. Chaque facette a de multiples valeurs et l'utilisateur sélectionne une valeur pour obtenir les éléments pertinents dans l'espace d'information. La théorie des facettes peut être directement associée à la navigation dans les données RDF : les éléments d'information sont des sujets RDF, les facettes sont des prédicats RDF et la restriction des valeurs sont des objets RDF. Dans *WikiBridge*, elle permet aux utilisateurs d'explorer, par filtration successive de la structure de l'ontologie, l'information disponible ;
2. la recherche par formulaire permet de saisir des paramètres pour les requêtes portant sur des conjonctions et identifiées lors de la phase d'analyse (voir figure 4.14) ;
3. une vue agrégée affiche dans une factbox, pour chaque article, toutes les annotations qui lui sont liées.

Trois types de résultats peuvent être affichés :

1. les résultats peuvent apparaître sous la forme d'une liste contenant des liens vers les articles, à l'endroit même où se situe l'annotation ;
2. un utilisateur peut manuellement naviguer *via* les liens entre les articles ;
3. les utilisateurs peuvent sélectionner une annotation et obtenir la liste des articles associés à la même annotation. Ce type de résultat est une combinaison des listes de résultats et des factbox. Le principe consiste ensuite à lire en parallèle les articles possédant des annotations communes.

Il est à remarquer, qu'une fois sélectionnées à l'aide d'annotations, des parties de documents peuvent être réutilisées pour construire un nouveau document.

Il est toujours possible pour un expert d'écrire des requêtes SPARQL en ligne comme le montre le texte suivant :

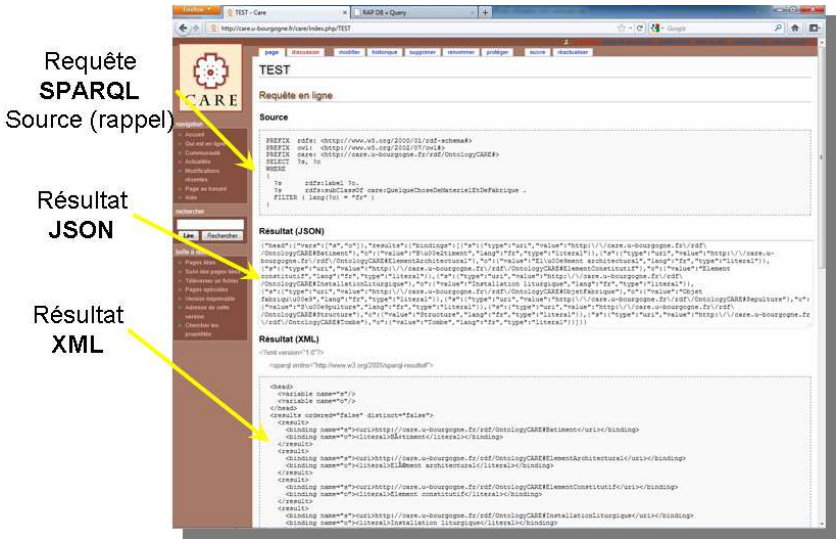


FIGURE 4.15 – Résultat d'une requête SPARQL dans *WikiBridge*

```
<smwt_ilsq type="db" model="http://care.u-bourgogne.fr/rdf/CAREO">
prefix ns: http://www.w3.org/1999/02/22-rdf-syntax-ns#
prefix careo http://care.u-bourgogne.fr/rdf/CAREO#
select ?s, ?longitude, ?latitude, ?pays, ?region, ?departement, ?batiment
where
{
   ?s       <aPourLongitude>    ?longitude ;
            <aPourLatitude>     ?latitude ;
            <aPourPays>         ?pays ;
            <aPourRegion>       ?region ;
            <aPourDepartement>  ?departement ;
            <aPourBatiment>     ?batiment ;
            <aPourRegion>       'Bourgogne' .
}
</smwt_ilsq>
```

La figure 4.15 montre le résultat d'une requête SPARQL dans différents formats (JSON, XML et HTML).

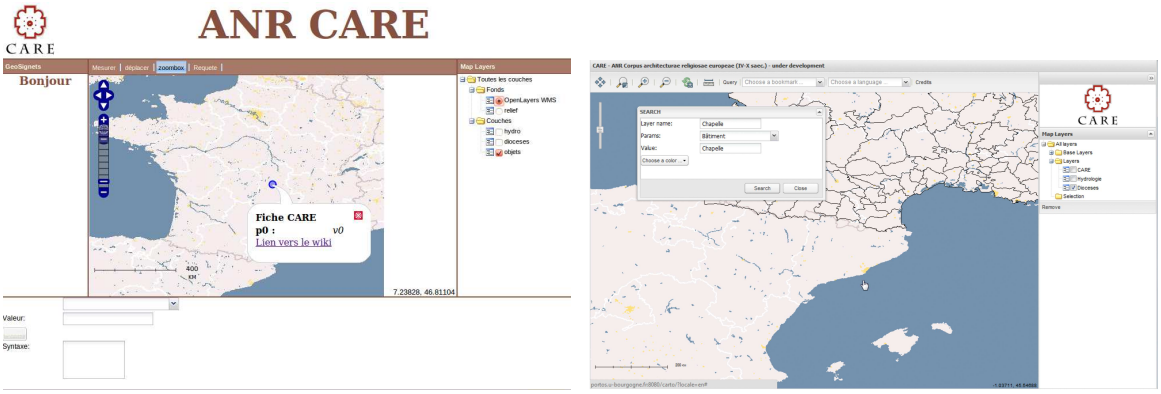
4.7 Services Web

Afin de permettre l'analyse spatio-temporelle des annotations relatives aux édifices (voir figure 4.16), un ensemble de Services Web a été développé en PHP. Un service spécifique permet d'établir la liste des coordonnées des édifices relatifs à une conjonction de propriétés. Un service générique renvoie les édifices et leurs propriétés par une requête SPARQL. Ces résultats sont exploités par l'équipe de géomatique de la Maison des Sciences de l'Homme de Dijon (http://msh-dijon.u-bourgogne.fr/pole-geomatique-cartographie.html) afin de créer une application de cartographie en ligne [Gra12]. Son développement, débuté en janvier 2011, comprend les éléments suivants :

– définition des types de zonage par entités géographiques, économiques, politiques, administratives, culturelles, etc. ;
– définition des niveaux d'entrée thématiques (général, spécifique, etc.) ;

– définition des possibilités de combinaisons entre zonage et niveaux d'entrée ;
– un module d'impression et la diffusion de fonds de carte (relief, hydrologie, image satellite, etc.).

Deux captures d'écran sont présentées en figure 4.16.



(a) Informations disponibles sur un édifice

(b) Formulaire de recherche

Figure 4.16 – Deux captures d'écran de l'application de cartographie CARE

Les choix technologiques sont basés sur des solutions Open Source : le serveur cartographique MapServer a été utilisé, l'interface graphique a utilisé les bibliothèques de fonctions OpenLayers [10], ExtJS [11] et GeoExt [12].

4.8 Pages de discussion

Les pages de discussion, proposées nativement par le wiki, permettent aux chercheurs de faire des interprétations et de confronter leurs idées. Cela est indispensable dans un projet où la multiplicité de points de vue sur un même article provoque le débat et devient la garantie d'un travail exhaustif.

4.9 Droits et permissions des utilisateurs

La gestion des articles a été conçue de manière à supporter différentes catégories d'utilisateurs en proposant ou masquant des fonctionnalités. MediaWiki permet de mettre en place des groupes d'utilisateurs disposant d'un ensemble de droits personnalisés. Un groupe d'utilisateurs est défini par un ensemble de droits auquel on donne un nom. On s'y réfère par la suite par la syntaxe `$wgGroupPermissions[nom du groupe] [nom du droit] = true ou false` dans le fichier `LocalSettings.php` de MediaWiki. Dans le wiki, l'affectation d'un utilisateur à un groupe se fait à travers la page spéciale *Gestion des droits des utilisateurs*, `Special:Permissions`.

Pour le projet CARE, la configuration actuelle (non figée) repose sur trois types d'utilisateurs (voir tableau 4.1) : 1) le responsable du pays qui possède un accès complet à l'ensemble des ressources ; 2) le responsable de région qui possède un accès complet à sa région et qui peut définir des annotations et 3) l'éditeur local qui n'a accès qu'aux fiches de sa région et qui ne peut pas définir d'annotation.

10. http://openlayers.org
11. ExtJS : bibliothèque de fonctions JavaScript permettant de créer des interfaces Web riches http://www.sencha.com/products/extjs
12. GeoExt : bibliothèque de fonctions JavaScript permettant de faire dialoguer ExtJS et OpenLayers

Nom du groupe	Droits associés
Region_Chief	edit (modifier) : intervention sur les documents du wiki upload : téléchargement de document
Country_chief	edit (modifier) : intervention sur les documents du wiki delete : suppression des documents

TABLE 4.1 – Groupes d'utilisateurs spécifiques au projet CARE

Si le mécanisme d'ACL (*Access Control List*) fourni de base par MediaWiki se révèlait insuffisant, il sera toujours possible de construire une ontologie dédiée à la représentation des droits d'accès aux documents, d'annoter les documents avec cette ontologie et d'exprimer les stratégies d'accès à l'aide de SPARQL. Nous pourrions nous servir de l'ontologie AMO [BFZK10] utilisée par le wiki sémantique SweetWiki [BGE+08] développé à l'INRIA Sophia-Antipolis.

4.10 Panorama des applications informatiques dans le domaine du patrimoine culturel

Dallas [Dal09] et Lock [Loc09] ont étudié les liens entre données, théories et interprétations dans les divers courants de l'archéologie en montrant comment les outils informatiques orientent leur évolution. La fin des années 1970 a vu apparaître les premières bases de données dans le domaine du patrimoine culturel [13]. Ce sont des bases thématiques (iconographie, mobilier, etc.) à des fins d'inventaire. La majorité des travaux, à partir du début des années 1990, s'est ensuite dirigée vers le couplage d'une base de données avec un Système d'Information Géographique (SIG) permettant des requêtes générant des représentations spatiales dynamiques. Pour dépasser les limites des SIG, Rivett [Riv97] a proposé ensuite de se concentrer sur la modélisation des données.
La tendance dominante à présent est aux « entrepôts de données ouverts » de données archéologiques et aux Services Web mis à disposition des utilisateurs. Le projet E-Culture [SAA+08] utilise les outils du Web Sémantique (OWL, RDF et SPARQL) pour annoter des collections de musées et faciliter leur recherche.

Wikis sémantiques pour l'archéologie

Isto Huvila [Huv09, Huv12] conseille d'utiliser les wikis sémantiques comme plate-forme pour les applications archéologiques. Nous présentons plusieurs projets archéologiques qui ont fait ce choix.
Le projet German Handbuch der Architektur a pour objectif de construire un wiki au moyen de la numérisation d'un volume (506 pages) du même nom [WKKL10]. Les auteurs cherchent à représenter deux sous-domaines par des ontologies : le sous-domaine de la gestion des documents (phrase, nom, numéro de page, etc.) et le sous-domaine de l'architecture (murs, matériaux de construction, etc.). Le traitement automatique des langages permet de connecter des concepts architecturaux avec un document spécifique, par exemple, les phrases qui mentionnent des éléments de construction utilisant un matériau donné. Une version publique est disponible à http://durm.semanticsoftware.info/wiki. Dans la même perspective, Plantec et al. [PRV09] utilisent le traitement automatique des langages pour transformer les pages d'un wiki en pages d'un wiki sémantique dans le domaine des collections scientifiques d'un musée. Les auteurs utilisent Semantic MediaWiki [KVV06] et CIDOC-CRM.

13. http://www.culture.gouv.fr/culture/inventai/patrimoine/

Le projet HermesWiki [RLB+10] est un plugin du wiki sémantique KnowWE [RBP08]. L'objectif est de fournir un aperçu concis et fiable de l'histoire grecque à des étudiants. Une ontologie, inspirée par le projet VICODI [aNDO05], pour le domaine historique a été élaborée. Une version publique est disponible à http://hermeswiki.informatik.uni-wuerzburg.de.

Le projet 3C2MA (« Climat, Catastrophes naturelles et Crises sanitaires des Mondes périméditerranéens dans l'Antiquité et au Moyen-Âge ») a pour objectif de collecter toute information historique concernant le climat, les événements tectoniques (séismes, éruptions volcaniques, etc.) et les crises sanitaires (épidémies, épizooties) ayant touché les pays riverains de la Méditerranée dans l'Antiquité et au Moyen Âge (http://www.3c2ma.com). Extraites et analysées par des historiens spécialistes de l'Occident musulman médiéval associés à des experts archéologues, géographes, médecins et vétérinaires épidémiologistes, les informations recueillies au sein des textes d'origines multilingues et polygraphes (arabe, hébreux, latin et grec) ont permis de construire une base de ressources termino-ontologiques. Cette base est développée à partir du wiki sémantique SweetWiki.

Dans la même idée, NavEditOW est un framework pour construire des sites Web basés sur des ontologies. Il a été testé dans deux projets : 1) des notices bibliographiques dans un portail sur la Préhistoire et la Protohistoire en Italie [BMV06], et 2) le projet SilkRoDE dont l'objectif est de recueillir, structurer et diffuser toutes les connaissances sur le patrimoine culturel de l'Asie centrale dans des domaines tels que l'archéologie, la géographie ou l'histoire. NavEditOW a intégré un moteur de wiki pour la visualisation des documents stockés [BMPV08]. MANTIC est un portail Web sur les informations archéologiques de la ville de Milan [MPV10]. Il intègre différentes sources de données et son schéma global est basé sur CIDOC-CRM. Porphyry (http://www.porphyry.org) propose un outil de mise en ligne de corpus, un système hypermédia pour la manipulation et l'annotation des documents par les experts. Deux domaines ont permis de valider les concepts : l'histoire sociale du monde chinois du XX^{e} siècle et des photographies de vases anciens.

4.11 Synthèse

Dans cette section, nous proposons un bilan de l'utilisation de *WikiBridge* par des archéologues dans le cadre du projet CARE et une comparaison entre l'utilisation d'une Base de Données relationnelles et d'un wiki sémantique comme plate-forme applicative.

4.11.1 Bilan de l'utilisation de *WikiBridge* dans le cadre du corpus CARE

Dans un premier temps, les utilisateurs saisissent dans *WikiBridge* du texte, téléchargent des images, des plans. À ce stade les utilisateurs peuvent faire des recherches plein-texte et naviguer à travers le contenu disponible. Une première formalisation est proposée par une structure définie *via* des formulaires qui par analogie avec les schémas de bases de données permettent de décrire les concepts fondamentaux manipulés sans décrire précisément leurs propriétés.
Nous avons mis en place un mécanisme d'annotations automatiques obtenues à partir de ces concepts. Un mécanisme d'annotation manuelle qui travaille avec une granularité plus fine permet d'annoter, par des experts, des portions de texte et des ressources internes ou externes. Afin de fournir une sémantique valide, les annotations sont définies en utilisant les termes d'une ontologie. Avec de telles annotations, une version performante de la recherche et de la navigation devient possible. *WikiBridge* permet aux chercheurs de créer librement et de faire évoluer leur cadre d'analyse en fonction de leurs hypothèses. Il rend aussi possible le partage de commentaires et l'expression des points de vue afin d'enrichir l'analyse des informations.

WikiBridge offre un cadre à la démarche des chercheurs associés au projet CARE. Leur démarche demande différentes étapes, en interaction les unes avec les autres qui se concentrent sur :
- la notion de documents (texte, image) qui doivent être produits de façon à obtenir une description exhaustive de l'objet étudié. Cette description signée par son auteur peut être amenée à évoluer par des révisions ;
- la notion d'interprétation car les documents produits doivent être analysés et interprétés au regard d'une ontologie qui joue le rôle d'une heuristique. Les démarches favorisant la découverte sont ainsi facilitées ;
- la notion d'interdisciplinarité car l'interprétation rassemble des chercheurs de différentes disciplines qui émettent des points de vue différents.

Ces étapes sont assurées par le wiki sémantique qui permet de faire apparaître la connaissance cachée derrière l'ambiguïté ou la polysémie du document. L'utilisation des technologies standards du Web Sémantique (RDF, OWL et SPARQL) permet l'exportation et la réutilisation des données et des annotations de *WikiBridge*.
L'objectif de *WikiBridge* est de permettre de construire la connaissance à partir du contexte exprimé par une ontologie et par des annotations : par un mécanisme d'analyse et de déduction des pages du wiki et des annotations posées, de nouvelles connaissances peuvent émerger et remettre en cause la conceptualisation effectuée (voir figure 4.17).

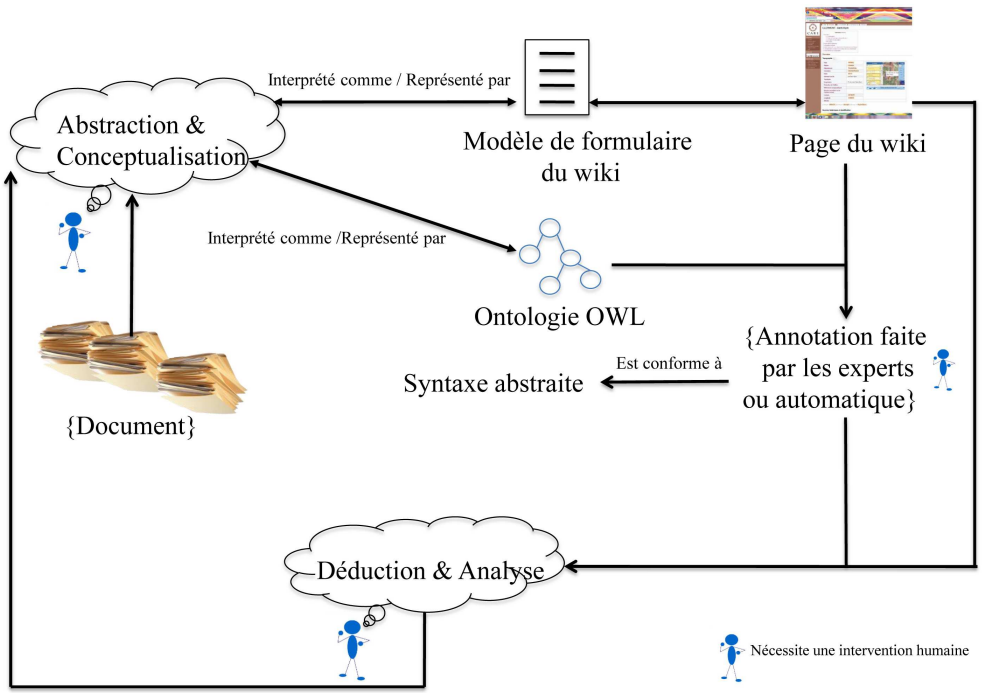


FIGURE 4.17 – Structure sémantique mise en place dans *WikiBridge* pour produire de la connaissance

4.11.2 Comparaison entre Base de Données relationnelles et wiki sémantique

La conception d'une application s'appuyant sur une Base de Données (modèle exécutable) impose : 1) de structurer les concepts sous la forme d'attributs simples et mono-valués, ainsi le schéma résultant de la normalisation et de l'adaptation au SGBD cible est en général éloigné

du modèle conceptuel ; 2) de construire le modèle dès la phase d'analyse en s'appuyant sur une connaissance du domaine à un instant donné ce qui fige la connaissance par les éléments modélisés ; 3) d'effectuer une modélisation par un petit nombre d'experts du domaine. De plus, l'expérience a montré que des modèles conçus par deux équipes différentes pour deux Bases de Données différentes avec des champs d'application voisins, seront difficiles à intégrer *a posteriori*, chaque concepteur possèdant son propre style de modélisation.
Pour gérer une connaissance informelle, la structure orientée document des wikis possède un avantage par rapport à une approche centrée base de données. En outre, les wikis sont des plates-formes qui fournissent des fonctionnalités collaboratives, ils sont simples à mettre en œuvre, ils supportent l'édition en ligne de documents. Les utilisateurs peuvent saisir et mettre en forme des articles, importer des ressources multimédia et lier les articles et les ressources au moyen de liens hypertextes. Les wikis proposent également un système de gestion de versions ainsi qu'un moteur de recherche plein-texte. Ces caractéristiques contribuent largement à leur succès. Cependant, le système de gestion d'articles et les liens ne sont pas suffisants pour modéliser finement la connaissance et garantir une cohérence sémantique des informations. Par exemple, le besoin d'une version structurée de Wikipédia s'est fait sentir : le projet DBpedia [14] [AL07] consiste à extraire des données de Wikipédia, à les stocker en RDF et à les interroger à l'aide de requêtes SPARQL.
Les wikis sémantiques proposent des solutions afin de pouvoir exprimer la sémantique, ils peuvent combiner le meilleur des deux mondes : la structure issue des Bases de Données et la flexibilité apportée par les wikis (voir figure 4.18). La connaissance est ainsi représentée sous une forme semi-structurée gardant les avantages à la fois des wikis et des Bases de Données. La modélisation est effectuée selon un processus collaboratif, dynamique et évolutif. En effet, par rapport à une application s'appuyant sur une Base de Données, les wikis sémantiques proposent : 1) une extension de la structure des documents *via* les formulaires et les annotations ; 2) une utilisation de connaissances établies *via* les ontologies ; 3) un support pour la collaboration *via* les débats autour des articles et le suivi des versions permettant entre autres une émergence du modèle de document et de la sémantique à partir des usages. Un wiki sémantique renvoie une connaissance minimale par les requêtes plein-texte, ensuite *via* les annotations il élargit la base de connaissances qui ne demeure pas figée puisqu'en permanence de nouvelles annotations peuvent être ajoutées. De plus dans le wiki, l'intégralité du corpus est stocké d'où la possibilité de le faire évoluer facilement permettant ainsi potentiellement de découvrir de nouvelles connaissances. Cependant, la représentation de la connaissance et la vérification de la sémantique des annotations est un des enjeux majeurs des solutions basées sur les wikis sémantiques car ils n'offrent pas encore le même niveau de contrôle que les SGBD.

Figure 4.18 – Wiki sémantique à la frontière des wikis et des Bases de Données relationnelles

14. http://www.dbpedia.org

Publications associées

WikiBridge a donné lieu à publication dans le domaine archéologique :

[1] Pascale Chevalier, Ludovic Granjon, Éric Leclercq, Arnaud Millereux, Marinette Savonnet et Christian Sapin, Base de Données annotées et Wiki pour la constitution du corpus numérique CARE, *Hortus Artium Medievalium*, Vol.18, n^{o}1, pp.27-35, 2012

[2] Éric Leclercq et Marinette Savonnet, Système d'Information pour la production de connaissances : l'approche wiki sémantique, *Ingénierie des Systèmes d'Information (ISI)*, Vol. 17, n^{o}3, pp. 143-166, 2012.

[3] Éric Leclercq et Marinette Savonnet, Structured Wiki with Annotation for Knowledge Management : an Application to Cultural Heritage, *International Journal of Digital Information and Wireless Communications (IJDIWC)*, Vol.1, n^{o}1, pp. 264-280, 2011

[4] Éric Leclercq et Marinette Savonnet, Adding Semantic Extension to Wikis for Enhancing Cultural Heritage Applications, *in Digital Information and Communication Technology and Its Applications (DICTAP)*, pp. 348–361, France, Juin 2011.

[5] Éric Leclercq et Marinette Savonnet, Système d'Information pour la production de connaissances : l'approche wiki sémantique, *INFormatique des Organisations et Systèmes d'Information de Décision (INFORSID)*, pp. 233-248, France, mai 2011

[6] Éric Leclercq et Marinette Savonnet, Semantic Wiki for the Protection, Emergency Management and Knowledge of Cultural Heritage, *Archeomatica*, Vol.1, n^{o}3, pp.46-48, 2010

[7] Éric Leclercq et Marinette Savonnet, Access and Annotation of Archaelogical Corpus via a Semantic Wiki, *Fifth Workshop on Semantic Wikis - Linking Data and People (Semwiki)*, Crète, Mai 2010

[8] Pascale Chevalier, Éric Leclercq, Arnaud Millereux, Christian Sapin et Marinette Savonnet, WikiBridge : a Semantic Wiki for Archaeological Applications, *XXXVIII Annual Conference on Computer Applications and Quantitative Methods in Archaeology - Fusion of Culture (CAA)*, ISBN : 978-84-693-0772-4, pp. 193-196, Espagne, Avril 2010

WikiBridge a donné lieu à publication dans le domaine bio-informatique :

[9] Éric Leclercq et Marinette Savonnet, Scientific Collaborations : Principles of WikiBridge Design, *Proceedings of the Workshop on Semantic Web Applications and Tools for Life Sciences*, Germany, December, 2010, http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-698/

[10] Éric Leclercq et Marinette Savonnet, Emerging Tools for Bioinformatics : The Semantic Wiki Approach, *Conférence I3-CRB, Biobanks : Interoperability, Database and Ontology*, France, novembre 2010

Chapitre 5

La plate-forme *eClims* : un exemple d'application dans le domaine des Sciences du Vivant

L'omniprésence de l'informatique, dans le quotidien du biologiste, est telle qu'on parle aujourd'hui de **biologie *in silico*** au même titre que les dispositifs *in vivo* et *in vitro*.

Sommaire

L'objectif de ce chapitre est de présenter le composant que nous avons développé, dans le cadre du co-encadrement d'une thèse, pour travailler sur des données cliniques en protéomique. La plate-forme de protéomique CLIPP [1] a permis d'expérimenter ce composant. Le suivi des échantillons de protéomique clinique nécessite tant pour le chercheur que pour le patient une gestion sans la moindre faille. L'utilisation d'un LIMS (*Laboratory Information Management System*) permet de prendre en compte les données en amont, pendant et en aval des expériences. Notre travail a consisté à gérer les données au moment de leur entrée dans le LIMS *via* le composant *eClims (experiments Clinical Information Management System)*.

5.1 Un LIMS comme support du Système d'Information des plate-formes de protéomique

Les LIMS sont des Systèmes d'Information Scientifique spécialisés, développés pour les chercheurs afin de les aider à faire face à la complexité croissante de leurs tâches [CHLBS98, Woo07].

1. CLIPP : CLinical and Innovation Proteomic Platform http://www.clipproteomic.fr/

Un LIMS peut être vu comme un ERP (*Enterprise Resource Planning*) qui intègre les principales activités d'un laboratoire en permettant à des utilisateurs de différents métiers (biologistes, statisticiens, bio-informaticiens) de collaborer. Un LIMS doit gérer les expériences biologiques en proposant le paramétrage des appareils et des protocoles expérimentaux, la gestion des échantillons initiaux et de leurs dérivés, le support du transfert d'informations entre partenaires mais aussi la mise en place d'un traitement semi-automatique des rapports d'expériences et leur communication [Pig08]. Utiliser un LIMS assure la traçabilité des données et des manipulations effectuées, automatise les tâches de collecte, de traitement et de stockage des données. La centralisation des données et des traitements favorise leur réutilisation et améliore la qualité des résultats au moyen de contrôles automatisés.

5.1.1 Caractéristiques essentielles des LIMS

Une des caractéristiques essentielles pour le développement de l'utilisation des LIMS est l'interopérabilité. L'utilisation de standards, de normes et de recommandations doit permettre de faciliter les échanges de données entre différents systèmes. D'un point de vue technique, l'interopérabilité peut être définie à plusieurs niveaux :
- au niveau syntaxique par l'utilisation de formats d'échange standards (langages XML, *Domain Specific Language* par exemple) pour favoriser le traitement automatique des données ;
- au niveau sémantique par l'utilisation de modèles formels des connaissances du domaine, par exemple sous la forme d'ontologies, pour une compréhension automatique des données échangées.

Quelques LIMS assurent l'interopérabilité des données. OpenLIMS [TGDS04] offre la possibilité technique d'échange de données au niveau syntaxique. Le LIMS Geneus développé par la société GenoLogics améliore l'interopérabilité en incorporant le standard de données FuGE.

Stephan et al. dans [SKT+10] proposent un état de l'art des LIMS et de leurs outils dans le cadre de la protéomique sans intégrer cependant les spécificités de la protéomique clinique. Les LIMS présentés stockent les données générées et assurent leur traçabilité sans offrir d'association entre les données acquises et les données sur les patients. Dans cette catégorie on trouve par exemple ProteinScape [TGH10] ou myProMS [PCB07].
Des LIMS aux fonctionnalités plus évolués existent dans des domaines très spécifiques tels que la recherche sur les végétaux (LIMS MassProt'INRA et la base de données PROTICdb [FDHM+05]) ou les études sur les médicaments (STARLIMS). Pour les LIMS utilisés dans les CRO (société de recherche sous contrat ou *Contract Research Organisation*), les données des études pharmacologiques sont très ciblées, le plus souvent ces logiciels ne permettent pas d'intégrer des données multi-centres et les méthodes statistiques sont standardisées, ce qui n'est pas le cas pour la recherche de bio-marqueurs en protéomique clinique.

5.1.2 Les LIMS en protéomique clinique

En conclusion, les LIMS existants ne répondent donc pas aux besoins de la protéomique clinique car ils sont construits pour des besoins très spécifiques et ils n'intègrent pas la gestion des données cliniques. De plus, les études concernent donc des données variées multi-modèles issues de multiples sources.
OLes LIMS en protéomique clinique doivent être capable de supporter :

1. la variabilité de la structure des données qui découle des questions abordées dans les études et de la diversité des outils techniques utilisés pour collecter les données. La variabilité de la structure des données dans les LIMS est plus importante que celle des données dans les Systèmes d'Information d'entreprises pour lesquels on utilise plutôt le terme

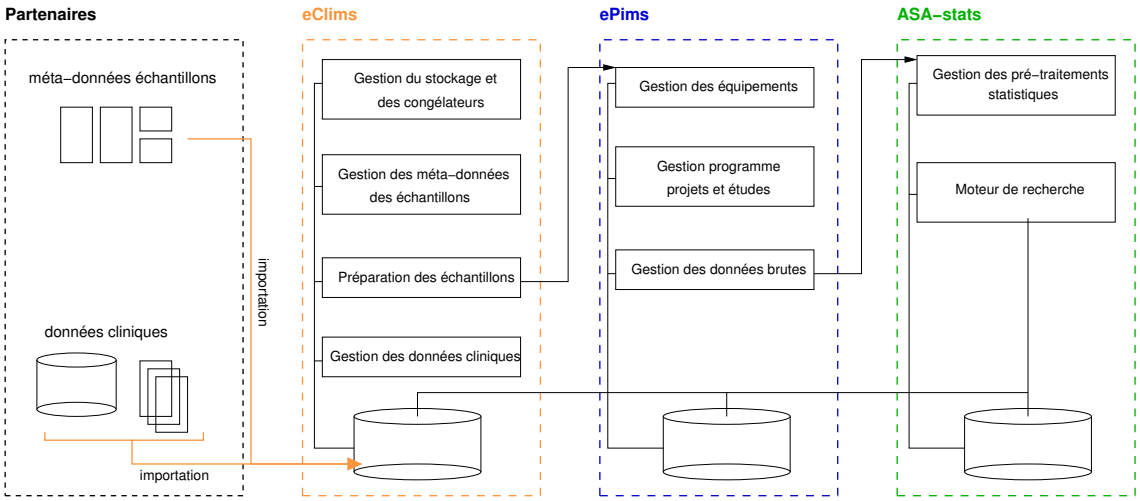


FIGURE 5.1 – Prise en compte des différentes étapes des études protéomiques grâce aux trois composants du LIMS choisi par CLIPP

d'hétérogénéité. D'un point de vue plus technique, la variabilité de la structure des données nécessite des évolutions coûteuses des schémas de bases de données ;

2. la variabilité de la sémantique induite par le caractère multidisciplinaire des recherches et des partenaires, nécessite l'utilisation d'ontologies.

Les LIMS en protéomique clinique requièrent donc une interopérabilité syntaxique et sémantique, des mécanismes d'extension de structure de données et des mécanismes de contrôle de qualité qui ne sont pas présents dans la plupart des solutions existantes. La qualité des données dans les LIMS est une autre caractéristique essentielle amplifiée par la nécessité des échanges entre partenaires. Les ontologies et des règles métier formalisées peuvent permettre de contrôler la qualité des données importées ou exportées par le LIMS. Par ailleurs, un contrôle qualité est nécessaire lorsque l'on veut fusionner des données présentes dans le LIMS provenant de plusieurs études.

5.2 Une solution modulaire de LIMS en protéomique clinique

La figure 5.1 présente les différents composants nécessaires pour gérer les données relatives aux recherches en protéomique clinique, en amont, pendant et en aval des expériences :

1. le composant clinique *eClims*, que nous avons développé, gère l'ensemble des données disponibles en amont des expériences réalisées au sein de la plate-forme. Il apporte à la plate-forme :
   - un outil de suivi et de contrôle de la qualité des données cliniques associées aux échantillons reçus ;
   - un mécanisme d'importation des données provenant des différents partenaires (la flèche orange de la figure 5.1). Le mécanisme d'importation traite les variabilités et les problèmes d'incompatibilité pouvant exister entre les différents systèmes. Le traitement de la variabilité sémantique issue des différents partenaires repose sur des ontologies. Le traitement de la variabilité de la structure de données repose sur un mécanisme d'annotation ;
   - la gestion du stockage des échantillons fournis à CLIPP transitant par des congélateurs de stockage avant d'être préparés en vue d'une expérience protéomique, le stockage des

données relatives à la préparation pré-expérimentale des échantillons est réalisé (voir figure 5.2) ;

2. le composant laboratoire *ePims* fournit les outils nécessaires à la configuration des expériences et au suivi des données brutes issues des spectromètres [DBB09] ;

3. le composant statistique, en cours de développement, permettra de gérer en partie les études statistiques, réalisées en aval des expériences, et nécessaires à la réalisation des conclusions des études protéomiques (c'est-à-dire l'identification de nouveaux bio-marqueurs). Le composant *ASA-stats* mettra en place des *workflows* évolutifs et réutilisables automatisant certaines tâches redondantes au sein des études statistiques. De plus, une fois les spectres prétraités, *ASA-stats* proposera aussi un assistant pour l'analyse statistique conduisant à la sélection de bio-marqueurs ainsi qu'un moteur de recherche multicritères sur les données de *ePims* et de *eClims*.

Parmi les objectifs visés par CLIPP, l'utilisation conjointe de ces trois composants doit fournir un système d'information permettant : 1) de gagner en reproductibilité, grâce à la création de réplicats de plaques de travail au sein de *eClims* et aux routines statistiques qui seront mises en place dans *ASA-stats*, 2) d'améliorer le contrôle qualité des données gérées et importées par une liaison avec l'ontologie, 3) tout en garantissant une traçabilité des données aussi bien en amont, pendant et en aval des expérimentations protéomiques.

Figure 5.2 – Placement des échantillons sur des plaques qui seront introduites dans le spectromètre de masse

5.3 *eClims* : qualité des données biomédicales

Pour assurer la qualité des données dans un Système d'Information biomédical, deux mécanismes doivent être particulièrement contrôlés : l'importation qui permet d'introduire de nouvelles données dans le Système d'Information et l'annotation qui permet de compléter des données existantes avec des données non prévues initialement. Le composant *eClims* que nous proposons est un composant spécifique du LIMS *ePims* afin de traiter le problème de la qualité des données biomédicales. Il a été développé dans le cadre de la thèse Région-Entreprise de Pierre Naubourg que j'ai co-encadrée.

5.3.1 Exigences auxquelles répond *eClims*

Lors de l'importation dans un environnement caractérisé par une forte variabilité inter-partenaires et inter-études, les composants mis en place doivent permettre de :

1. connaître et exploiter la sémantique des données des Systèmes d'Information source et cible. Pour cela, nous proposons d'utiliser des ontologies pour :
   - aligner les informations dans un référentiel où elles peuvent être comparées et
   - définir les règles de transformation des données source avant leur importation dans le système cible ;
2. améliorer l'extensibilité du stockage pour répondre à la variabilité des données utilisées par les diverses études. Si *eClims* ne peut pas prédire les évolutions il doit être conçu pour les prévoir ;
3. contrôler avant l'introduction dans la base de données, que les données source respectent les règles imposées par le système cible. En effet de nombreux partenaires fournissent des données de qualité inégale, cependant une fois importées avec les données présentes dans *eClims*, les données des partenaires doivent avoir la même qualité que les données existantes.

5.3.2 Architecture de *eClims*

La figure 5.3 présente les différents éléments fonctionnels pour gérer l'importation des données et leur annotation. Les données sont divisées en deux catégories : les données référentielles [2] [DHM+08] qui sont communes aux études et les données spécifiques qui sont traitées sous forme d'annotations stockées sous la forme clé-valeur [Str11]. Les données référentielles sont identifiables au sein d'une application, d'un système ou d'un ensemble de systèmes, par leur importance primordiale au bon fonctionnement des processus mis en œuvre. Les données référentielles présentent trois caractéristiques essentielles : partagées, stables et fréquemment utilisées. La figure 5.4 montre les données référentielles stockées dans un SGBDR et les données complémentaires stockées sous la forme d'annotations, ces deux types de données reposent sur l'utilisation d'une ontologie d'application.

Le processus d'importation effectue un contrôle de la qualité en vérifiant la complétude, la consistance et la cohérence. Le problème de complétude peut exister à deux niveaux : 1) un concept peut avoir des attributs obligatoires absents lors de l'importation, 2) un concept peut être associé à un autre concept de façon obligatoire et cette association n'existe pas lors de l'importation. L'inconsistance des données survient lorsque des caractéristiques différentes apparaissent pour un même concept (par exemple, deux échantillons prélevés sur un même patient dont l'un est associé à une donnée de sexe féminin et l'autre à une donnée de sexe masculin).

2. Les données référentielles sont aussi connues sous le nom anglais de *Master Data Management*

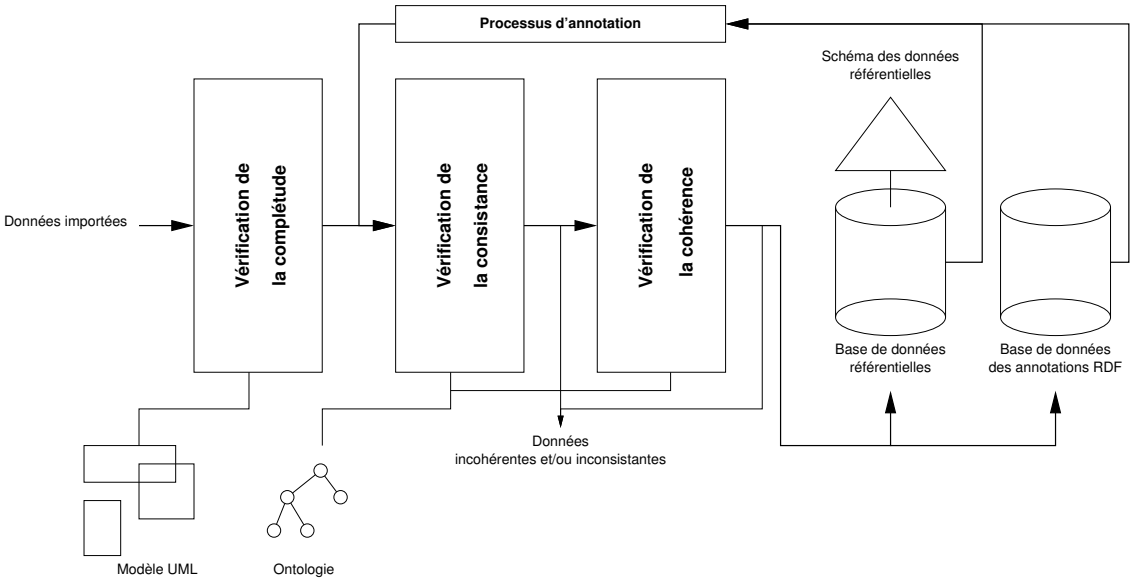


Figure 5.3 – Architecture du composant *eClims*

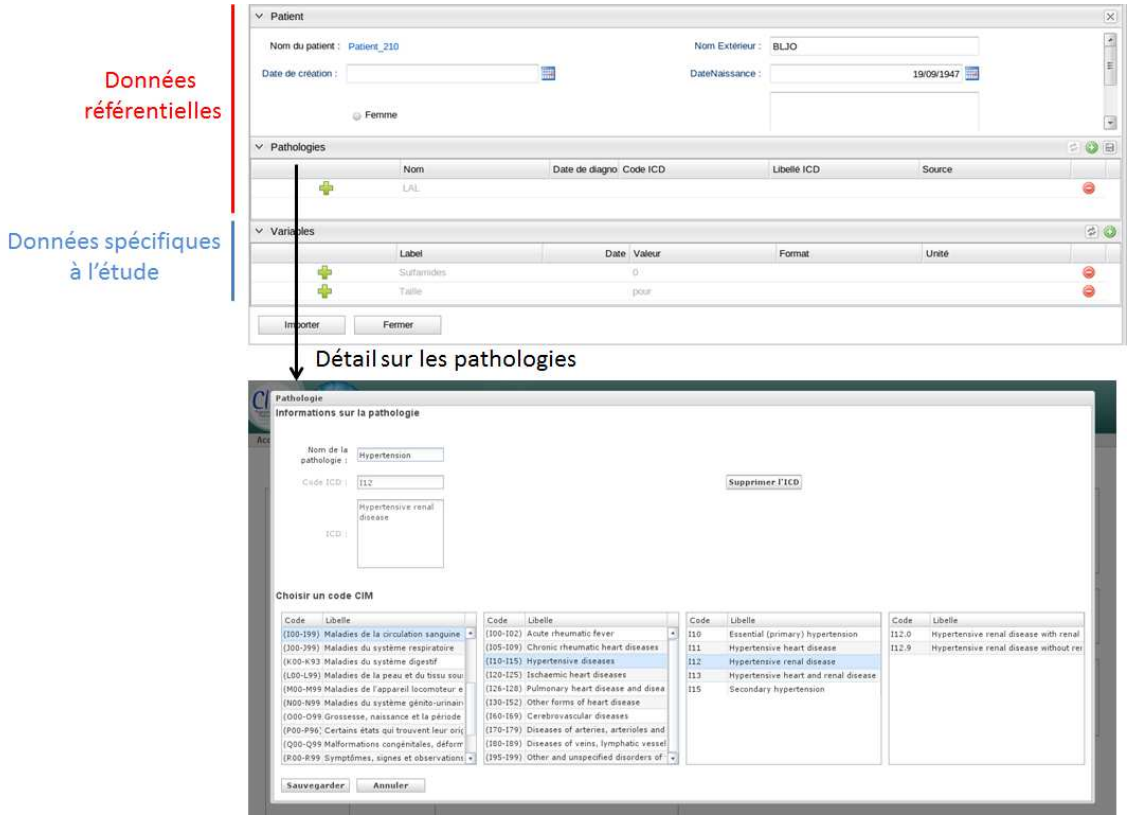


Figure 5.4 – Données cliniques traitées dans *eClims*

La cohérence des données touche à la connaissance du domaine (par exemple, est-il cohérent dans une étude sur le cancer du sein d'étudier un échantillon provenant d'un foie ?).
Le processus d'annotation travaille sur des données déjà existantes mais nécessite seulement

une vérification de la cohérence et de la consistance.
Dans la suite, nous présentons le processus d'importation et les mécanismes de contrôle de la complétude, de la consistance et de la cohérence des données, puis nous présentons le processus d'annotation.

5.4 Processus d'importation des données

Le processus d'importation des données que nous proposons, exploite un modèle UML représentant les données référentielles du système (voir figure 2.17 du chapitre 2) et une ontologie représentant la connaissance du domaine (détaillée en section 2.5.4 du chapitre 2). Cette ontologie d'application est utilisée comme médiatrice entre les schémas sources (c'est-à-dire les schémas des partenaires, essentiellement des fichiers CSV ou XML) et le schéma de *eClims*. Elle met en relation les jeux de données des sources avec le modèle de données de *eClims*, traite les problèmes de formats, de domaines et d'échelles. L'ontologie d'application *eClims* est structurée en trois branches distinctes :

1. la première branche décrit les concepts qui seront utilisés afin de mettre en correspondance les descripteurs des schémas des partenaires avec le modèle de données de *eClims* selon leur signification. Il s'agit d'une représentation ontologique des données de *eClims* ;
2. la deuxième branche est utilisée afin de définir quels types et formats de données sont concernés pour chaque descripteur de données des schémas ;
3. la troisième branche définit les opérations de conversion entre les différents types et formats de données.

La mise en correspondance est basée sur le concept de mapping. Un mapping schéma-ontologie MSO est un couple $< \{D\}; C_o >$ constitué d'un ensemble de descripteurs D et d'un concept C_o de l'ontologie o. Par exemple, un descripteur de type chaîne appartenant au système cible dont le label est *NumPatient* sera mis en correspondance avec le concept de l'ontologie *IdPatient* grâce au mapping MSO1 : $<< NumPatient; Chaîne >; IdPatient >$. Un descripteur de type entier appartenant à un jeu de données du partenaire P1 dont le label est *NumDossier* sera mis en correspondance avec le concept de l'ontologie d'application *IdPatient* grâce au mapping MSO2 : $<< NumDossier; Entier >; IdPatient >$. En effet, P1 étant un CHU, le patient est connu par son numéro de dossier. Une fois les mappings réalisés, nous proposons un processus reposant sur trois étapes (voir figure 5.5).

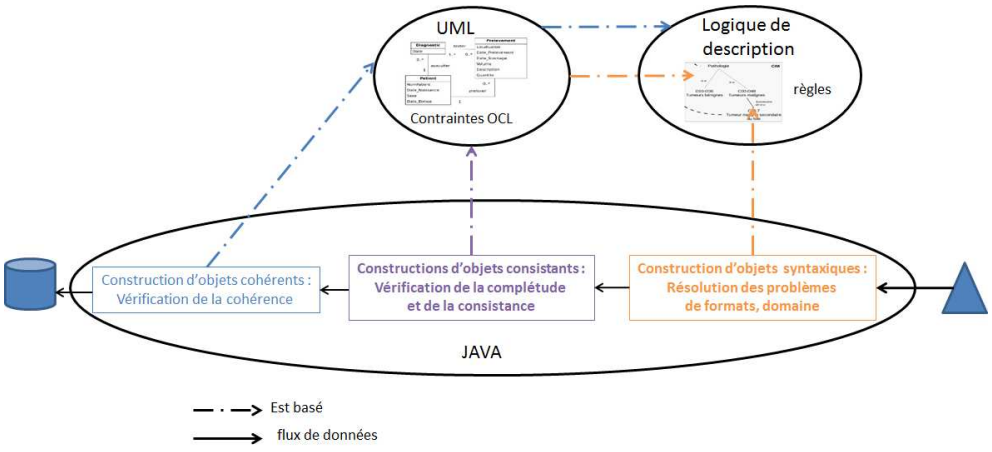


FIGURE 5.5 – Processus d'importation

La première étape consiste en la création d'objets Java correspondant aux données. Pour construire de tels objets syntaxiques : 1) nous associons des mappings schéma-ontologie afin de découvrir les couples de mappings partenaire-ontologie et *eClims*-ontologie qui correspondent, 2) nous convertissons éventuellement des valeurs des données du partenaire en des valeurs admises par *eClims* et 3) nous affectons ces valeurs au sein d'objets Java. Ces objets peuvent alors être manipulés afin de vérifier leur complétude, leur consistance et leur cohérence.

La deuxième étape traite des problématiques de complétude et de consistance. Le diagramme de classes UML de *eClims* permet de spécifier des contraintes portant sur :
- la complétude des données comme les attributs et les associations obligatoires ;
- la consistance des données comme l'unicité de valeur (deux objets possèdent les mêmes valeurs pour les attributs identifiants, ils doivent aussi posséder les mêmes valeurs pour les autres attributs), la comparaison de chemins. Par exemple, une contrainte détermine deux chemins depuis la classe `Patient` jusqu'à la classe `Pathologie`. Le premier chemin définit l'ensemble des pathologies d'un patient *via* la classe `Diagnostic`. Le deuxième définit l'ensemble des pathologies d'un patient *via* la classe `Hospitalisation`. Les deux ensembles d'objets doivent être égaux.

Nous avons utilisé les typologies proposées par Costal [CGQ+08] et Miliauskaite [MN05] sur les types de contrainte exprimables au sein des diagrammes UML pour réaliser ce niveau. Nous travaillons à partir d'objets complets afin d'assurer leur consistance par rapport au diagramme de classes UML.

La dernière étape porte sur la vérification de la cohérence et s'appuie sur une modélisation de la connaissance du domaine. Le processus de vérification de la cohérence repose sur un moteur d'inférence prenant en compte les faits, c'est-à-dire les objets nouvellement créés, la connaissance représentée sous la forme de l'ontologie et les règles métiers. Afin de ne travailler qu'avec des règles décidables, nous devons respecter les recommandations DL-Safe. Ces recommandations portent principalement sur la définition de règles travaillant sur des individus connus appartenant à des concepts nommés [MR10]. Pour respecter ces deux points, nous devons créer les individus représentant des objets Java, ces objets sont transformés en plusieurs individus représentant ses attributs. Puis nous devons affecter chaque individu à un concept de l'ontologie. Une fois les individus correspondant aux objets Java créés, le moteur de règles vérifie les règles portant sur l'ontologie. Par exemple, le moteur d'inférence utilise ces individus, l'ontologie et la règle énonçant « *qu'un prélèvement est valide si la pathologie pour lequel il est étudié et l'organe dont il provient sont mutuellement pertinents* » pour vérifier, pour chaque individu du concept `Prelevement` lié à un individu du concept `Organe` et à un individu du concept `Pathologie`, qu'une relation de type `organeTouché` existe au sein de l'ontologie. Si le moteur d'inférence ne trouve aucune relation définissant que le foie est un organe touché par la pathologie néoplasme du sein, il déterminera que le prélèvement est invalide ainsi que les données s'y référant.
À la fin de ce processus, nous obtenons soit des objets qui ont passé les trois vérifications avec succès soit des objets non valides. Tous les objets sont insérés dans le système cible, les objets non valides sont annotés comme "incertain" empêchant leur utilisation par les autres modules du LIMS. Ce choix est dicté par le fait que le matériel biologique humain est rare et qu'il est inconcevable de s'en séparer sans rechercher auprès des différents partenaires des informations complémentaires pour lever l'incertitude. Ce processus est présenté en détail dans [NSLY11].
La figure 5.6 est une capture d'écran présentant les cinq étapes nécessaires à l'importation 1) le choix des options d'importation, 2) la sélection du fichier à importer, 3) la sélection du programme de recherche (c'est-à-dire d'une étude, dans la figure 5.6 nommée "infection fongique"), 4) la visualisation du fichier et 5) la réalisation des mappings schéma-ontologie. La visualisation des données du fichier autorise leur modification avant importation. Enfin la réalisation des mappings schéma-ontologie permet de lier les descripteurs du fichier d'importation au schéma

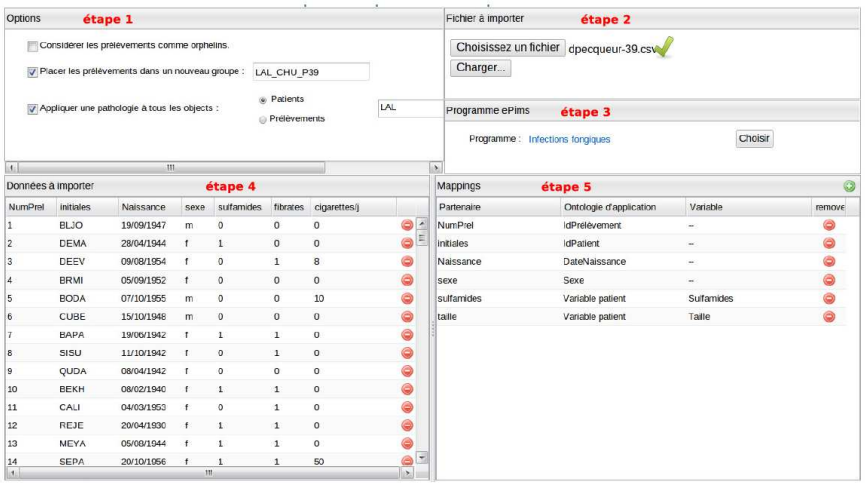


FIGURE 5.6 – Interface d'importation des fichiers cliniques

de *eClims*. Une fois la définition de l'importation réalisée, les données sont créées et présentées à l'utilisateur qui peut les modifier avant leur stockage en base de données.

Lors de l'importation, les données pour lesquelles la structure d'accueil (c'est-à-dire les tables du SGBD) est déjà existante sont stockées dans la base de données, les autres données sont transformées en annotations. Le mécanisme d'annotation peut être déclenché par l'importation ou par les utilisateurs du système souhaitant compléter des données. Dans ces deux cas, les composants mis en place doivent permettre de créer des annotations en utilisant la connaissance du domaine et de contrôler la consistance et la cohérence des annotations ajoutées (entre elles et par rapport aux annotations existantes).

5.5 Traitement de la variabilité de la structure de données

Nous traitons la variabilité de la structure de données par un mécanisme d'annotation qui permet d'ajouter dynamiquement, sans modifier les applications, des données complémentaires aux données existantes.
La modélisation d'applications complexes comprend différentes couches contenant des modèles spécialisés. Une couche peut être dédiée à la modélisation des interfaces utilisateur. Une autre peut être centrée sur la modélisation des composants internes des applications. Une dernière couche rassemble les modèles de données qui correspondent aux schémas de Bases de Données. Même si des couches d'abstraction permettant de réaliser le mapping objet-relationnel sont utilisées pour découpler les applications des données, la modification d'un modèle d'une application impacte les autres modèles de cette application du fait de leur interconnexion. Par exemple, au sein de la figure 5.7, présentant un Système d'Information composé de trois modèles, la modification de l'élément T2 du modèle des données impacte en cascade l'élément C2 du modèle des composants et la fonctionnalité U1 du modèle des fonctionnalités. De même, la création d'un nouvel élément T4 au niveau du modèle des données doit être prise en compte au niveau du modèle des composants par la création d'un nouvel élément C4 de ce modèle et présenté à l'utilisateur *via* la modification de la fonctionnalité U2. Ainsi, une modification d'un élément au niveau du modèle de données nécessite de faire évoluer les autres modèles. Ce travail d'évolution des modèles est long et complexe, il demande un haut niveau de compétence technique.

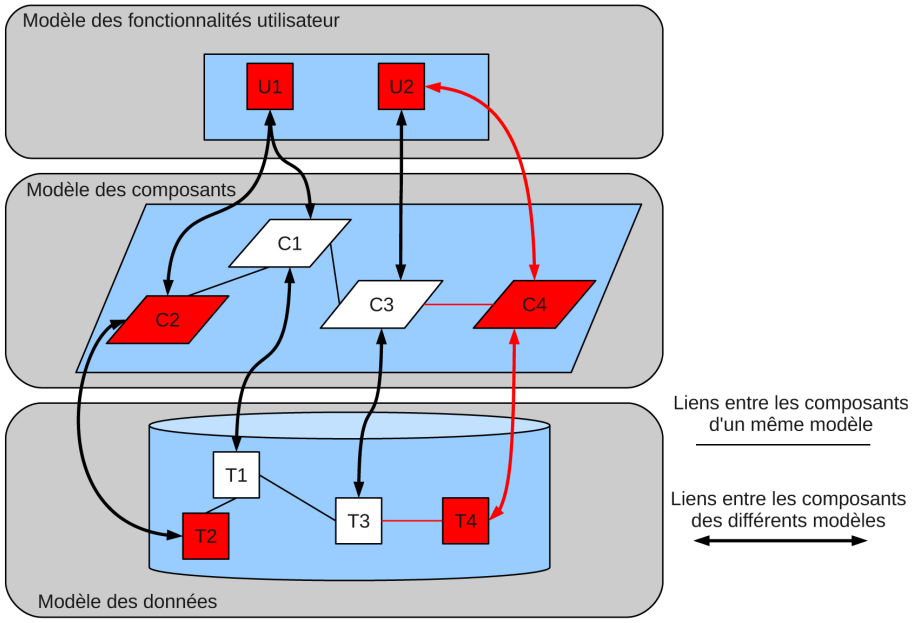


Figure 5.7 – Impact des évolutions des modèles sur les différentes couches composant une application (extrait de [Nau11])

L'objectif de notre proposition est d'intégrer au sein d'un LIMS, destiné à tous les types de plate-formes de protéomique, un mécanisme d'extension de schéma sous la forme d'annotations et sa prise en compte dans les autres modèles. En effet, beaucoup de petites structures de recherche ne possèdent ni les compétences ni les moyens de faire évoluer leur Système d'Information à chaque fois que les données stockées évoluent.
Tout comme la relation dans le modèle relationnel, l'annotation est une structure simple et quasi universelle qui permet de développer des composants génériques pour traiter l'extensibilité nécessaire à *eClims*.
L'annotation des données de *eClims* peut être effectuée, lors de la consultation et lors de l'importation des données. L'utilisateur peut choisir le prédicat et le type parmi une liste. Si le prédicat souhaité n'existe pas, il a la possibilité d'en créer un nouveau.
Les annotations mises en place dans *eClims* ne s'appliquent qu'aux données portant sur des patients, des prélèvements et des échantillons. L'implémentation a permis de valider les fonctionnalités suivantes :

– l'ajout de données complémentaires durant : 1) le processus d'importation, quand elles sont présentes dans les jeux de données des acteurs ou 2) "à la volée" lorsqu'elles sont fournies *a posteriori* par les partenaires dans le cadre par exemple d'une demande complémentaire d'information par la plate-forme CLIPP ;
– l'interrogation des annotations et des données en utilisant les `Criteria Queries` du framework Hibernate [3].

3. http://docs.jboss.org/hibernate/core/3.3/reference/en/html/querycriteria.html

5.6 Synthèse

La protéomique clinique est un domaine singulier par l'exhaustivité des données traitées (allant du patient aux résultats statistiques en passant par des données brutes) et par les analyses qu'elle requiert. Par ailleurs, le développement d'un LIMS prend également toute son importance dans les orientations actuelles de la protéomique clinique avec l'analyse statistique multi-modèles unifiée et l'apparition constante de nouveaux types de données (données "omiques", comme le génome, transcriptome, protéome, mais aussi spectroscopie infra-rouge, etc.) et en volumes toujours plus importants. Dans ce cadre, la plate-forme de protéomique CLIPP a fait appel au laboratoire LE2I et à la société ASA pour concevoir et mettre en place un *workflow* applicatif facilitant la gestion et l'analyse des données de spectrométrie de masse depuis la réception des prélèvements de patients jusqu'à la sélection de marqueurs pertinents. Le premier composant du *workflow*, *eClims* que nous avons développé, a pour objectif de prendre en charge le flux de données clinico-biologiques en provenance des cliniciens et des Centres de Ressources Biologiques. Le deuxième composant du *workflow* assure le suivi des traitements analytiques subis par les prélèvements pour leur analyse en spectrométrie de masse. *ePims*, développé par le laboratoire EDyP du CEA Grenoble, a été mis en place sur la plate-forme CLIPP afin de tracer le passage des échantillons sur les instruments, sauvegarder et organiser les acquisitions générées. L'interfaçage entre *eClims* et *ePims* garantit un suivi global des prélèvements. L'interopérabilité, l'extensibilité et le contrôle de la qualité sont des caractéristiques incontournables que nous nous sommes imposés tout au long du projet. Ces caractéristiques sont transversales à d'autres domaines d'application des LIMS. Ainsi la conception modulaire et Open Source de notre proposition contribue à sa réutilisation.

Une première perspective est de réaliser l'interfaçage entre les trois composants du LIMS. En effet, un troisième composant du *workflow*, *ASA-stats* développé par la société ASA, se charge de pré-traiter les acquisitions stockées dans *ePims*, de mettre à disposition des bio-statisticiens un moteur de recherche leur permettant d'extraire à la fois les données clinico-biologiques stockées dans *eClims* et les données protéomiques stockées dans *ePims* pour pouvoir ensuite les analyser avec des scripts R qu'ils ont écrits. Une fois l'interfaçage fait, CLIPP aura à sa disposition un moteur de *workflow* car les analyses pourront être refaites avec d'autres données ou avec les mêmes données mais les scripts R auront été paramétrés différemment ou encore avec d'autres scripts R, ce qui apportera plus de souplesse dans le *workflow* et plus de pertinence dans les analyses.

Une seconde perspective est le développement d'un composant cahier de laboratoire ou *Electronic Lab Notebook* (ELN). Alors qu'un LIMS assure la traçabilité des mouvements d'un échantillon dans le *workflow*, un ELN assure la traçabilité de ce qui est fait à un échantillon à une étape du *workflow*. Nous proposons d'utiliser l'approche wiki sémantique car elle répond directement aux besoins d'accès à des données multimédia (les scientifiques prennent des notes sur leurs expériences et les résultats obtenus, complétées par des schémas, des photographies, etc.), d'une plateforme collaborative, de sécurité des données, de gestion des utilisateurs. Chaque étape du *workflow* est vue comme une catégorie du wiki, un protocole expérimental est décrit par un formulaire. La création de formulaire est suffisamment simple pour permettre une évolution aisée des protocoles expérimentaux. Chaque expérience est décrite dans un article du wiki, elle correspond à une catégorie (c'est-à-dire à une étape dans le *workflow*) et à un protocole (c'est-à-dire au formulaire qui lui est associé à sa création), les échantillons concernés sont obtenus *via* des Services Web depuis le LIMS. Les notes prises lors de l'expérience sont annotées avec la même ontologie que celle construite pour le LIMS, la base de données stockant les annotations peut être commune aux deux outils.
Nous avons transposé ce concept de cahier de laboratoire à l'archéologie : dans le cadre du

projet *Corpus Lapidum Burgundiae*[4], l'élaboration du corpus est vue comme un ELN déployé sous la forme d'un wiki sémantique.

Publications associées

[1] Éric Leclercq, Marinette Savonnet, Pierre Naubourg, Traitement des variabilités métier dans les Systèmes d'Information biologiques, *INFormatique des Organisations et Systèmes d'Information de Décision (INFORSID)*, pp. 173-188, 2012

[2] Pierre Naubourg, Marinette Savonnet, Éric Leclercq, Kokou Yétongnon, An Approach to Clinical Proteomics Data Quality Control and Import, *Second International Conference on Information Technology in Bio- and Medical Informatics (ITBAM)*, LNCS 6865, pp. 168-182, 2011

[3] Pierre Naubourg, Marinette Savonnet, Éric Leclercq, Kokou Yétongnon, Approche préventive de la qualité des données d'importation dans le contexte de la protéomique clinique, *Revue des Nouvelles Technologies Informatiques (RNTI)*, E.22, ISBN 9782705682866, pp. 189-205, 2011

[4] Pierre Naubourg, Marinette Savonnet, Éric Leclercq, Kokou Yétongnon, Quality of clinical data in a proteomics LIMS, *Conférence I3-CRB, Biobanks : Interoperability, Database and Ontology*, Novembre 2010, France

[5] Pierre Naubourg, Marinette Savonnet, Marie-Noëlle Terrasse, Réalisation d'un LIMS protéomique - Modélisation des informations cliniques, *Second International conference on E-Medical Systems (E-MEDISYS)*, 2008, Tunisie

4. L'objectif du projet est de construire et de publier un corpus numérique sur l'extraction, l'usage et l'acheminement de la pierre en Bourgogne. Ce projet, financé par la région Bourgogne (CPER) et l'Union Européenne (FEDER), se termine en décembre 2013.

Chapitre 6

Conclusion & Perspectives

Sommaire

6.1 Conclusion

Les Systèmes d'Information Scientifique (SIS) sont des Systèmes d'Information (SI) dont le but est de produire de la connaissance et non pas de gérer ou contrôler une activité de production de biens ou de services comme les SI d'entreprise. Les SIS se caractérisent par des domaines de recherche fortement collaboratif impliquant des équipes pluridisciplinaires et le plus souvent géographiquement éloignées, ils manipulent des données aux structures très variables qui vont au-delà de la simple hétérogénéité : nuages de points issus de scanner 3D, modèles numériques de terrain, cartographie, publications, données issues de spectromètre de masse, de technique de thermoluminescence, etc. La gestion de données scientifiques nécessite une architecture de SIS ayant un niveau d'extensibilité plus élevé que dans un SI d'entreprise. Afin de supporter l'interopérabilité, l'extensibilité tout en contrôlant la qualité des données, nous avons proposé une architecture de SIS reposant sur un unique paradigme, l'annotation sémantique :

– des données référentielles fortement structurées, identifiables lors de la phase d'analyse et amenées à évoluer rarement. Elles constituent un pivot pour établir des liens avec les autres données ;
– des données complémentaires multi-modèles (matricielles, cartographiques, nuages de points 3D, documentaires, etc.) représentées sous la forme d'annotations.

Dans ce cadre, les annotations offrent ainsi une contextualisation des données qui permet de vérifier la cohérence des annotations ajoutées, par rapport à la connaissance du domaine. Nous avons proposé un modèle d'annotation pour construire des annotations sémantiques à base ontologique dont la cohérence et la consistance peuvent être contrôlées par une ontologie, des règles.

Nous avons montré, à travers deux collaborations transdisciplinaires, que l'annotation est une structure universelle permettant de développer des composants génériques prenant en charge la

variabilité nécessaire aux SIS.
La première collaboration avec le laboratoire ARTeHIS a débuté en 2009 dans le cadre de l'ANR CARE (*Corpus Architecture Religiosae Europeae* ANR-07-CORP-011). Cette ANR réunissait plus de soixante archéologues en poste dans une vingtaine d'universités, des dessinateurs topographe, le pôle de géomatique de la MSH de Dijon et notre équipe. L'objet de notre participation était le développement d'une plate-forme collaborative gérant des connaissances relatives au corpus européen des édifices religieux du IV^e siècle au X^e siècle. Cette plate-forme offre les outils nécessaires au travail de synthèse sur le référencement des édifices religieux et sur leurs évolutions au cours des siècles à travers un modèle spatio-temporel spécifique. La connaissance des archéologues est modélisée à travers une ontologie d'application qui spécialise une ontologie de domaine. L'application est basée sur MediaWiki que nous avons étendu afin d'y intégrer la sémantique des domaines impliqués et la démarche scientifique spécifique à la discipline. Cette plateforme s'inscrit dans le cadre du Web 2.0 avec l'utilisation des recommandations du Web Sémantique (RDF, OWL, SPARQL). Les aspects contributifs sont couverts par le wiki et les aspects sémantiques sont couverts par les annotations et la modélisation de la connaissance du domaine.
La seconde collaboration concerne la plate-forme de protéomique CLIPP et la société ASA de Montpellier. Elle a abouti à l'application *eClims* qui met en œuvre un outil d'importation basé sur le couplage entre des modèles représentant les sources et le système protéomique, et des ontologies utilisées comme médiatrices entre ces derniers. Les différents contrôles que nous mettons en place garantissent la validité des domaines de valeurs, la complétude, la consistance des données et leur cohérence. Le stockage des annotations est assuré par une Base de Données orientées colonnes associé à une Base de Données relationnelles.
Une nouvelle collaboration, dans le domaine archéologique, est mise en place avec le centre d'études médiévales Saint Germain d'Auxerre. Elle concerne le développement d'un Système d'Information Scientifique traitant la nature et l'usage de la pierre bourguignonne à partir des données fournies par le Bureau de Recherches Géologiques et Minières [Büt11]. Ce projet est financé par la Région Bourgogne (CPER) et l'Union Européenne (FEDER).
Ces collaborations représentent un large champ d'expérimentation en terme d'objectifs, de partenariats et de plate-formes technologiques (wiki, Java/J2EE, Services Web).

6.2 Perspectives

Dans la suite, quelques prolongements possibles sur les annotations et le wiki sémantique sont présentés.

6.2.1 Validité des annotations

Notre objectif est d'améliorer la qualité des annotations posées. L'assistant d'annotation, développé dans la version actuelle de *WikiBridge*, offre déjà un contrôle de premier niveau des annotations où deux types de contraintes structurelles sont vérifiés en utilisant l'ontologie : 1) les domaines de valeurs des objets ; 2) la consistance structurelle des prédicats : par exemple une cathédrale peut avoir une nef mais ne peut pas posséder un atrium. Cependant les contraintes ne peuvent toutes être vérifiées avec la structure de l'ontologie et le mécanisme d'inférence des Logiques de Description. En effet au delà de la syntaxe des annotations :
– deux annotations portant sur le même sujet ne peuvent pas être contradictoires ;
– il existe des règles de cohérence plus globales dépendantes fortement du domaine : par exemple, il n'existe pas en France, pour les siècles concernés, d'église dont les murs ont été construits avec des techniques utilisant la terre crue alors que ce type de technique peut

avoir été employée en Irlande.
Nous projetons de travailler sur un ensemble de mécanismes basés sur des règles écrites en SWRL (un sous-ensemble). Pour ce faire nous utiliserons différents moteurs d'inférence (Pellet, HermIT) pour raisonner soit avec l'hypothèse du monde clos (contraintes) soit avec l'hypothèse du monde ouvert, ceci pour déduire de nouvelles connaissances.

6.2.2 Système de recommandations

Notre objectif est de mettre en place un système de recommandations par profil d'utilisateur (archéologue, médiéviste, historien d'art ou clinicien, biologiste) pour proposer à chaque utilisateur les termes utilisés par sa communauté ou au contraire déterminer des termes peu usités dans sa communauté. Ceci permettra de détecter d'éventuelles incohérences ou déterminer des termes pivots entre communautés.

6.2.3 Annotation des traitements

Les traitements et leurs enchaînements peuvent être vus par le Système d'Information sous la forme de données, par exemple, des documents BPEL pour les enchaînements et des scripts R ou Matlab pour la réalisation des traitements. La réification du *workflow* en données permet l'annotation de ses différents composants en utilisant le modèle que nous avons développé. Cela devra favoriser la réutilisation des traitements et permettre de contrôler la qualité et la provenance des résultats. D'ailleurs, Tireau et al. [TDC+10] ont proposé une ontologie pour annoter des scripts R.

6.2.4 Confluence entre annotation et fragmentation

Les annotations génèrent une masse de données difficilement traitable par des systèmes de gestion de Bases de Données Relationnelles. Le problème repose principalement sur le passage à l'échelle de la réification dans le mécanisme d'annotation (générant ainsi des requêtes récursives) face à un nombre de tuples très important. Des publications récentes montrent que le traitement de dix millions de tuples RDF nécessite plusieurs centaines de secondes [KGSS12]. De surcroît, l'ajout de règles logiques au sein de ces tuples pour matérialiser des contraintes métiers, et pour améliorer la qualité des réponses d'une requête SPARQL peut très vite donner lieu à des temps de réponse inacceptables. Les techniques habituelles sont donc peu efficientes dans des applications interactives.
Mon objectif est l'étude des méthodes de placement des annotations, représentées par des triplets RDF, sur bases de données verticales avec prise en compte de la sémantique portée par les annotations et les contraintes exprimées sur RDF. Deux sémantiques sont à travailler : l'une, intrinsèque au graphe d'annotations portant sur un sujet donné, l'autre, référentielle portant sur les relations entre les termes de l'annotation et ceux de l'ontologie (les annotations utilisent le plus souvent des termes issus d'une ontologie).
Je projette l'écriture d'algorithmes exploitant les structures de graphes et les travaux sur la fragmentation dans les Bases de Données Relationnelles et Objets [1], afin de fragmenter les annotations et de les regrouper.
Je propose deux grandes approches :

1. la **fragmentation verticale** consiste à découper un graphe d'annotation complexe de manière à regrouper les annotations simples utilisées conjointement. Cette approche fera appel à des techniques de *graph clustering* [Sch07] qui permettent de partitionner un

1. Ma thèse et mes travaux jusqu'en 2001 ont porté sur la fragmentation et l'allocation de Bases de Données Orientée Objets

graphe (en clusters) de sorte que les nœuds ayant des caractéristiques communes fassent partie d'un même cluster. Nous devrons valider la pertinence de cette approche, utilisée jusqu'à présent essentiellement dans le domaine des réseaux informatiques et autres réseaux sociaux, et l'adapter ;

2. la **fragmentation horizontale** consiste à répartir sur différents nœuds de calcul/stockage des graphes d'annotations souvent manipulés ensemble car ils ont des propriétés communes. J'envisage de développer deux stratégies :
   (a) une stratégie de mesure de distance entre deux graphes s'appuyant sur des algorithmes de *graph matching* [Bun00] qui permettent d'estimer une distance de similarité entre deux graphes en fonction des besoins exprimés. Il s'agit d'une question difficile à résoudre pour laquelle de nombreuses heuristiques ont été mises au point dans différents contextes. Le point crucial sera ici de définir correctement les paramètres de similarité pour qu'ils correspondent au mieux au monde réel.
   (b) une stratégie qui consiste à typer un graphe (au sens du typage des langages) afin de déterminer l'ensemble des graphes de même type.

6.2.5 Extension de *WikiBridge* : Wiki sémantique distribué

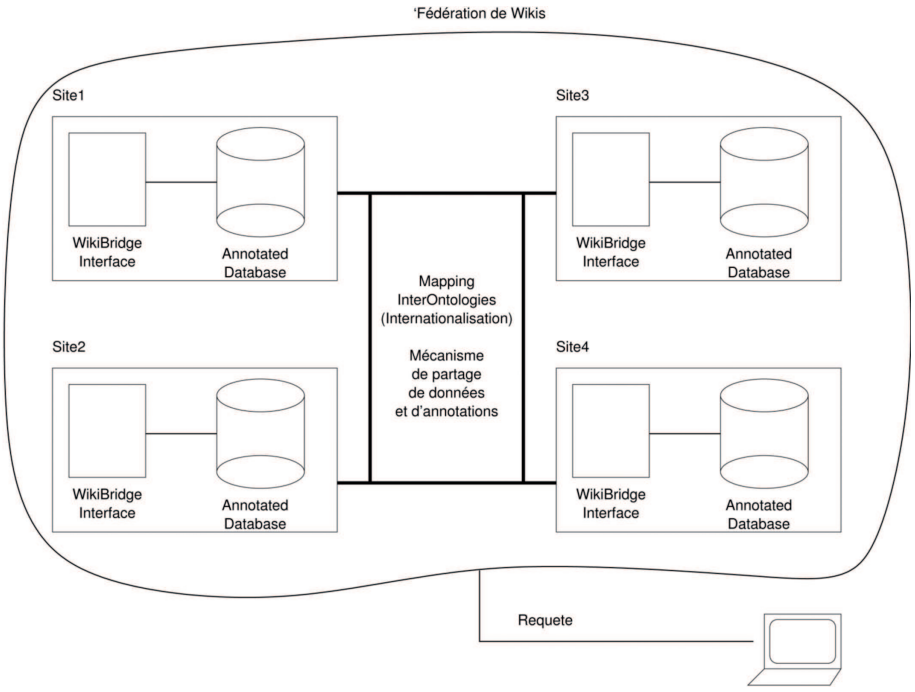


Figure 6.1 – Wiki distribué

La coopération internationale dans un environnement multi-langues et multi-cultures nécessite d'étendre *WikiBridge*. Une suite logique des travaux de l'ANR CARE est de créer les fonctionnalités permettant de gérer cette distribution (voir figure 6.1). En supposant que chaque participant possède son propre système, sauvegardé par la grille ADONIS [2], et le gère de manière autonome, les problématiques scientifiques mises en exergue par ces nouvelles fonctionnalités sont de deux types : d'une part l'alignement des ontologies (avec les liens inter-ontologies) et

2. http://www.tge-adonis.fr/

d'autre part le partage de données à grande échelle dans une coopération de systèmes sans schéma, au moyen d'interrogations SPARQL multi-systèmes. Nous détaillons dans les paragraphes suivants les verrous scientifiques et technologiques associés à ces problématiques.

L'alignement d'ontologie, problématique traitée depuis les années 2000, a donné lieu à de nombreux travaux dans le domaine des Systèmes d'Information. Les solutions développées abordent peu les aspects collaboratifs et dynamiques de l'alignement. Dans le contexte des applications que nous avons développées, les utilisateurs du système doivent pouvoir définir leur propre ontologie d'application. Une ou plusieurs ontologies de domaine peuvent également être définies de façon collaborative en se basant sur les usages des concepts. La distribution implique donc de pouvoir réaliser un alignement évolutif et dynamique entre les ontologies des différents systèmes. Cet alignement pourra par exemple être piloté par les communautés d'utilisateurs afin de faire émerger les liens entre des concepts communs.

Une coopération sans schéma nécessite l'échange des annotations, de leur contexte et de leur sémantique entre les systèmes. Notre objectif consiste à déterminer des ensembles minimaux et auto-suffisants d'informations à partager. Le modèle des tuples sémantiques (*semantic tuples*) peut nous servir de modèle d'échange [NOVS11]. L'ensemble des tuples sémantiques pourra alors être interrogé au moyen de requêtes SPARQL. Chaque système devra ensuite valider chaque tuple sémantique obtenu par rapport à sa connaissance locale. Nous projetons dans cette partie d'utiliser des techniques de vérification de modèles. D'un point de vue plus technique, il s'agira d'intégrer dans SPARQL la capacité d'interroger simultanément plusieurs ontologies et bases d'annotations.

Annexe A

Modélisation du corpus CARE (*Corpus Architecturae Religiosae Europeae - IV-X saec.*) : du conceptuel à l'exécutable

« *Tu est Pierre, et sur cette pierre je bâtirai mon église* . . . »Matthieu 16–18

Sommaire

1 Corpus CARE

L'objectif du projet international CARE (*Corpus Architecturae Religiosae Europeae - IV-X saec.*) est la constitution d'un corpus des monuments chrétiens antérieurs à l'an Mil. Le caractère novateur du projet est caractérisé par :

– l'exhaustivité : à terme, la quasi totalité des sources disponibles aura été dépouillée. La délimitation d'une zone géographique suffisamment vaste (la majeure partie des pays européens) permettra de travailler par dessus les frontières et pays contemporains ;

– l'interdisciplinarité : le programme de dépouillement et d'interprétation des données rassemble des archéologues, des historiens, des historiens de l'art, des dessinateurs topographes. Il est conçu comme un processus collaboratif permettant non seulement d'échanger de l'information entre les champs disciplinaires, mais aussi d'enrichir les pratiques respectives par la confrontation, autour d'un édifice, des problématiques et des méthodes propres à chacun. Il s'agit de traiter un édifice dans son ensemble, en identifiant et en intégrant toutes les relations entre les différents éléments impliqués et donc de comprendre un objet complexe dans sa globalité. L'objectif est de synthétiser et relier chaque savoir disciplinaire aux autres ;
– la transdisciplinarité : cette coopération se double d'un partenariat entre la recherche en Sciences Humaines et Sociales (ici l'archéologie) et la recherche informatique. La création d'une plate-forme informatique fournit ainsi aux informaticiens un terrain d'expérimentation réel des concepts qu'ils élaborent en matière d'ingénierie des ontologies et du Web Sémantique ;
– l'innovation : la mise en place d'une plate-forme informatique vise à moderniser les méthodes de travail en archéologie en libérant l'archéologue de routines manuelles fastidieuses. Il permet également de limiter les risques introduits par une sélection lacunaire des données ;
– la valorisation de la recherche : la plate-forme informatique est conçue de manière à pouvoir être transposé pour répondre aux besoins d'autres domaines de recherche.

Ce projet nous a aussi permis d'appliquer les notions de modèle conceptuel et modèle exécutable à une application concrète.

2 Validation du modèle conceptuel UML : catégorisation de la fiche de dépouillement de l'église Saint-Pierre-Estrier à Autun (71)

Dans le chapitre 2, nous avons présenté le modèle conceptuel UML du projet CARE que nous rappelons en figure A.1. Afin de le vérifier, nous avons catégorisé une fiche.
Les caractéristiques principales de la fiche sont prises en compte dans la figure A.2.
Dans la classe topographie
Références, topographie, sources et bibliographie
1. DONNEES
1.1. Topographie
Région : Bourgogne
21
Autun
Adresse : rue de l'Ermitage
Référence cartographique :

Dans les trois instances de la classe documentSourceAnalysee avec comme statut sourceHistorique
1.2. Sources historiques et identification
Sources archéologiques :
SAPIN C., "Autun Saint-Pantaléon", Archéologie Médiévale, t. 14, 1984, p. 312-313.
SRA Bourgogne : Rapports annuels (à compléter)
Cartulaire de l'église d'Autun, éd. A. de Charmasse, t. I-III.
Cart. Eglise I, p. 146, mention de Saint-Pierre en 1233 où l'on parle d'une villa Sancti Petri de Lestrée.
Grégoire de Tours, Liber in gloria confesorum 73. In Monumenta Germaniae historica, Scriptores rerum merovingicarum, 1885, t. I, 2, p. 791.

Dans des instances de la classe documentSourceAnalysee avec comme statut bibliographie

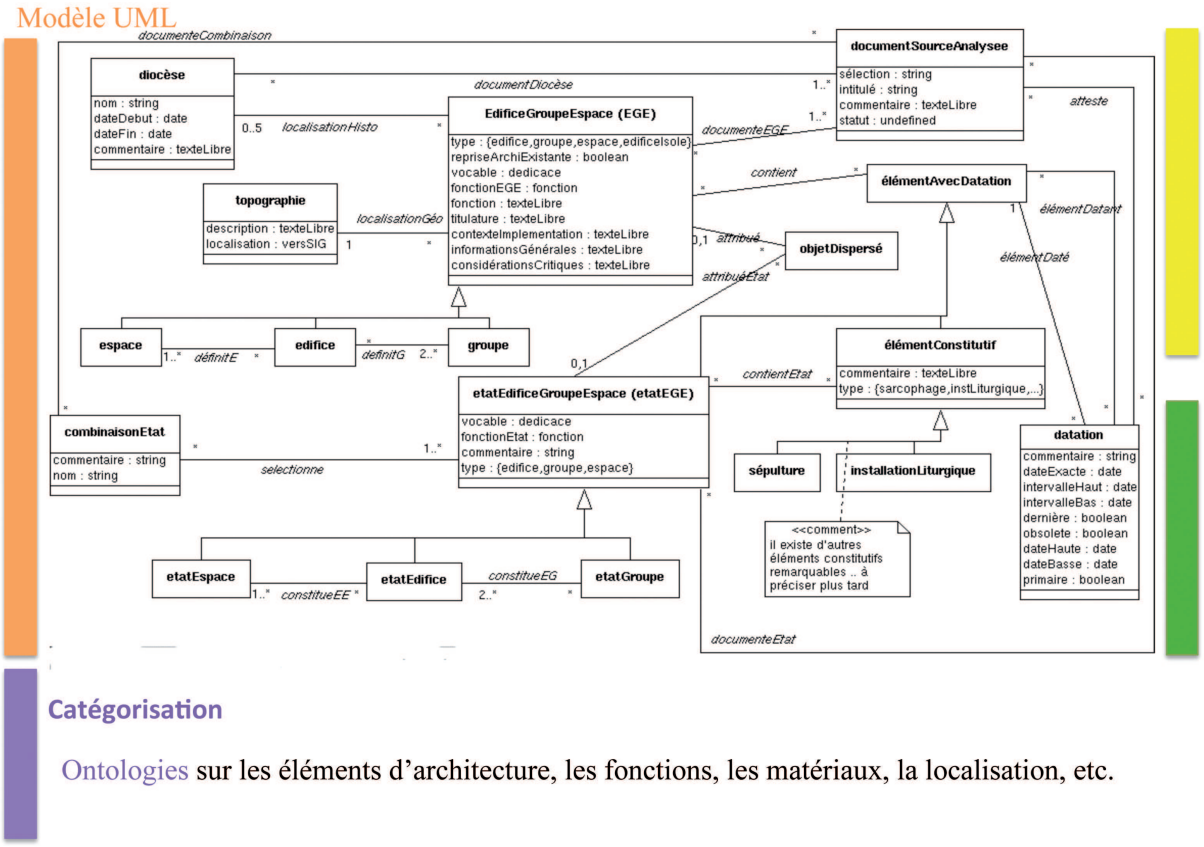


FIGURE A.1 – Vision globale du modèle conceptuel correspondant au projet CARE

9. BIBLIOGRAPHIE
BERTHOLLET J., La question des premiers, 1945, p. 38-44.
BERTHOLLET J., La restauration d'Autun, 1949, p. 115-123.
BERTHOLLET J., Esquisse de l'évolution, 1941-1944, p. 213-226.
CHARDON-PICAULT P. et HAMBLIN M., 1999, p.303.
DEVAUGES J.-B., 1979, p. 459
FONTENAY H. de, Autun et ses monuments, 1889.
GAILLARD DE SEMAINVILLE, 1983, P. 406 ET 1985, P. 266.
LECLERCQ Dom H., Autun, col. 3190-3203, 1907.
LEMPREUR Père J., Dissertations historiques, 1706.
MUNIER J., Recherches et mémoires, 1660.
PIETRI C., PICARD J.-C., Autun, 1986, p. 37-45.
REBOURG A., 1986, p. 72-73.
REBOURG A., 1996, p. 86-88.
RICHARD J, Sur les débuts du christianisme, 1948, p. 70-72.
SAPIN C., "L'ancienne église de Saint-Pierre-L'Estrier à Autun", Archéologie Médiévale, t. 12, 1982, p. 51-105.
SAPIN C., "Autun Saint-Pantaléon", Archéologie Médiévale, t. 14, 1984, p. 312-313.
SAPIN C., avril 1984, p. ?.
SAPIN C., 1984, p. 113-129.
SAPIN C., BERRY W., "Un sarcophage de plomb de Saint-Pierre-l'Estrier", Mémoires de la

Société Eduenne, LIV, fasc. 4, 1984, p. 285-289.
SAPIN C., BERRY W. et YOUNG B.K., Saint-Pierre-l'Estrier, Gesta, 25, n^o 1, 1986, p. 39-46
SAPIN C., 1986, p. 24,124 à 132, 161, 165-166, 178, 239, 245, 252, 264 n, 277n, 279 n.
SAPIN C., 1987, p. 364-375.
SAPIN C., 1987-1988, p.34.
SAPIN C., 1998, t. 3, p. 64-69.

Dans l'attribut titulature de la classe EGE
1.2. Sources historiques et identification
Titulature connue : Le vocable de Saint-Pierre apparaît en 843 : Monasterium Sancti Petri seu Sancti Stephani in suburbio civitatis (Cart. Eglise, p. 47) et sera rattaché à Saint-Nazaire d'Autun, au XIX^e siècle elle prendra le nom de Saint-Pierre-l'Estrier. La restauration d'un monastère dédié à l'évêque Cassien par le roi Robert le Pieux au début du XI^e siècle pourrait correspondre à ce site (Cf. Vie de Robert le Pieux par Helgaud de Fleury, R.H.Bautier et G.Labory éd.et Trad., Paris, CNRS, 1965, p.89)

Dans la classe diocèse
1.2. Sources historiques et identification
Diocèse : Autun

Dans l'attribut contexteImplantation de la classe EGE
1.3. Contexte d'implantation
L'église se trouve dans une nécropole antique très importante sous le bas-Empire. Les tombes des premiers chrétiens viennent s'implanter entre les tombes païennes. C'est là qu'on trouve les tombeaux ou mausolées des premiers évêques de la cité, aux abords de la basilique Saint-Etienne, extra muros. Lors de son passage à Autun, saint Germain verra la tombe de l'évêque Cassien en 430. Un premier récit sur la présence d'un cimetière est fourni par Constance de Lyon. Grégoire de Tours mentionne également le cimetière sans donner le nom de ce dernier. L'église Saint-Pierre se situait à cent mètres à l'Est de la basilique Saint-Etienne. Cette basilique a été entièrement détruite à la fin du XVII^e siècle. L'église Saint-Pierre se situe à proximité de la voie romaine conduisant d'Autun par la porte Saint-André vers Langres et Besançon. Cette voie a été remplacée par une route communale. Elle est plus proche d'une voie secondaire situé à cinquante mètres de Saint-Pierre. Elle est entourée de quelques maisons du XIX^e siècle formant un hameau. Une partie du "mausolée" se situe sous la rue de l'Hermitage qui passe devant l'église. Au XV^e siècle, apparaît la mention Saint-Pierre de l'Estrier et au XIX^e siècle Saint-Pierre-l'Estrier, pour le différencier de Saint-Pierre-Saint-Andoche. L'ensemble du site se nomme "l'Estrier" qui dérive de la "Strata". Lors de fouilles archéologiques (1976-1986) des structures d'une construction antique (villa) ont été mises au jour datant des I^{er} et II^e siècle sur lesquelles l'église va venir se substituer.

Dans l'attribut fonction de la classe EGE
1.3. Contexte d'implantation
Fonction :
D'abord église cimetériale à partir du IV^e siècle, elle est signalée comme église paroissiale dans un texte de 1328. L'église garde son utilisation funéraire à côté de son rôle d'église paroissiale. Dès le X^e siècle, la mention de chanoines titulaires du chapitre et en même temps abbés de Saint-Pierre et de Saint-Etienne confirme son attachement à l'Église d'Autun. Jusqu'au XVII^e siècle, la mention honorifique d'abbé de Saint-Pierre est présente dans le cartulaire des évêques d'Autun. À la Révolution, elle est transformée en grange.

Dans l'attribut informationsGénérales de la classe EGE
2. Informations générales

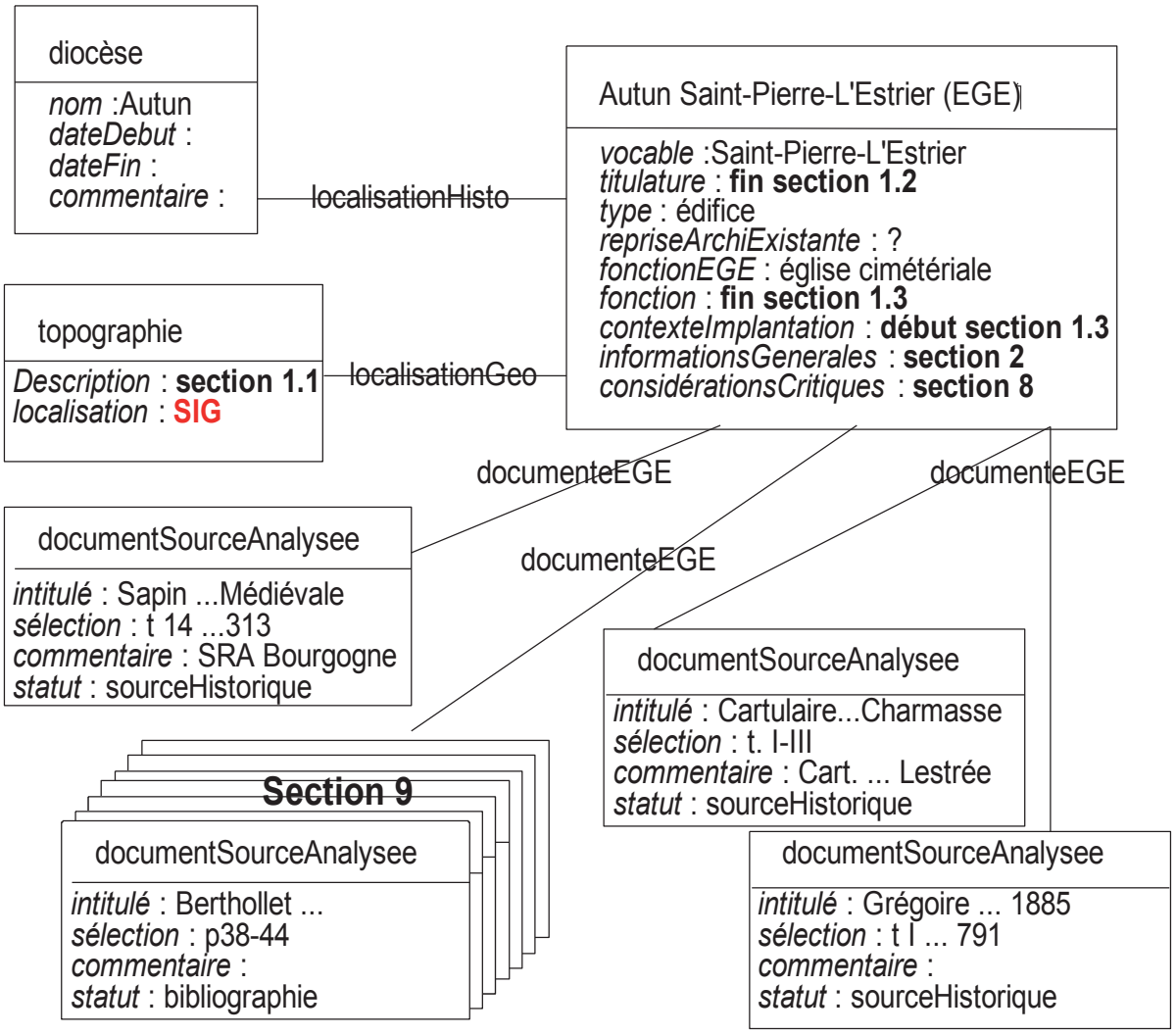


FIGURE A.2 – Partie principale de la fiche Saint-Pierre-L'Estrier (sans les états)

L'édifice actuel a été transformé en grange et ne conserve de l'église que la nef et la façade occidentale. La façade est avec l'arc triomphal supporté par des colonnes, chapiteaux et surmonté d'une ouverture à petites arcades hautes appartient à un état originel. Les arcades donnant accès aux bas-côtés sont bouchées depuis le XVIII^{e} siècle. L'édifice actuel repose sur une construction antique datant des I^{er} et II^{e} siècle apparue lors des fouilles de 1976. Un mur nord-sud est venu au III^{e} siècle compléter ce premier état de construction. Dès la fin du III^{e} siècle ou au début du IV^{e} siècle une construction occidentale rectangulaire a été ajoutée pouvant être un grand mausolée funéraire. Une première abside a pu exister dès le V^{e} siècle. Plusieurs modifications ont lieu entre le VI^{e} et le IX^{e} siècle alors que se développe une occupation funéraire.
Dimensions : Longueur générale actuelle : 17 m ; largeur : 8,50 m

Dans l'attribut considérationsCritiques de la classe EGE
8. CONSIDÉRATIONS CRITIQUES SUR LES PHASES ET SUR LA CHRONOLOGIE
Deux phases de constructions se sont déroulées pendant l'antiquité, avec une période de destruction à la fin du III^{e} siècle (Stratigraphie et mobilier). Le deuxième état correspond à une phase immédiatement postérieure, dans les premières années du IV^{e} siècle avec une réutilisation

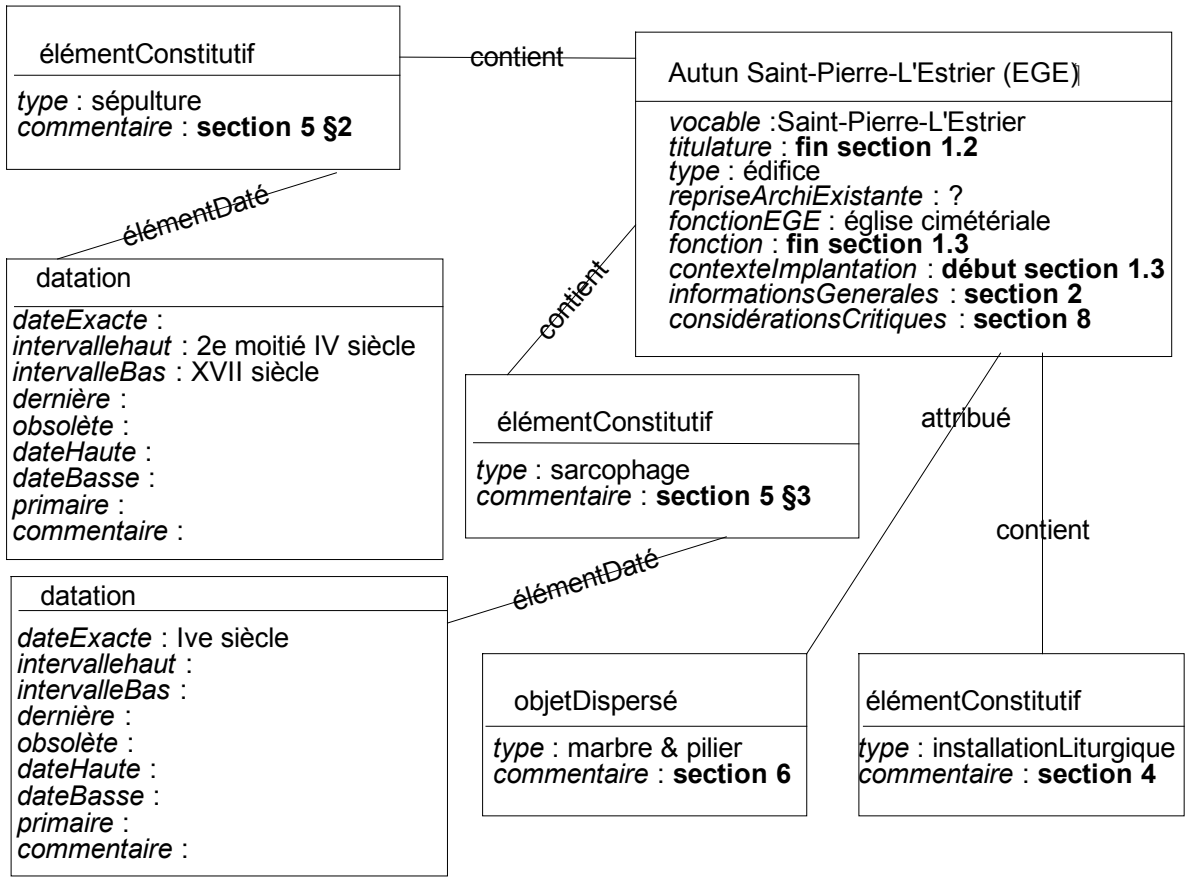


Figure A.3 – Suite de la fiche Saint-Pierre-L'Estrier

du bâtiment du III^e siècle (Chronologie relative et mobilier). Il correspond à la construction d'une double chambre rectangulaire occidentale, interprété comme un mausolée semi-hypogée (Typologie). La présence d'inhumations au nord-est et à l'est incline à croire à l'établissement dès le IV^e siècle d'un culte chrétien attesté également par la proximité des mausolées des premiers évêques d'Autun. Dans un deuxième temps, le plan de l'église se précise avec l'adjonction d'un bas-côté nord ou un portique et la construction d'une abside. Au troisième état correspond la reprise de façade, la construction de l'abside et par la suite l'établissement d'une église à trois nefs à une date entre la fin du VI^e ou au plus tard le VIII^e siècle. Le quatrième état est celui de la reconstruction située au IX^e-X^e siècle de la nef (Trace de taille, mise en œuvre, chronologie relative). Le cinquième est situé au début du XI^e siècle avec une reprise de la nef et la constitution d'un transept et les éléments sculptés (Datation par le type de chapiteaux élaborés autour de 1000-1020, la mise en œuvre, et les sources textuelles avec l'intervention du roi Robert Le Pieux).
La catégorisation de ces éléments est présentée en figure A.2.

Dans une classe élémentConstitutif sans datation (voir figure A.3)
4. INSTALLATIONS LITURGIQUES
À la limite du collatéral nord et de la nef centrale, une grosse pierre présentant une mortaise témoigne d'un tracé de clôture.

Dans une classe élémentConstitutif avec datation
La datation correspond à l'extrait de texte en bleu. Les éléments en vert dans le texte ne sont pas représentés sur la figure A.3.
5. SÉPULTURES
5.1 Emplacement et relation avec l'édifice :
Les sépultures sont présentes aussi bien à l'est et à l'ouest marquant une pérennité de l'usage funéraire du lieu. Dans le secteur fouillé nord-est, sur des fosses correspondant peut-être à la fondation est du transept, trois groupes d'inhumations ont été identifiés entre Antiquité tardive (seconde moitié du IV^{e} s.) et l'époque moderne (XVII^{e} s.) .
Plusieurs sarcophages et fragments de sarcophages du IV^{e} siècle ont été retrouvés en relation avec l'édifice. Dans l'angle Nord-Ouest, les sépultures sont proches de l'édifice. Dans des couches remblais du matériel a été découvert : colliers, anneau, bracelet et monnaies. Ces dernières de l'époque de Magnence appartiennent à différents ateliers, tout comme la céramique et confirment la présence d'une occupation continue des inhumations même après une période de destruction d'éléments d'une première occupation. Des sépultures ont été découvertes dans la zone du collatéral nord, lors des fouilles à cet emplacement dont un petit sarcophage d'enfant en plomb avec des croix gravées non décoratives (IV^{e}s. ?). D'autres sépultures d'enfants semblent avoir été alignées le long du bas-côté nord. Dans l'angle intérieur nord-ouest de la façade, un massif de maçonnerie peut constituer le soubassement d'une tombe à arcosolium. Des sarcophages ont été retrouvés dans le collatéral nord et à l'ouest ont été découverts en 1985 sept sarcophages dont cinq avaient encore leur couvercle. Le type et le mobilier permettent de les situer à la fin du VI^{e} et au début du VII^{e} siècle.

Dans une classe objetDispersé
6. OBJETS DISPERSÉS NON RATTACHABLES À L'ARCHITECTURE DE L'ÉGLISE
Des fragments de marbre ont été découverts hors contexte composés de : plaques, bases, moulures antiques. Des fragments de pierre ou de marbres proviennent de chapiteaux et de colonnes ou tailloirs. Un fragment de pilier a été découvert réemployé. Il pourrait provenir d'un pilier d'extrémité de plaque de chancel par ses dimensions et sa feuillure d'encastrement.

La description des états et leur catégorisation sont données ci-dessous.

Pour l'état I - Section 3. Articulation en état - État I (voir figure A.4)
(a) : Deux phases antiques de constructions et une période de destruction à la fin du III^{e} siècle. Deux murs à ouest épais respectivement de 0,55 m et 0,90 m se prolongeant sur plus de 20 m de l'angle sud de la nef au-delà du bas-côté nord avec une ouverture dans l'un des murs de 2 m Ces deux murs appartiennent à la première phase de construction datable du I^{er}-II^{e} siècle. Dans la seconde phase de construction, des fragments de murs sont visibles dans le mur du collatéral nord, dans l'alignement est observable le négatif du mur conservé dans la cave. Un second mur orienté est-ouest est à noter de même qu'à l'est, un mur de retour rejoint l'alignement du mur nord. Ces éléments sont à dater de la première moitié du III^{e} siècle. Enfin, à l'est subsiste les restes d'un mur sur lequel s'appuiera l'abside.
(b) II. Maçonnerie : Des sols de tuileaux on été découverts.
(b) VI. Décor appliqué aux murs et maçonneries : Lors des fouilles côté est et dans l'espace détruit qui a servi de sacristie, des enduits colorés ont été découverts. Les plus gros fragments appartenaient à des angles. Leur technique de réalisation et les couleurs des motifs ont permis de les rapprocher des techniques antiques de réalisation.

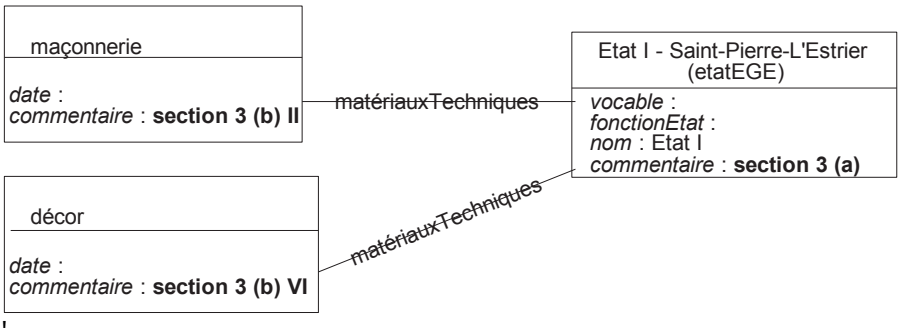


Figure A.4 – État I de Saint-Pierre-L'Estrier

Pour l'état II - Section 3. Articulation en état - État II (voir figure A.5)
(a) : À l'ouest en façade, un hypogée était semi-enterré, sur des fondations de 1,20 m, il mesurait 4 m x 8 m intérieur et été divisé en deux chambres funéraires voûtées de 3 x 4 m, selon toute probabilité à usage funéraire, un reste de sarcophage a été retrouvé et plusieurs sépultures en pleine terre plus tardives. À l'extrémité nord-ouest du collatéral nord ou du portique, le "mausolée" ou hypogée rejoint le mur antique le plus épais et semble avoir été repris en élévation et utilisé pour constituer la façade ouest.
(b) II Maçonnerie :
Les murs sont composés de moellons rectangulaires non taillés, posés en assises horizontales avec un mortier à forte densité de tuileau très résistant.

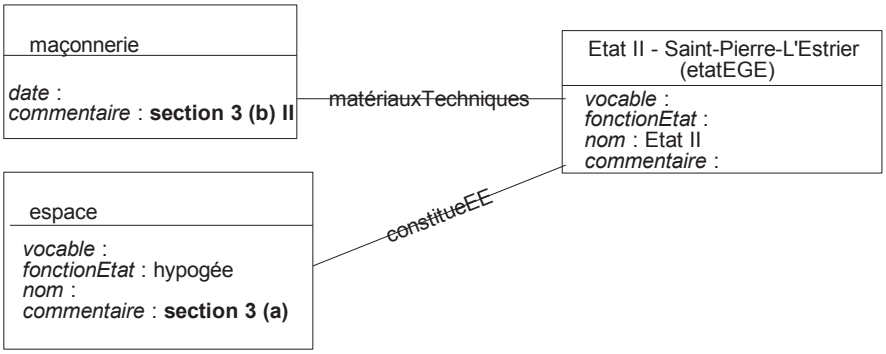


Figure A.5 – État II de Saint-Pierre-L'Estrier

Pour l'état III - Section 3. Articulation en état - État III (voir figure A.6)
(a) : L'église est formée d'un plan basilical à trois vaisseaux réutilisant les maçonneries antiques du 1er état avec l'ajout d'une abside unique. La longueur de l'édifice de l'abside reconnue en fouille jusqu'au parement extérieur présumé du mur ouest du "mausolée" occidental est de 30 m. La largeur supposée à la naissance de l'abside avoisine 18 m. L'abside est circulaire à l'intérieur et à l'extérieur. Elle possédait une ouverture intérieure de 5 m pour une profondeur de 3,50 m ; les fondations sont épaisses de 0,90 m. La nef mesure 6,40 à 6,60 m de large et possède des collatéraux très larges puisque l'on retrouve les 4 m entre les piliers dans la largeur du collatéral. À l'ouest, un mur a été identifié plus à l'ouest des deux premiers (murs antiques) pouvant correspondre à la façade du collatéral nord et indiquant une reprise dans la disposition du massif occidental. Il devait y avoir deux niveaux d'utilisation.
(b) II Murs et maçonneries Dans cette phase de construction un mortier gris a été utilisé.

FIGURE A.6 – État III de Saint-Pierre-L'Estrier

Pour l'état IV - Section 3. Articulation en état - État IV (voir figure A.7)
Cette partie du texte introduit des éléments datants (traces de taille et tailloirs) extraits du texte en bleu ci-dessous.
(a) : La modification du plan de la façade et la reprise de la nef semble prendre sa forme définitive. Il s'avère nécessaire de reprendre les arcades donnant sur les bas-côtés. Ainsi, l'arcade sud-ouest est reprise plus bas suivant un profil outrepassé avec au-dessus une ouverture encore visible. Cette dernière devait donner accès à une tribune en bois. Dans le collatéral nord, une reconstruction est observable grâce à l'utilisation d'un mortier gris. Tout le mur nord du bas-côté a été aussi repris et élargi de 20 cm. Aux deux tiers de la longueur, on constate dans cette phase l'existence d'un retour vers le sud. Cela peut être l'amorce d'un transept.
(b) II Murs et maçonneries Dans cette phase on peut signaler la présence d'un mortier rose de tuileau. Les traces de tailles visibles sur les claveaux se rapprochent de celle de Saint-Germain d'Auxerre au milieu du IX^{e} siècle et perdurant jusqu'au début du XI^{e} siècle : peut-être préciser. Le profil des tailloirs appartient à des formes dérivées de l'antique et utilisées au milieu du IX^{e} siècle.

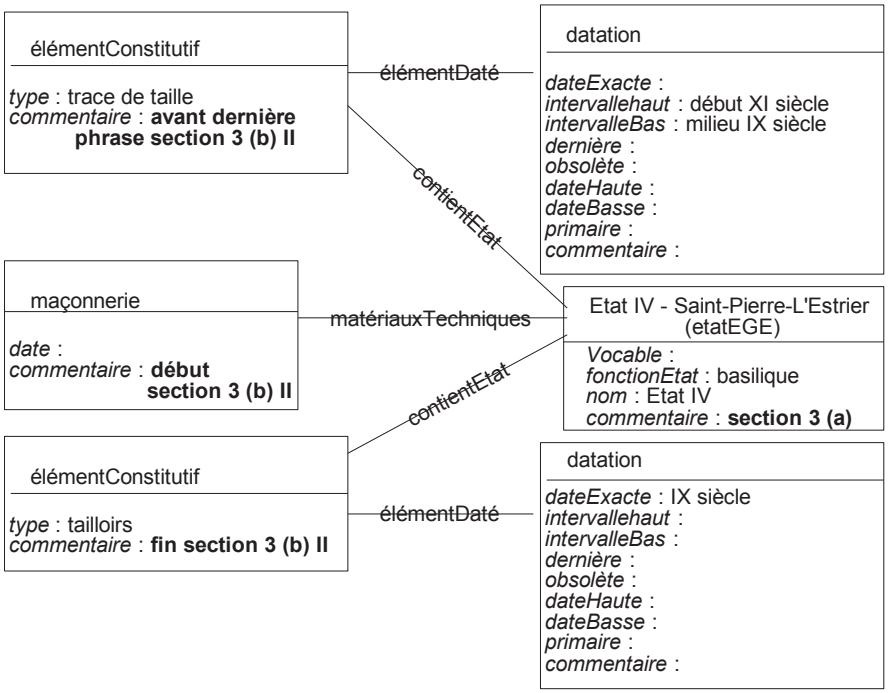


FIGURE A.7 – État IV de Saint-Pierre-L'Estrier

Pour l'état V - Section 3. Articulation en état - État V (voir figure A.8)
Pour cet état un espace est introduit (pour le chœur) qui correspond à l'extrait de texte en bleu ci-dessous.
(a) Le plan de l'édifice se présente comme un plan à croisée régulière.
Le plan du chœur trop étroit a été modifié à partir du premier arc triomphal pour être agrandi vers l'est par de nouvelles arcades. Puis par une croisée de transept et probablement une nouvelle abside. À la nef, deux travées ont été ajoutées donnant une longueur entre 6 m et 6 m 50 de large. Les collatéraux ont été à plusieurs reprises transformés, leur largeur définitive était de 3 m à 4 m 50. Le transept a été modifié lors de la construction d'un nouvel arc-triomphal. Un pilier pouvant former un carré avec le grand arc, donc une croisée de transept a été observé lors des fouilles du secteur nord est.
(b) II Murs et maçonneries : Un mortier ocre-jaune a été observé pour les structures en élévation et dans la structure du pilier trouvé en fouille pouvant former avec l'arc un carré donc une croisée du transept. On note l'alternance de brique et de pierre On trouve l'utilisation de plomb dans la mise en œuvre des colonnes et chapiteaux.
(b) VI. Décor appliqué aux murs et maçonneries : Les chapiteaux de la façade orientale sont d'une grande qualité d'exécution. Ils reposent, ainsi que les colonnes, sur d'épaisses feuilles de plomb nécessaire pour la stabilité et pour éviter les remontées d'humidité. Ils ont été réalisés par un atelier que l'on retrouve avec les chapiteaux des arcades supérieurs et avec ceux déposés aujourd'hui au musée Rolin d'Autun. Ils sont tous sans astragale composés d'une rangée de feuilles d'eau ou plus simplement épannelés, avec une échancrure qui les sépare sans descendre jusqu'à leur base. Deux feuilles formant volute se superposent aux angles. Au centre de chaque face, un motif est surmonté d'un dé marqué par la présence au centre d'une fleur à bouton et pétales ou par un disque en hélice. L'abaque est souligné par trois baguettes parallèles et striées par alternance.

Pour l'état VI - Section 3. Articulation en état - État VI (voir figure A.9)
Pour cet état un espace est introduit (pour la cave) qui correspond à l'extrait de texte en bleu ci-dessous.
(a) L'état actuel de l'édifice est constitué des quatre murs de la nef, transformé au XIX^{e} siècle en grange ; c'est un plan rectangulaire d'environ 16 m sur 6 à 6,40 m. Dans les murs latéraux, les arcades ouvrant sur les bas-côtés détruits sont encore visibles mais bouchées depuis le XVIII^{e} siècle. Une cave a été aménagée par l'agriculteur propriétaire du lieu au XIX^{e} siècle supprimant tous les niveaux antérieurs de construction. Sur le mur est, la grande arcade à double rangée de larges claveaux qui devait faire office d'arc triomphal est aujourd'hui obturée par une cloison de briques. Il est supporté par deux colonnes antiques de 2,50 m réemployées à chapiteaux. Une porte a été aménagée ainsi qu'une fenêtre dans la partie inférieure bouchée. Au-dessus de l'arc, une suite de cinq arcades a été conservée mais obturée. La toiture devait être plus haute.
(b) II Matériaux et techniques de construction : Pour la construction de l'édifice sont employés des grès, granit et calcaire de deux types. Les calcaires tendres de couleur blanche. On les trouve dans la nef et les baies ouvrant sur les collatéraux, ainsi que dans la sculpture des chapiteaux. Pour les claveaux des arcades hautes, une alternance entre le calcaire blanc s'effectue avec des briques moulées en forme de claveaux. Pour les mortiers, l'utilisation de tuileau est relativement importante notamment pour les arcades de la nef. Ce n'est pas le cas pour la reconstruction du chœur. Dans le chevet, entre bases et colonnes et colonnes et chapiteaux des plaques de plomb ont permis d'assurer à l'ensemble une meilleure répartition des charges.

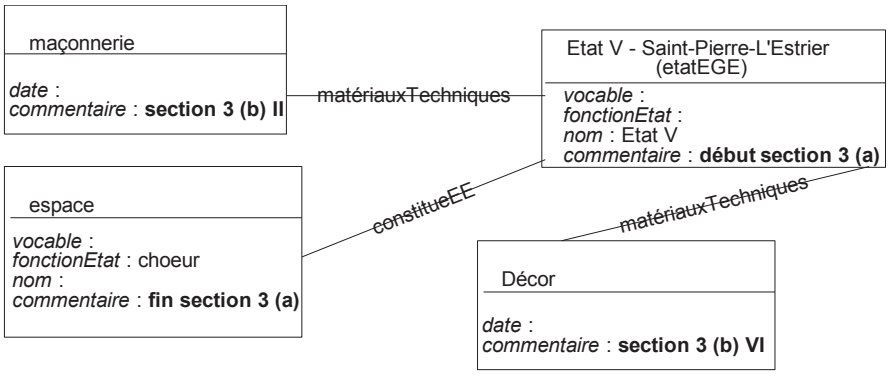


FIGURE A.8 – État V de Saint-Pierre-L'Estrier

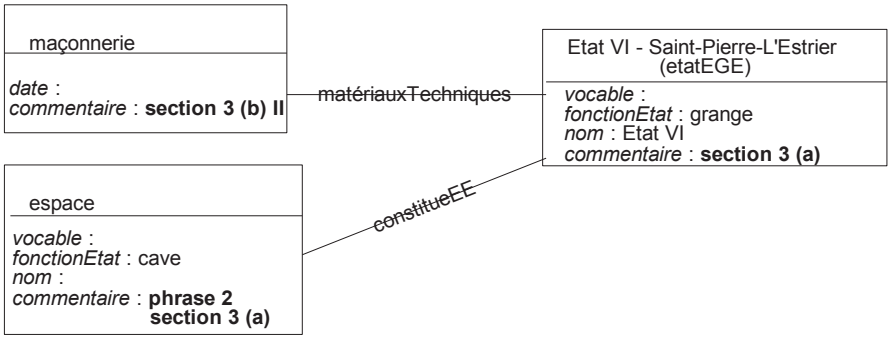


FIGURE A.9 – État VI de Saint-Pierre-L'Estrier

3 Modélisation du projet CARE à l'aide d'un diagramme de classes UML exécutable

À partir de la modélisation conceptuelle donnée en figure A.1, nous avons élaboré le diagramme de classes UML donné en figure A.10. Dans ce diagramme, nous avons :

1. explicité les concepts de groupe d'édifices, d'édifice, d'espace et d'élément constitutif ce qui nous a permis de supprimer les associations récursives et les relations d'héritage ;
2. utilisé *executable* UML [MB02] permettant ainsi une traduction automatique vers le modèle relationnel (par exemple).

Ce modèle exécutable comporte plus d'une trentaine de classes, plus de cinquante associations pour une modélisation partielle qui s'est révélée difficile à stabiliser après trois mois de travail avec les archéologues.

4 Ontologies dans le domaine du patrimoine

Il existe un grand nombre de vocabulaires contrôlés pour décrire et indexer des objets du patrimoine culturel, nous les présentons dans le prochain paragraphe. Nous présenterons ensuite des ontologies propres au domaine du patrimoine culturel dont CIDOC-CRM qui est l'ontologie de référence.

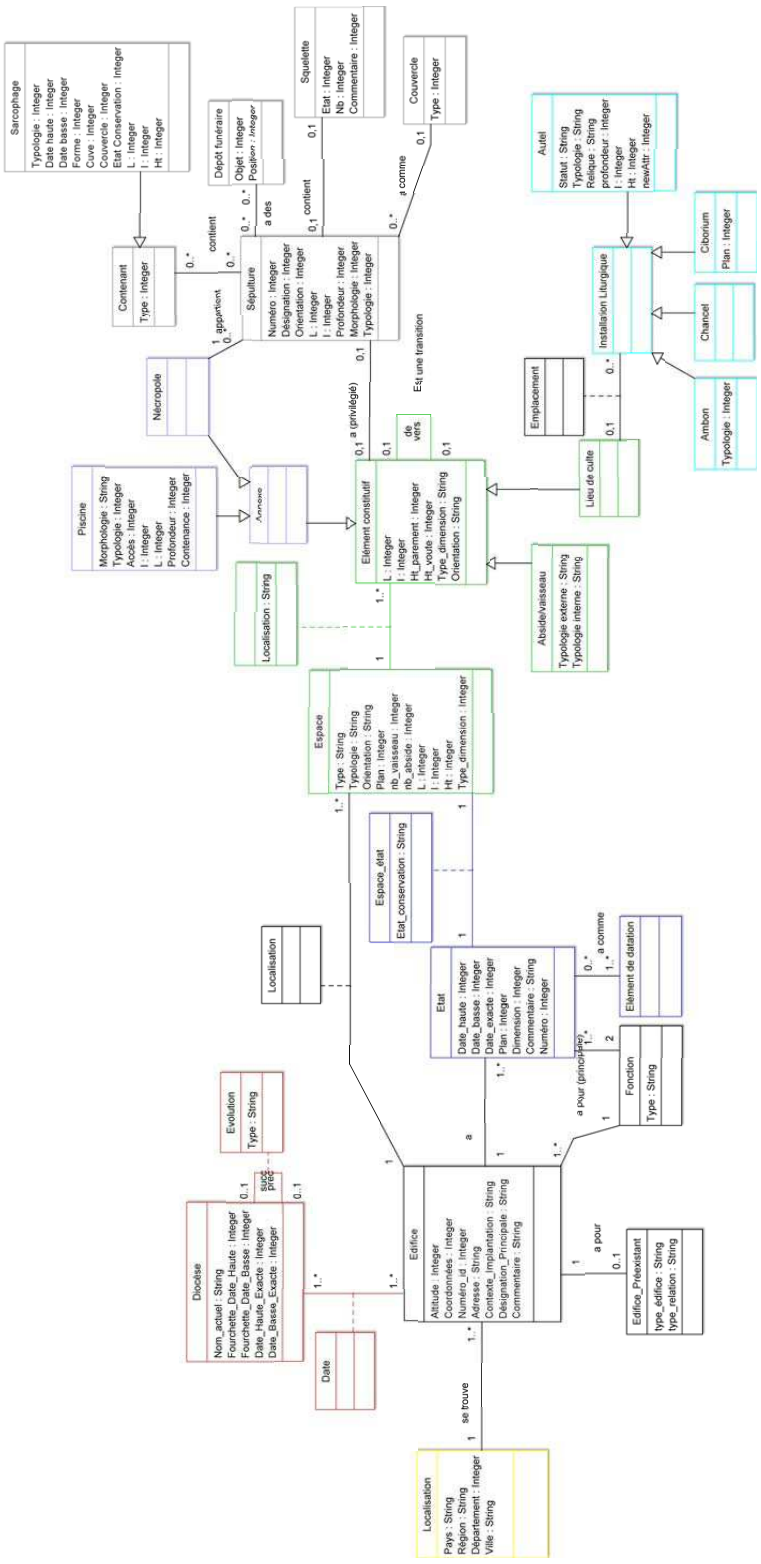


Figure A.10 – Diagramme de classes UML exécutable du projet CARE

4.1 Vocabulaires contrôlés dans le domaine du patrimoine culturel

La page sur les normes de vocabulaire pour la description du patrimoine d'un site canadien [1] répertorie une vingtaine de vocabulaires contrôlés, classés selon leur contenu et leur organisation. Dans la catégorie sur laquelle nous travaillons, archéologie et architecture, seule la base de données *Thésaurus* est en français http://www.culture.gouv.fr/culture/inventai/patrimoine/. Ce thésaurus d'architecture a été élaboré par la Direction de l'architecture et du patrimoine du ministère français de la Culture et de la Communication. Il contient 1 135 termes décrivant des œuvres architecturales, leurs définitions et des notes sur leurs usages et leurs équivalents en anglais et en italien. Ce thésaurus sert à indexer le contenu de *Mérimée*, la base de données nationale du patrimoine de France. Son interface n'est disponible qu'en français. Les figures A.11(a) à A.11(c) montrent les parties qui nous intéressent : "objets religieux", "meubles religieux" et "architecture religieuse".

4.2 Thésaurus PACTOLS

PACTOLS, acronyme de "Peuples et cultures, Anthroponymes, Chronologie relative, Toponymes, Oeuvres, Lieux et Sujets", est un thésaurus géré par le programme FRANTIQ [2]. Les termes ont été sélectionnés et organisés selon des relations de synonymie, d'antonymie, d'adjacence et/ou hiérarchiques. Il se compose de six micro-thésaurus :

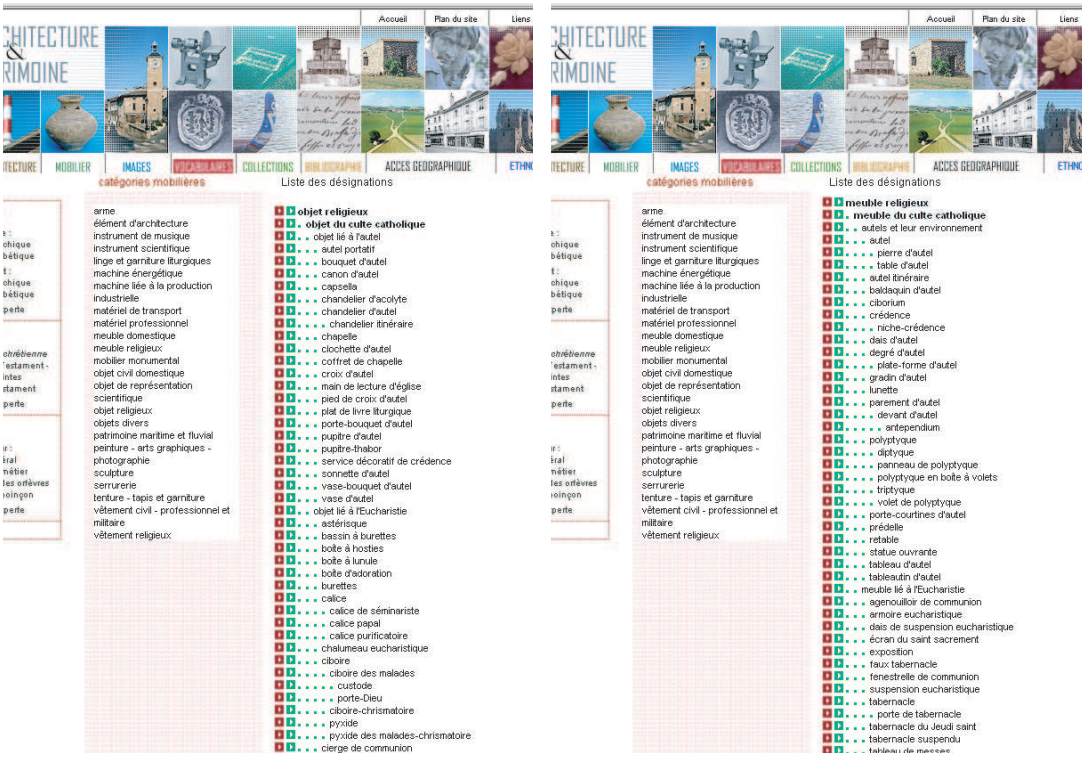
1. "Peuples" regroupe tous les noms des entités culturelles reconnues en Préhistoire et Protohistoire, les peuples antiques et actuels et aussi les noms d'habitants d'une cité mentionnée dans un document historique ;
2. "Anthroponymes" regroupe tous les noms de personnes (dieux, héros, écrivains, etc.) cités dans un document ;
3. "Chronologie" regroupe les termes de chronologie relative et les ères géologiques. La chronologie absolue n'est pas traitée ;
4. "Oeuvres" regroupe des œuvres artistiques et littéraires qu'elles soient religieuses, juridiques ou poétiques ;
5. "Lieux" regroupe tous les noms de lieux hiérarchisés (du continent à la commune), la géographie physique et hydrographique des continents ;
6. "Sujets", détaillé en vingt-six domaines, regroupe les domaines d'études sur l'Antiquité même si pour les archéologues les domaines droit, paléographie, philosophie et politique peuvent être d'un intérêt moindre. On retrouve sous le domaine "architecture/édifice" les édifices religieux mais ces termes sont trop imprécis pour le projet CARE. La hiérarchie du micro-thésaurus "Sujets" est présentée dans la partie gauche de la figure A.12 et le détail des termes pour les édifices religieux en partie droite.

PACTOLS est disponible à l'adresse http://frantiq.mom.fr/fr. Il a fait l'objet d'une transformation en ontologie par Marchaix en 2008 [Mar08].

4.3 Ontologie dans le projet Arkeotek

Le projet Arkeotek (www.arkeotek.org) a produit un prototype de recherche d'information au sein de collections documentaires en sciences humaines. Ce prototype s'appuie sur une ontologie légère comportant deux cents concepts décrivant les principales notions de l'archéologie des techniques [AG06]. Cette ontologie comporte deux hiérarchies de concepts :

1. http://www.pro.rcip-chin.gc.ca/normes-standards/vocabulaire_vocabulaires-vocabulary_vocabulary-fra.jsp
2. FRANTIQ – Fédération et Ressources pour l'antiquité – regroupe des centres de recherche CNRS, des Universités et le ministère de la Culture, son objectif est de mettre en commun des bases de données sur les sciences de l'Antiquité c'est-à-dire de la préhistoire jusqu'au Moyen Age.



(a) Objets religieux (b) Meubles religieux

(c) Architecture religieuse

Figure A.11 – Base de données *Thésaurus*

– une hiérarchie relative aux paramètres décrivant les objets archéologiques : nature, usage et matériaux qui les composent ;
– une hiérarchie relative aux paramètres d'une étude archéologique : méthodes et techniques mises en œuvre, leurs buts et leurs résultats.

Cette ontologie utilise l'aspect temporel de l'ontologie CIDOC-CRM.

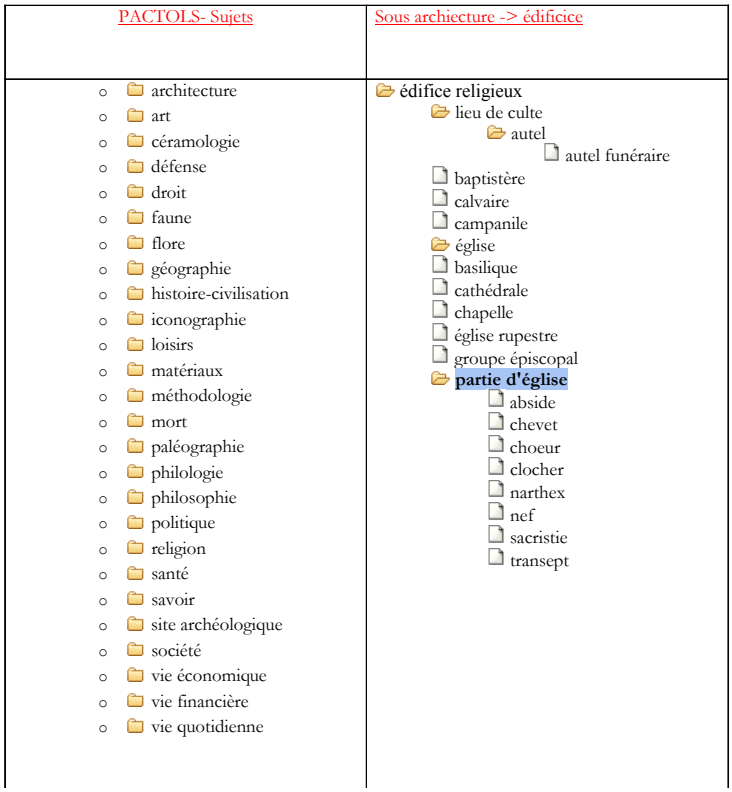


FIGURE A.12 – Extrait de PACTOLS (micro-thésaurus "Sujets")

4.4 Ontologie CRM et projet CIDOC

Le CIDOC (Comité International pour la Documentation) soutenu par l'ICOM (International Council of Museums) [CID94] a pour objectif d'améliorer la gestion des collections, des archives et des produits scientifiques ou administratifs liés au patrimoine artistique et culturel. L'idée qui préside est la création d'un modèle de données standard pour décrire des objets de musées et des informations culturelles (comme des œuvres d'art), des vestiges archéologiques (comme une céramique) et tout ce qu'il est par nature impossible de conserver à l'intérieur d'un musée (comme un monument ou une grotte préhistorique). Ce modèle doit être exploitable par des systèmes informatiques et être suffisamment riche pour rendre compte de la diversité des analyses et des interprétations. Un modèle appelé CRM (Conceptual Reference Model) a donc été élaboré depuis 1994, il a été publié en 2006 par l'ISO en tant que norme internationale (ISO 21127 :2006) (http://www.cidoc-crm.org) [CC02].
Au centre du modèle CRM, comme le montre la figure A.13, se trouve la notion d'*événement*. Intuitivement on pourrait en effet penser que ce qui prime est la description d'un objet matériel tel qu'il se présente aujourd'hui, alors qu'il s'agit avant tout de restituer cet objet dans son contexte historique. Ce qui importe à l'archéologue, c'est l'environnement de l'objet au cours du temps et l'énoncé de tout ce qui a pu lui arriver. Si le contexte joue un rôle primordial dans la constitution des connaissances, il prend une dimension particulière dans le domaine du patrimoine culturel : la signification des données perçues et des informations collectées sur un objet est fortement corrélée avec son contexte. Chaque objet trouvé et étudié en dehors de son contexte peut perdre une partie de ses attributs fonctionnels, par exemple un claquoir devient "deux planchettes reliées par une charnière" alors que replacé dans son contexte religieux, c'est

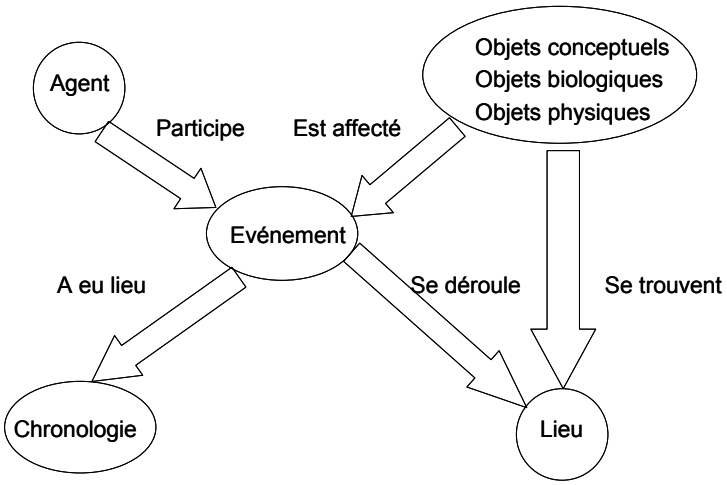


Figure A.13 – Les composants et leurs relations dans CIDOC-CRM (d'après [Doe03])

un instrument servant à coordonner les mouvements des enfants de chœur. Il faut alors relier les connaissances liées au contexte, celles associées à d'autres objets trouvés dans le même contexte et celles sur l'objet étudié.
Le cœur du modèle est constitué des *Événements*, qui expriment soit l'état de l'objet (classe *État de Conservation*), soit un événement (un *Événement*, une *Activité*, un *Début d'Existence* ou une *Fin d'Existence*, etc.). Cette notion centrale est complétée par le moment de l'événement grâce à la classe *Tranche Chronologique*, le *lieu*, qui est à l'origine de l'événement grâce à la classe *Agent* et l'objet décrit. CRM propose les notions de *Quelque Chose de Matériel*, qui peut être naturel ou fabriqué par l'homme et d'*Objet Conceptuel*.
CRM sert de référence pour établir d'autres ontologies. Par exemple, Navigli et al. [NV06] ont automatiquement traduit sous forme d'ontologie le *Art and Architecture Thesaurus*[3]. Le haut de l'ontologie est constitué de CIDOC-CRM, les concepts plus spécifiques correspondent aux termes du thésaurus. Le processus de transformation en ontologie a utilisé la structure du thésaurus et une analyse linguistique des définitions des termes pour identifier des relations lexicales et obtenir au final des relations entre concepts.

Une autre ontologie dans le domaine de l'archéologie a été proposée par Whitlow [Whi]. Elle se base sur l'ontologie de haut niveau BFO[4] qui pose les bases de descriptions spatio-temporelles nécessaires à la description des entités archéologiques.

5 Ontologie pour le corpus CARE

L'ontologie CARE est une spécialisation de CRM prise comme norme. Notre spécialisation, développée avec Protégé[5], comporte 124 classes et 715 individus (en janvier 2012).

3. Le *Art and Architecture Thesaurus* est un thésaurus sur le patrimoine culturel développé par le Getty Information Institute.

4. Basic Formal Ontology : http://www.infomis.org/bfo

5. Protégé est un éditeur d'ontologie développé par l'Université de Stanford. Des plugins notamment pour gérer les représentations sous forme graphique, OWLViz dans notre cas, peuvent être insérés. En quelques années, cet éditeur s'est imposé comme la référence, il gère des langages standards tels que RDF et OWL. Il est également possible de faire fonctionner des raisonneurs, comme Racer (Renamed ABox and Concept Expression Reasonner), pour vérifier la cohérence et la consistance de la structure ontologique.

5.1 Méthodologie

Notre matériau de base est constituée de phrases en langage naturel, produites par des archéologues, pour décrire les différents éléments d'un édifice. L'annexe B présente quelques uns de ces concepts. Nous avons traduit ces descriptions dans un fichier excel (voir annexe B), puis nous avons travaillé sur une représentation formelle des connaissances des éléments d'architecture en

1. classifiant les éléments architecturaux d'un édifice : éléments maçonnées (châteaux, fûts, etc.), charpentes, sols, etc. ;
2. définissant les relations entre les éléments. Goulette [Gou99] a énuméré les différentes relations entre éléments : relation partie-tout, relations spatiales, relations de composition. Nous avons aussi intégré les relations qu'un élément d'architecture peut avoir avec une technique de construction, un élément stylistique, etc. ;
3. établissant la correspondance entre les éléments d'architecture et le domaine religieux. Nous avons pour cela utilisé la représentation symbolique que fournit un élément comme une installation liturgique ou un espace identifié de l'édifice. La nef par exemple est l'endroit où sont rassemblés les fidèles, le chœur est l'espace où se trouve l'autel et où se déroule la liturgie ;

et finalement tenu compte de la dimension temporelle (voir figure A.14).

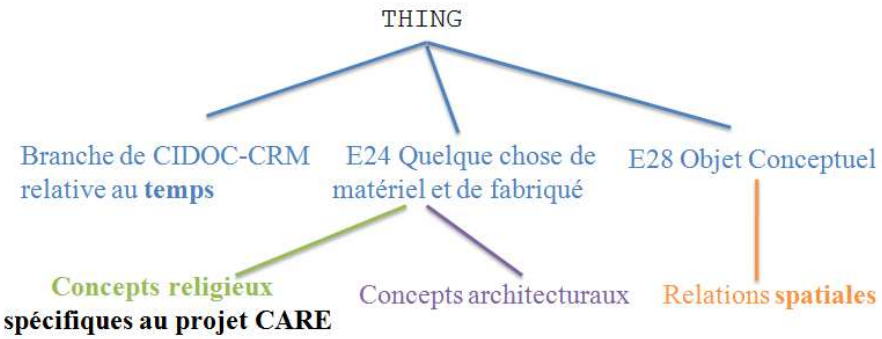


FIGURE A.14 – Grandes branches de l'ontologie CARE

Nous avons ainsi dégagé un ensemble de concepts connectés à des termes architecturaux, organisés par des relations méronomiques (décomposition morphologique) et représentationnelles (dimensions, matériaux, relations spatiales). La haute technicité des descriptions fournies (nom et organisation des parties d'un édifice, techniques de construction, éléments de décoration, etc.) a permis de créer une ontologie d'application. Les concepts ont ensuite été rapprochés de l'ontologie CIDOC-CRM qui apporte aussi la modélisation temporelle sous la forme d'états.

5.2 Modélisation des concepts religieux

Un des concepts religieux de CARE est l'édifice, représenté par le concept `Batiment`. Un édifice est décrit par les éléments suivants : la topographie (latitude, longitude, adresse, diocèse), un ou des usages religieux (fonction), un ou des plans, un type de bâtiment, des documents associés. Le concept de `Batiment` est équivalent au concept *Objet fabriqué E22* de CIDOC-CRM. En effet, CIDOC-CRM définit un objet fabriqué comme « *un objet bien délimité, réel, d'ordre matériel et résultat d'actions d'ordre technique* ». Nous pouvons rapprocher ces éléments du modèle

d'infobox[6] développé pour les édifices religieux. La figure A.15 reprend les concepts issus de CIDOC-CRM (dont le label se termine par EXX) et leurs spécialisations pour traiter l'édifice en lui-même.

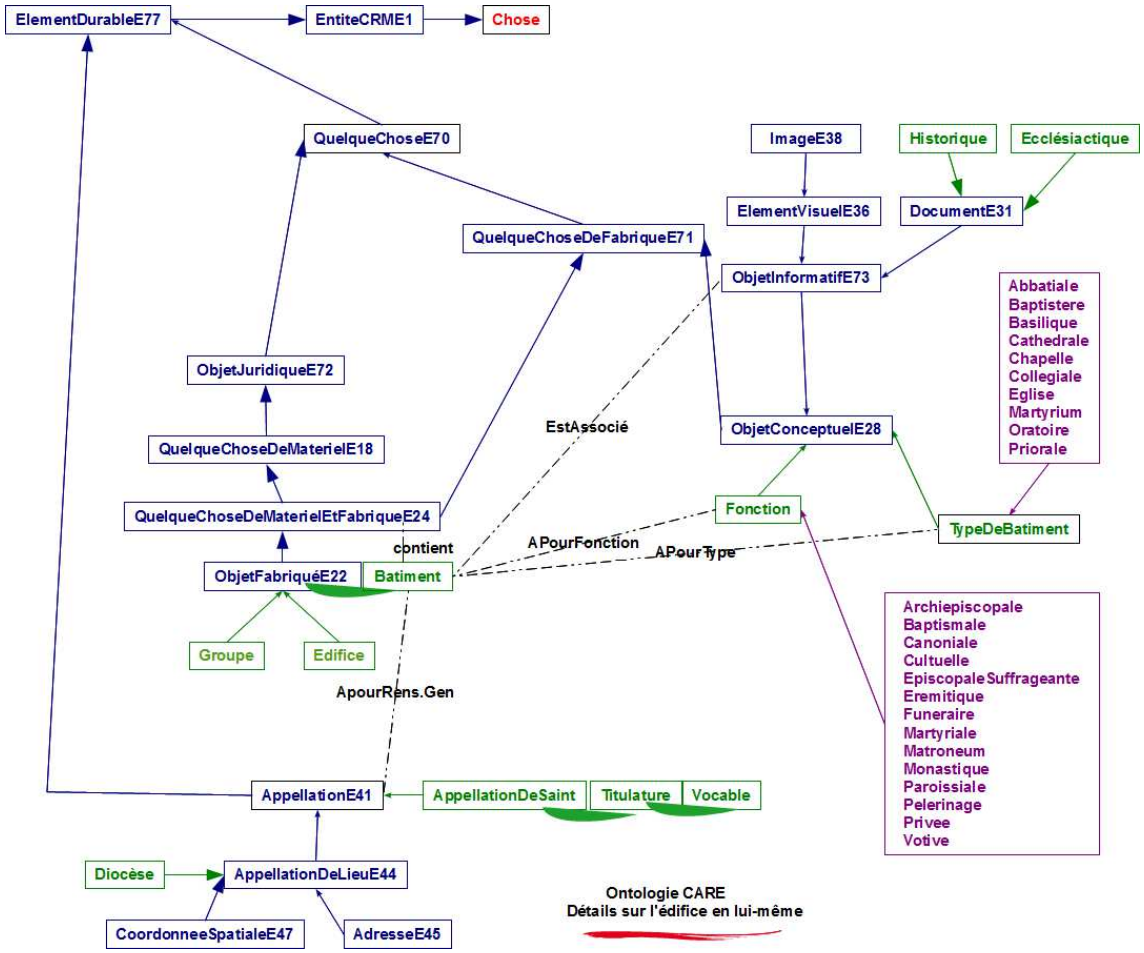


Figure A.15 – Partie d'ontologie relative à l'édifice en lui-même dans le projet CARE - En bleu les concepts issus de CIDOC-CRM, en vert les concepts propres à CARE, en rose les individus, en noir pointillé les propriétés, la demi-ellipse représente l'équivalence de concepts.

Le corpus mettant l'accent sur l'architecture, l'édifice est ensuite décomposé logiquement et physiquement en différents espaces, comme la nef, le transept, l'abside, le chevet, le portique, etc., représentés par le concept `Structure`. La figure A.16 montre des espaces d'une église romane : elle est en forme de croix latine, les deux bras de la croix formant le transept. Sur cette figure, ① représente le narthex, ② les collatéraux ou bas-côtés qui sont parfois doubles, ③ la travée qui est une division transversale de la nef entre deux piliers, ④ la nef, ⑤ le transept, ⑥ le chœur toujours "orienté" c'est-à-dire tourné vers l'Est ; ⑦ le déambulatoire qui prolonge les bas-côtés autour du chœur permettant de défiler devant les reliques dans les églises de pèlerinage. La section 2 de l'annexe B présente un glossaire et des illustrations des différents concepts manipulés. Un espace logique a pour configuration des espaces physiques, par exemple une nef peut avoir comme configuration un vaisseau central de x travées. Les espaces physiques sont construits à partir de différents éléments architecturaux, par exemple une colonnade est

6. Encadré présentant un condensé d'informations sur un sujet, et qui se retrouve sur toutes les pages parlant d'un sujet similaire.

une file de supports. Ces concepts sont des spécialisations de *Quelque Chose de Matériel et de Fabriqué E24*, concept générique qui regroupe « *des objets et des caractéristiques fabriqués par l'homme* ». La figure A.17 reprend ces différents espaces en distinguant les espaces logiques des espaces physiques.

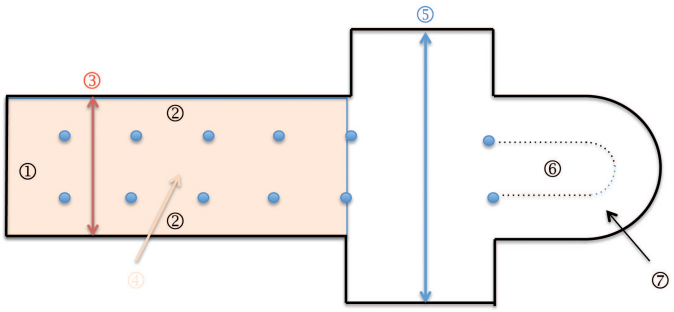


FIGURE A.16 – Espace type d'une église romane

Les installations liturgiques, comme l'autel, l'ambon, l'armoire liturgique, le ciborium, le bénitier, etc., sont représentées par le concept `InstallationLiturgique` et les sépultures sont représentées par le concept `Sepulture` (voir figure A.18).
Afin de détailler les éléments constituant un édifice, nous avons introduit le concept d'`ElementArchitectural` pour décrire par exemple les éléments maçonnés, les charpentes, les sols, les ouvertures (voir figure A.19).
Les objets dispersés sont des objets trouvés aux abords de l'édifice ou qui sont conservés dans un musée. Ce sont essentiellement du mobilier liturgique, des éléments sculptés comme un chapiteau, des fragments de décor. La figure A.20 les présente.

5.3 Modélisation des appellations

Nous devons représenter deux types d'appellations pour lesquelles nous avons utilisé le concept *Appellation E41* de CIDOC-CRM. Ce sont :
– les appellations de Saint que l'on retrouve dans la partie titulature ;
– les appellations de lieu (adresse de l'édifice) se situent dans la partie intitulée topographie de la fiche (voir figure A.21) et les appellations de diocèses. Ces deux dernières appellations sont basées sur le concept *Appellation de Lieu E44* de CIDOC-CRM.

5.4 Modélisation des connaissances spatiales

D'après Goulette [Gou99], « *la "géométrie" mise en œuvre dans les descriptions textuelles [...] est une géométrie complexe. En effet, ces descriptions ne réfèrent pas à un espace absolu et orthonormé : il s'agit plutôt d'un espace de la perception ou espace cognitif dont la structure repose en grande partie sur les aspects fonctionnels et symboliques des objets décrits, et sur le point de vue de l'archéologue. Pour décrire les éléments, ce dernier s'intéresse principalement à quatre caractéristiques : l'orientation, la délimitation, la distance et le positionnement relatif.* »
L'**orientation** représente un positionnement étudié à travers les six relations à gauche / à droite, devant / derrière et en-dessous / au-dessus. Ces relations lient un objet de référence et un objet cible. Laure Vieu [Vie97] distingue trois types d'orientation : 1) l'orientation absolue qui fait référence à un système de coordonnées externes comme les directions cardinales ; 2) l'orientation intrinsèque pour laquelle le repère est lié à l'objet de référence ; 3) l'orientation contextuelle pour laquelle le repère est lié à une entité différente de l'objet de référence. Dans

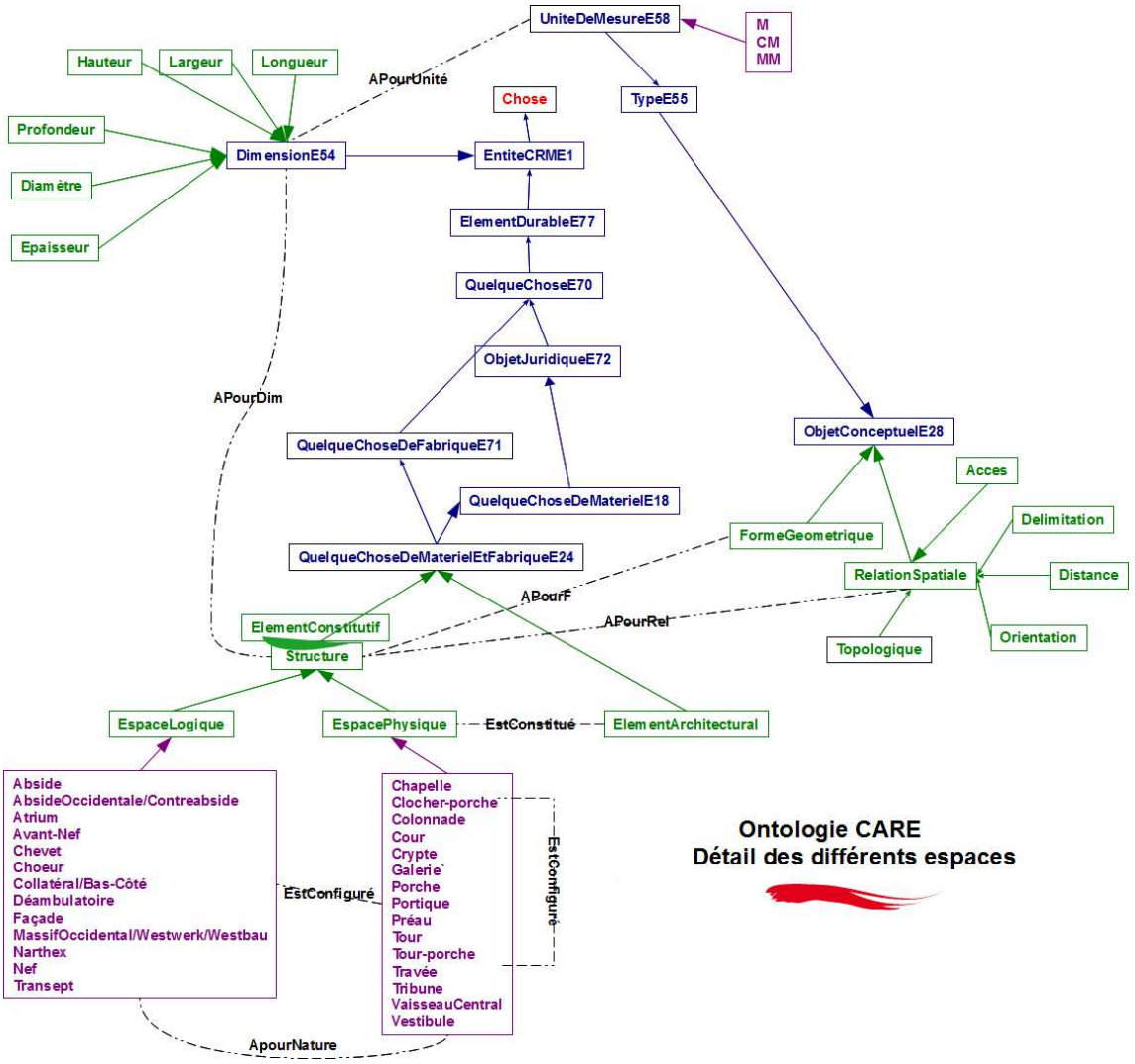


Figure A.17 – Partie d'ontologie relative aux espaces religieux dans le projet CARE

le projet CARE, l'orientation contextuelle est donnée par l'orientation du bâtiment. En effet, depuis les origines et jusqu'au XV^{e} siècle, dans tous les pays chrétiens, une église est orientée vers l'est.

La **délimitation** permet de définir les frontières des éléments avec les notions d'intérieur et d'extérieur.

La **distance** représente la notion de proximité ou d'éloignement entre deux éléments.

Le **positionnement relatif** permet de préciser la position d'un élément ou d'un attribut d'un élément, l'axe par exemple, relativement à un autre élément ou à un de ses attributs. Le deuxième élément devient alors le référentiel architectural qui permet de préciser la position du premier par le biais de prépositions ou de verbes de localisation. En français, nous trouvons des prépositions comme "sur, dans, sous, devant, derrière, etc." , des verbes tels que "encadrer, être en renfoncement, flanquer, couronner, limiter par le bas, etc.". Cela forme un vocabulaire d'une grande richesse, Borillo dans [Bor98] répertorie plus de deux cents compositions de prépositions. Laure Vieu dans sa thèse [Vie91] propose une analyse des relations entre la sémantique des relations spatiales dans le langage naturel et le raisonnement en Intelligence Artificielle. Fort heureusement, des recherches ont abouti à réduire les relations topologiques entre objets à

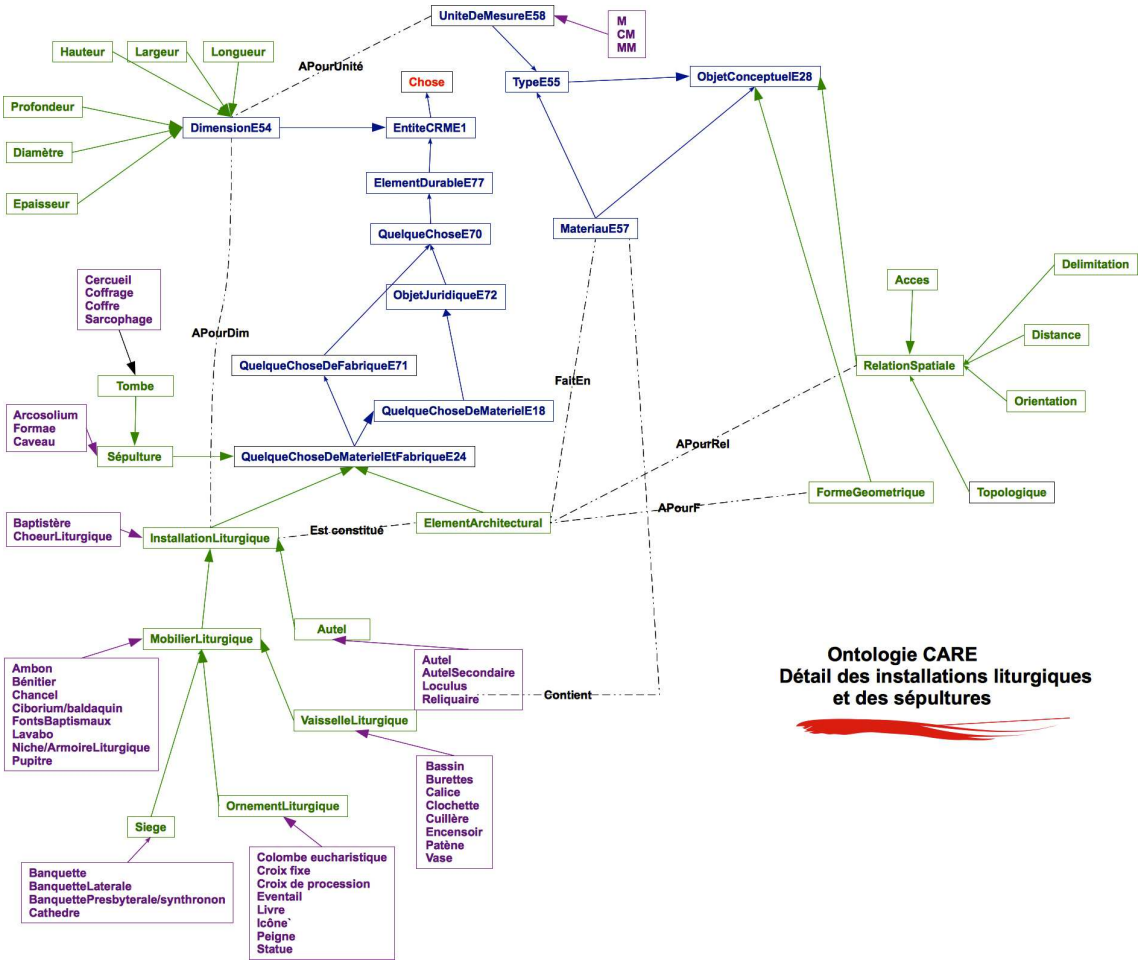


FIGURE A.18 – Partie d'ontologie relative aux installations liturgiques et aux sépultures dans le projet CARE

quelques relations considérées comme fondamentales. Egenhofer et Herring [HH91] ont défini un ensemble minimal de huit relations décrivant les relations entre deux régions. Chaque région est décrite par son intérieur, son extérieur et sa frontière. Chaque configuration est décrite par une matrice 3X3 ce qui aboutit à un modèle nommé 9-Intersection. La théorie RCC (*Region Connection Calculus*) [RCC92] propose les mêmes relations organisées sous la forme d'un treillis. La figure A.22 présente ces huit relations. Elles sont utilisées dans les ontologies pour les Systèmes d'Information Géographiques [Mir09] et ont fait l'objet de plusieurs travaux dans le Web Sémantique [HFK10].
Ces relations jouent le rôle de prédicats dans nos annotations, elles sont binaires (x à droite de y) mais la propriété "entre" est ternaire (x entre y et z). À partir de l'analyse textuelle de la description des concepts religieux, en particulier les parties concernant la position et la forme (voir annexe B), donnée par les archéologues nous avons extrait le vocabulaire correspondant aux quatre caractéristiques puis nous avons projeté ce vocabulaire sur les relations de base. Le tableau A.1 synthétise ce travail, la partie de l'ontologie résultante est présentée en figure A.23. Un travail similaire a été fait pour les accès, il s'est dégagé les propriétés "vers", "par", "à".

Vocabulaire archéologique	Relations de base
Propriétés d'orientation	
en avant, précédant	devant
en arrière	derrière
au fond	au fond
sur un côté, latéral	latéral
sous, au-dessous de, à un niveau inférieur, en bas	au-dessous de
au même niveau	au même niveau
au-dessus de, en haut	au-dessus de
à droite	à droite
à gauche	à gauche
à la corde	à la corde
au centre, centrée, axial	au centre
décentré	décentré
entre	entre
directions cardinales	N, S, E, O, NE, NO, SE, SO
propriétés de délimitation	
à l'extérieur de	à l'extérieur
à l'intérieur de	à l'intérieur
propriétés de distance	
à côté de, près de	à côté de
loin de	loin de
propriétés topologiques	
joignant, ouvrant, sur le flanc, flanquant, flanqué, attaché, accolé, lié à, au contact, adossé, enveloppant, encadré par, sur toute la longueur, inscrit dans	touche (EC)
surmontant	touche + au-dessus de
isolée	isolée

Table A.1 – Correspondance entre le vocabulaire archéologique et les relations de base

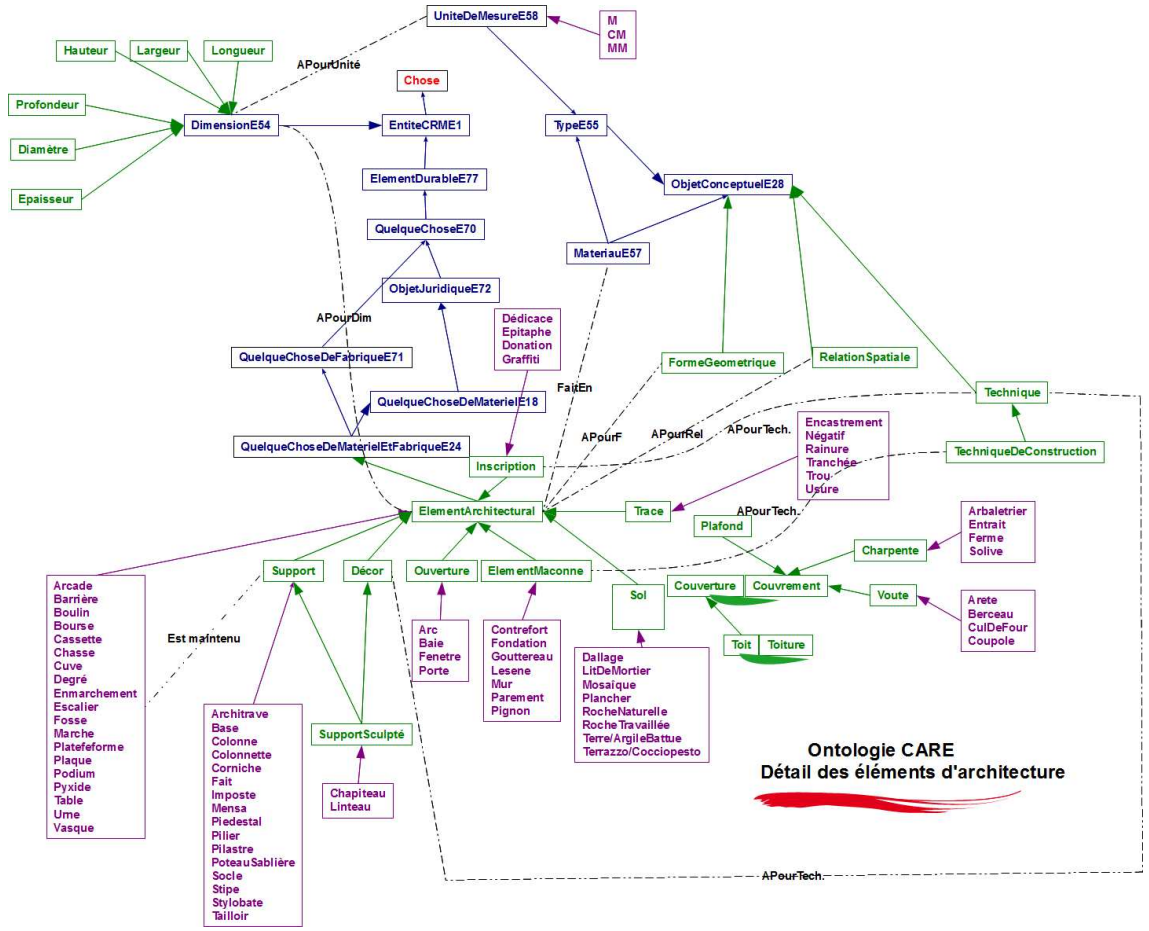


FIGURE A.19 – Partie d'ontologie relative aux éléments d'architecture dans le projet CARE

5.5 Modélisation des connaissances temporelles

Contrairement à une opinion répandue, le temps n'est pas une donnée mais le résultat de l'analyse d'indices spatiaux, stylistiques, naturels reposant sur des heuristiques.
Les archéologues ont créé des terminologies afin de classer les objets trouvés grâce à un schéma de chronologie relative, les périodes, lié aux séquences de colonies de peuplement et/ou aux séquences culturelles. L'utilisation d'une terminologie donnée repose sur un certain nombre d'opinions qui peuvent différer d'un chercheur à l'autre.

Le temps archéologique

En archéologie, le temps est donc construit à partir d'indices spatiaux (dater d'après la profondeur relative des vestiges), stylistiques (dater un objet manufacturé en se basant sur son style) ou naturels (dater par le carbone 14). Les indices sont ensuite croisés. Leur validité est sans cesse remise en question par l'émergence de nouvelles techniques.
L'archéologue manipule le temps à deux moments clefs. Tout d'abord sur le terrain avec l'analyse du site grâce par exemple à un diagramme stratigraphique [7]. Puis lorsqu'il rédige son rapport de fouille et utilise une frise chronologique qui synthétise graphiquement les résultats obtenus.

7. Le diagramme stratigraphique permet d'ordonner temporellement les différentes couches d'un site et de faire apparaître les phases de construction, d'occupation et de démolition.

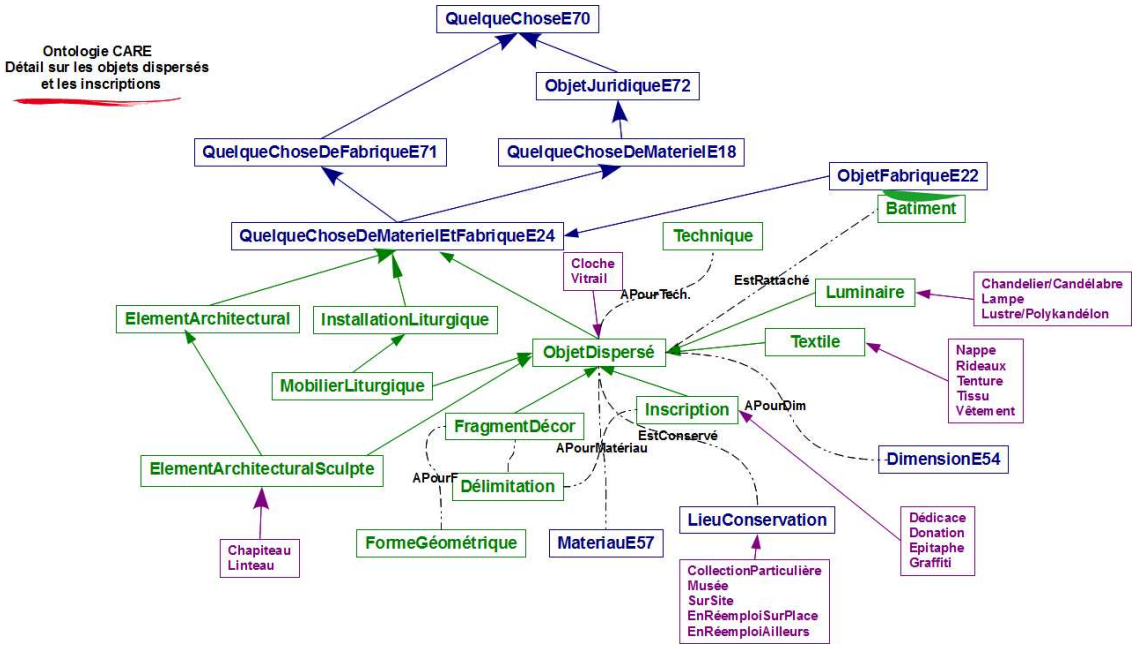


Figure A.20 – Partie d'ontologie relative aux objets dispersés dans le projet CARE

Si le temps manipulé par les archéologues et les historiens est bien en principe une fonction continue, les sources dont ils disposent permettent rarement la perception de ce continuum car tout n'est pas renseigné de manière homogène. De surcroît, la valeur de l'information archéologique est souvent relative. Comme les strates, la chronologie est souvent organisée par des relations d'antéro/postériorité, c'est-à-dire que les objets sont considérés les uns par rapport aux autres.

Doerr et al. [DPKB04] ont classé les éléments de preuve et les connaissances de base par leurs conséquences chronologiques :

– la **chronologie absolue** ne peut avoir que trois types de sources :

  1. les documents historiques datés par un calendrier connu, un événement astronomique ou tout autre événement lui-même daté dans l'absolu ;
  2. une correspondance avec un motif temporel unique dans une séquence de traces complètes et connues, telles que peuvent l'apporter la dendrochronologie ou les modèles tirés des glaces polaires ;
  3. un calcul de la distance temporelle entre aujourd'hui et l'objet à dater, comme la datation par le carbone 14, la datation au potassium-argon, la datation par thermoluminescence, le suivi des mutations de l'ADN mitochondrial, etc. Le concept *Type E55* a été spécialisé pour prendre en compte les différentes techniques de datation.

– la **chronologie relative par ordonnancement d'événements**. Une première forme de chronologie relative est donnée par une preuve directe de l'ordre temporel (séquençage) entre plusieurs événements. Elle ne peut avoir que trois types de sources :

  1. les documents d'archives qui permettent d'ordonner un événement par rapport à un autre événement, par exemple une liste de rois ;
  2. l'observation de l'ordre des traces laissées par différents événements comme la stratigraphie, des objets qui se trouvent dans un espace fermé, l'évolution des couches de glace

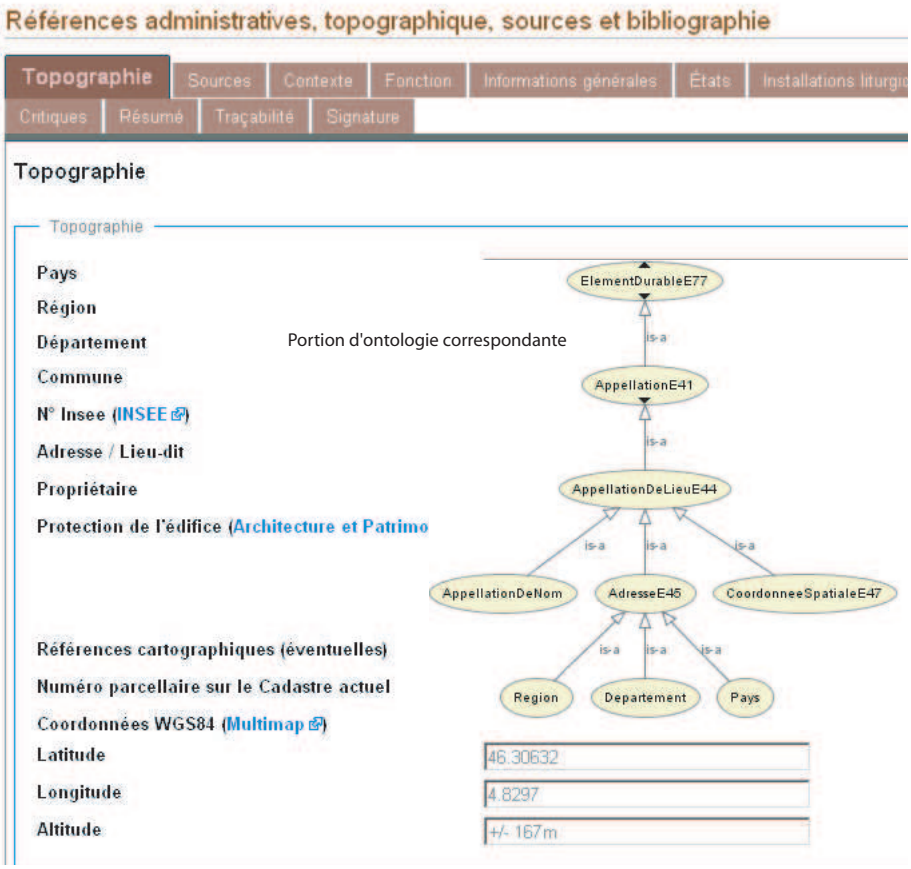


FIGURE A.21 – Partie d'ontologie relative aux appellations de lieu dans le projet CARE

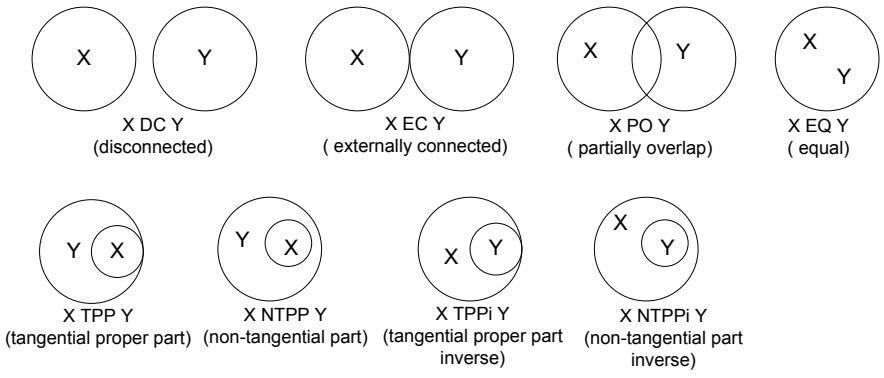


FIGURE A.22 – Relations topologiques de base

au sein des glaciers. L'ordre est déduit à partir d'une position relative comme la superposition, le remplacement partiel, l'obstruction et l'inclusion. L'ontologie CIDOC-CRM ne fournit pas ce type de relation ;

3. la causalité des relations entre les événements, c'est-à-dire les conditions préalables pour qu'un événement puisse avoir lieu.

– la **chronologie relative par inclusion d'événements** est une deuxième forme de chronolo-

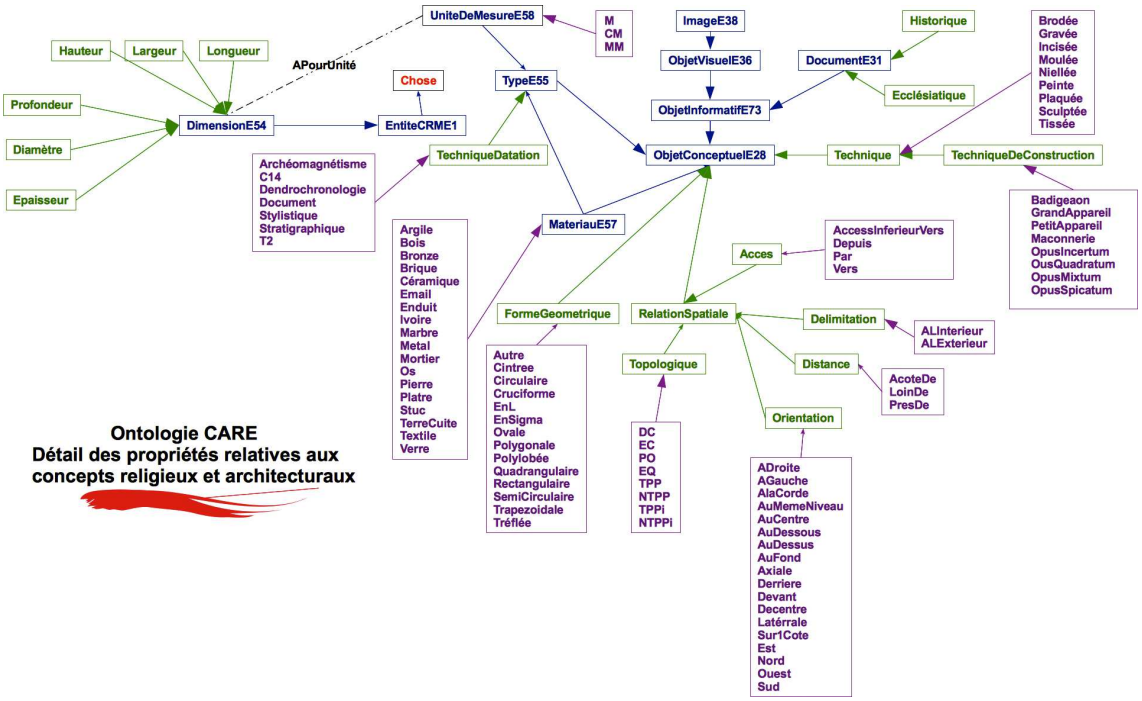


Figure A.23 – Partie d'ontologie relative aux propriétés des espaces religieux et des concepts architecturaux du projet CARE

gie relative : un processus contient des sous-processus qui peuvent être datés individuellement de façon relative ou absolue ;

– la **chronologie relative par distance temporelle** peut provenir :

1. de données historiques ;
2. de calcul de distance temporelle par l'estimation d'un taux de variation, comme les changements de style, la diffusion de compétences technologiques, l'abrasion des dents entre la naissance et la mort. Les estimations du temps de déplacement des personnes, des biens, des informations ou des technologies, basées sur la vitesse de communication, entrent également dans cette catégorie ;
3. de connaissances telles que l'espérance de vie, la durée d'utilisation moyenne des pots en argile, etc ;
4. la dernière forme de chronologie relative est basée sur le processus suivant : l'objet est classé dans une catégorie. La catégorie est associée à un type d'événements. L'objet peut donc être daté par la datation de la catégorie.

Les informations sur l'ordre des événements permettent la création d'un réseau temporel. Le réseau temporel peut être combiné avec des éléments de chronologie absolue. Le résultat est un ensemble renforcé qui peut affiner les intervalles jusqu'à transformer certains intervalles relatifs en intervalles absolus.

De ces constations, il découle que le modèle temporel dont nous avons besoin en archéologie doit être basé sur les critères suivants : quelques repères absolus mais surtout une chronologie relative basée sur des intervalles.

Le temps dans CIDOC-CRM

Le Groupe de travail archéologique de CIDOC a travaillé sur la partie "temps" de l'ontologie CIDOC-CRM [DKS04, DKS05]. Cette partie de l'ontologie doit permettre de :
- définir des périodes culturelles et des phases basées sur des caractéristiques différentes du contexte archéologique ;
- organiser l'information archéologique selon une chronologie cohérente et exploitable par une machine quelle que soit la provenance géographique des objets ;
- évaluer la chronologie pour les objets trouvés lors de fouilles archéologiques ;
- faciliter la communication et le partage des connaissances temporelles entre les archéologues, grâce à une représentation standardisée.

CIDOC-CRM offre une branche spécifique pour les concepts liés au temps (voir figure A.24). Le concept *Entité Temporelle E2* regroupe des notions telles que celles de période (*Période E4*) qui vaut pour une zone géographique donnée, d'événement (*Evénement E5*), d'état de conservation (*Etat de Conservation E3*) qui vaut pour un objet donné sur une durée variable et d'autres phénomènes limités dans le temps.
Un concept important dans CIDOC-CRM est la **période** définie de la façon suivante : « *cette classe (Periode E4) comprend des ensembles de phénomènes cohérents ou des manifestations culturelles limitées dans le temps et l'espace. C'est la cohésion sociale ou physique des phénomènes qui permet d'identifier une période et non les limites spatio-temporelles associées. Ainsi, des périodes différentes peuvent se recouper et coexister dans le temps et l'espace, comme lorsqu'une culture nomade existe dans la même région qu'une culture sédentaire* ... ». Cette définition est basée sur la notion de cohérence des phénomènes tels que des décisions politiques, une économie, etc. Comme la cohérence des phénomènes évolue progressivement, la définition d'une période est nécessairement floue vis-à-vis de l'espace-temps, mais les périodes n'en sont pas moins réelles. Par exemple, un phénomène peut apparaître dans une région et se propager lentement à d'autres régions, continuer à prospérer dans les régions éloignées de son bassin de création mais disparaître de son bassin d'origine. De plus, le degré de synchronisation entre différents types de phénomènes, tels qu'un style architectural et un système politique, peut varier considérablement. Cela donne lieu à de multiples points de vue divergeant dans les mêmes limites spatio-temporelles.

Certaines relations de Allen [All83] définissent les propriétés qui lient deux concepts *Entité Temporelle E2*. Accary et al. [ABC03, AC04] ont formalisé l'ensemble des relations observées sur les chronologies et ont défini quelques relations stratigraphiques. Ils ont ensuite utilisé les relations de Allen comme modèle pivot entre les diagrammes stratigraphiques et les frises chronologiques. Ce modèle pivot permet de comparer et de détecter les incohérences entre ces deux représentations (voir tableau A.2). Binding dans [Bin10] montre l'équivalence des relations temporelles de CIDOC-CRM avec celles de OWL-Time [8], Holmen et al. [HO09] montrent que CIDOC-CRM permet de traiter des dates floues.

Suivi des évolutions dans le projet CARE

Le projet CARE traite des états successifs (création, modification, destruction, etc.) d'un édifice et de ses éléments. Afin de représenter l'évolution de l'édifice, la notion d'état a été introduite par les archéologues : un état correspond à un changement substantiel de l'édifice (dans son architecture ou sa fonction), il est associé à un intervalle de temps déterminé par une technique de datation. De façon générale, le processus d'évolution des édifices a pour origine une relation cause-effet. Nous avons en conséquence traité :

8. OWL-Time : http://www.w3.org/TR/owl-time

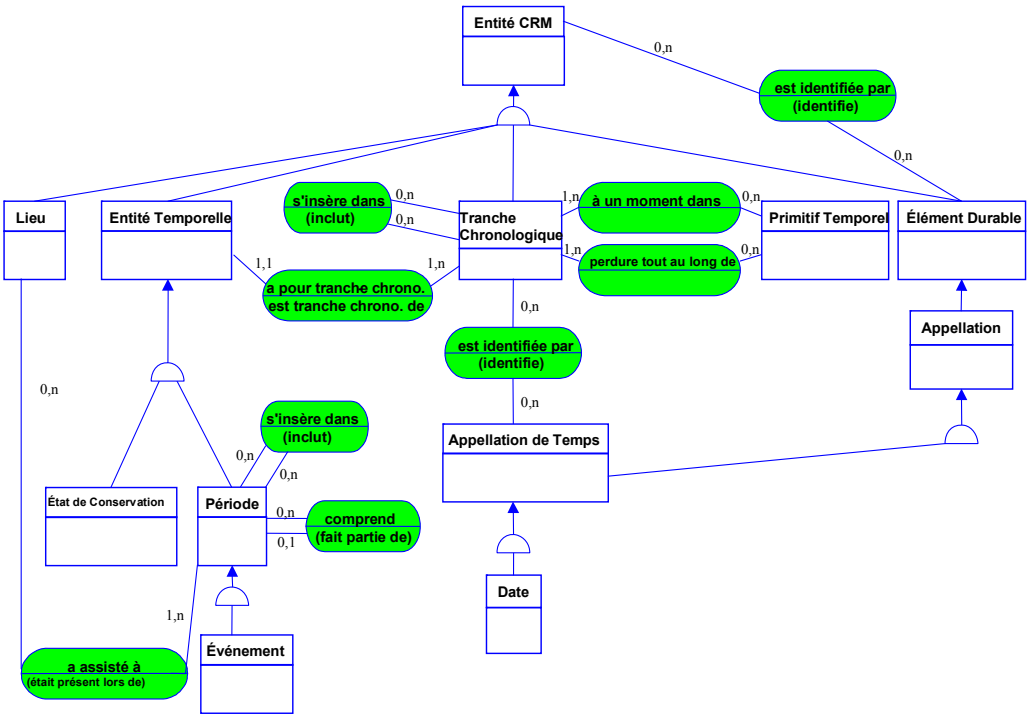


Figure A.24 – Les concepts et leurs relations liés au temps dans CIDOC-CRM (extrait de [CC02])

Vocabulaire descriptif	Expression en langage de Allen
A *sous* B	*avant*(A, B) OU *touche*(A, B)
A *même niveau que* B	*recouvre*(A, B) OU *recouvre*(B, A) OU *egal*(A, B)
A *remblais de* B	*après*(A, B)
A *tranchée de fondation de* B	*avant*(A, B)

Table A.2 – Relations dédiées aux stratigraphies proposées par Accary et al. et leur expression en langage de Allen (Extrait de [ABC03])

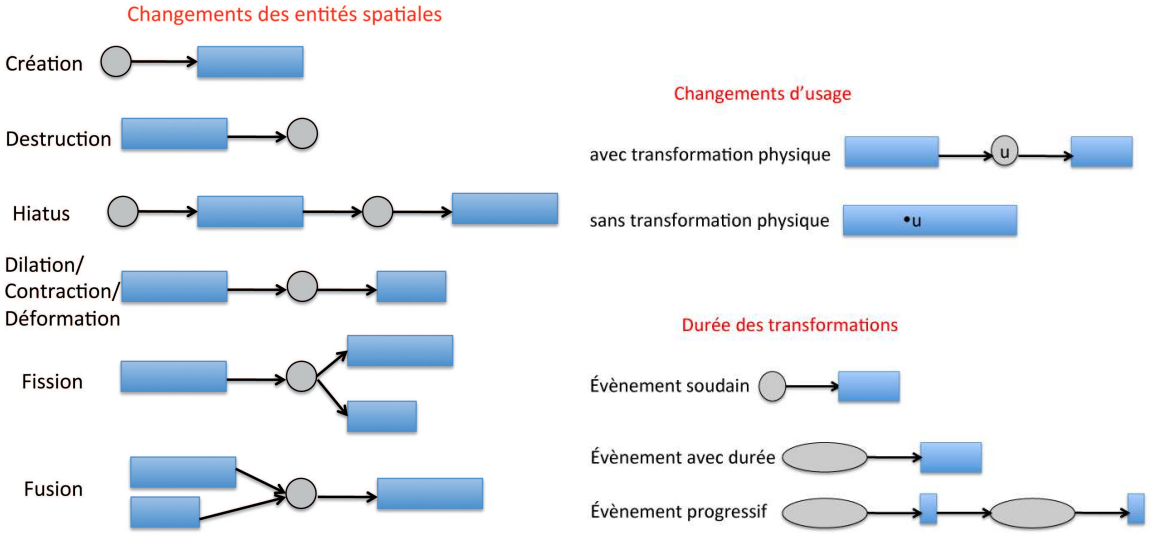
– les modifications d'usage de l'édifice. Par exemple l'église Saint-Pierre-Estrier à Autun a été transformée en grange au XIX^{e} siècle. Dans certains cas, ce type d'évolution peut impliquer des variations morphologiques ;
– les modifications morphologiques ;
– les modifications d'identité : un édifice change de nom, cette évolution dépend de l'histoire d'une société, ses traditions, sa langue ;
– les changements de propriétaires : par exemple l'édifice religieux vendu sous la Révolution.

Les facteurs qui contribuent à une évolution peuvent causer des variations graduelles et lentes (l'édification d'une cathédrale) ou des modifications soudaines (changement de propriétaire). Les évolutions peuvent prendre les formes suivantes :

– apparition : création d'un édifice ;
– disparition : destruction d'un édifice ;
– stabilité : l'édifice perdure dans le temps soit avec les mêmes usages, soit dans l'espace (même forme), soit les deux ;

– hiatus : un édifice apparaît, disparaît et réapparaît ;
– dilatation : l'édifice ou un espace dans l'édifice s'agrandit, se développe au cours du temps ;
– contraction : l'édifice ou un espace dans l'édifice se réduit au fil du temps ;
– déformation : le plan de l'édifice évolue ;
– fission : l'édifice se scinde en un groupe d'édifices ;
– fusion : un groupe se regroupe en un seul édifice.

Toutes ces évolutions sont schématisées en figure A.25 selon le formalisme proposé par Renolen dans [Ren00]. Celui-ci propose une notation à base de graphe qui prend en compte des évènements d'évolution tels que la création, la modification, la destruction, etc. Cette notation illustre les évolutions par une succession d'états représentés par des rectangles et de transitions représentées par des cercles si la durée est courte ou des rectangles aux bords arrondis si la durée de la phase est plus longue. Par exemple, en utilisant la notation de Renolen sur l'exemple de l'église Saint Clément de Mâcon, on obtient trois états entre la première moitié du VI^{e} siècle et le début du XI^{e} siècle. Dans le premier état (E1) l'église possède une nef unique à abside semi-circulaire saillante avec annexes latérales et portique orientée (voir ① dans la figure A.26). L'état 2 (E2) consiste à réduire cette nef, puis l'état 3 (E3) ajoute trois absides semi-circulaires saillantes orientée (voir ② dans la figure A.26). La figure A.26 montre les trois plans et les graphes historisés associés correspondant à la nef et aux absides.



(a) Changements morphologiques

(b) Changements des usages et impacts du temps sur les évolutions

FIGURE A.25 – Représentation des évolutions selon le formalisme de Renolen

Nous considérons le concept d'activité comme primordial car l'activité caractérise un état dans l'ontologie CIDOC-CRM. Le concept *Activité E7* est défini comme suit : « *Action ou enchaînement d'actions qu'exécutent des agents poursuivant un but déterminé, et qui débouchent globalement sur un changement d'état dans les systèmes culturels, sociaux, matériels, qui nous intéressent. Cette notion recouvre à la fois des actions complexes et durables telles que l'édification d'un établissement humain, ...* ».

Nous avons utilisé les concepts suivants de l'ontologie CIDOC-CRM pour modéliser les états :
– *Destruction E6* comprend les événements qui détruisent une ou plusieurs instances du concept *Quelque chose de Matériel E18*.
– *Événement de modification E11* comprend toutes les instances de *activité E7* qui créent, modifient ou changent *Quelque chose de Matériel et de Fabriqué E24*.

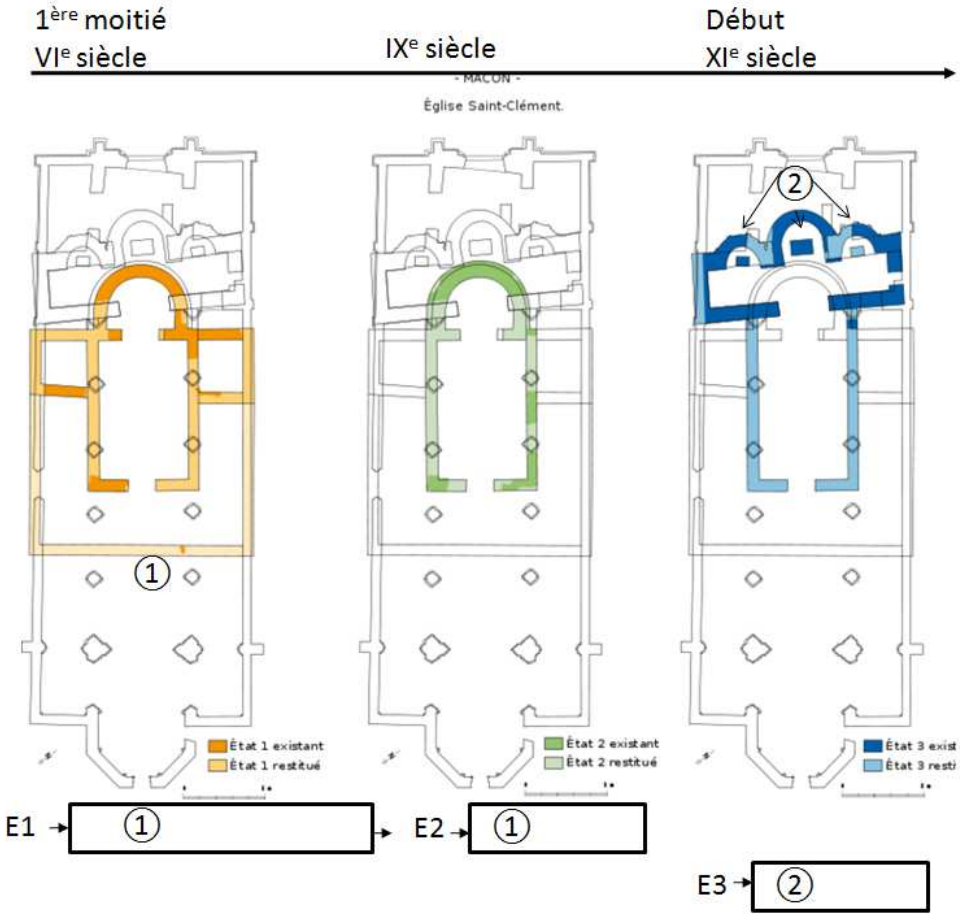


Figure A.26 – Graphes historisés de l'église Saint Clément de Mâcon (Plans élaborés par le Centre d'Études Mdédiévales d'Auxerre, 2010)

- *Début d'Existence E63* comprend les événements qui "apportent la vie" à tout *Élément Durable E77*, ainsi que l'événement *Evénement de Fondation E66*.
- *Fin d'Existence E64* comprend les événements qui mettent fin à l'existence de tout *Élément Durable E77* et sa spécialisation *Destruction E6*.
- *Ajout de Partie E79* comprend les activités qui se traduisent par le fait qu'une instance de *Quelque chose de Matériel et de Fabriquée E24* est élargie ou augmentée par l'ajout d'une partie.
- *Retrait de Partie E80* comprend les activités qui se traduisent par le fait qu'une instance de *Quelque chose de Matériel et de Fabriquée E24* est diminuée par la suppression d'une partie.
- *Transformation E81* comprend les événements qui entraînent la destruction simultanée d'un *Élément durable E77* et la(les) création(s) d'un autre *Élément durable E77*. La transformation préserve la substance identifiable du premier élément, mais exprime une nature différente. Les changements soit de titulature soit d'usage religieux correspondent à ce concept.

Nous étudions des édifices du IV^{e} au X^{e} siècle donc nous avons aussi besoin d'intervalles (siècle, début, fin et tiers de siècle). Une *Tranche Chronologique E52* est une plage temporelle ayant un début, une fin et une durée sans autres connotations sémantiques. Les tranches chronologiques sont utilisées pour définir la plage temporelle des instances de Période, d'Événement. Une même Tranche chronologique peut être identifiée au moyen d'une ou de plusieurs Appellations de Temps. À cela s'ajoute des périodes caractéristiques : Antiquité, Moyen Âge, etc. On obtient

donc la partie d'ontologie donnée en figure A.27.

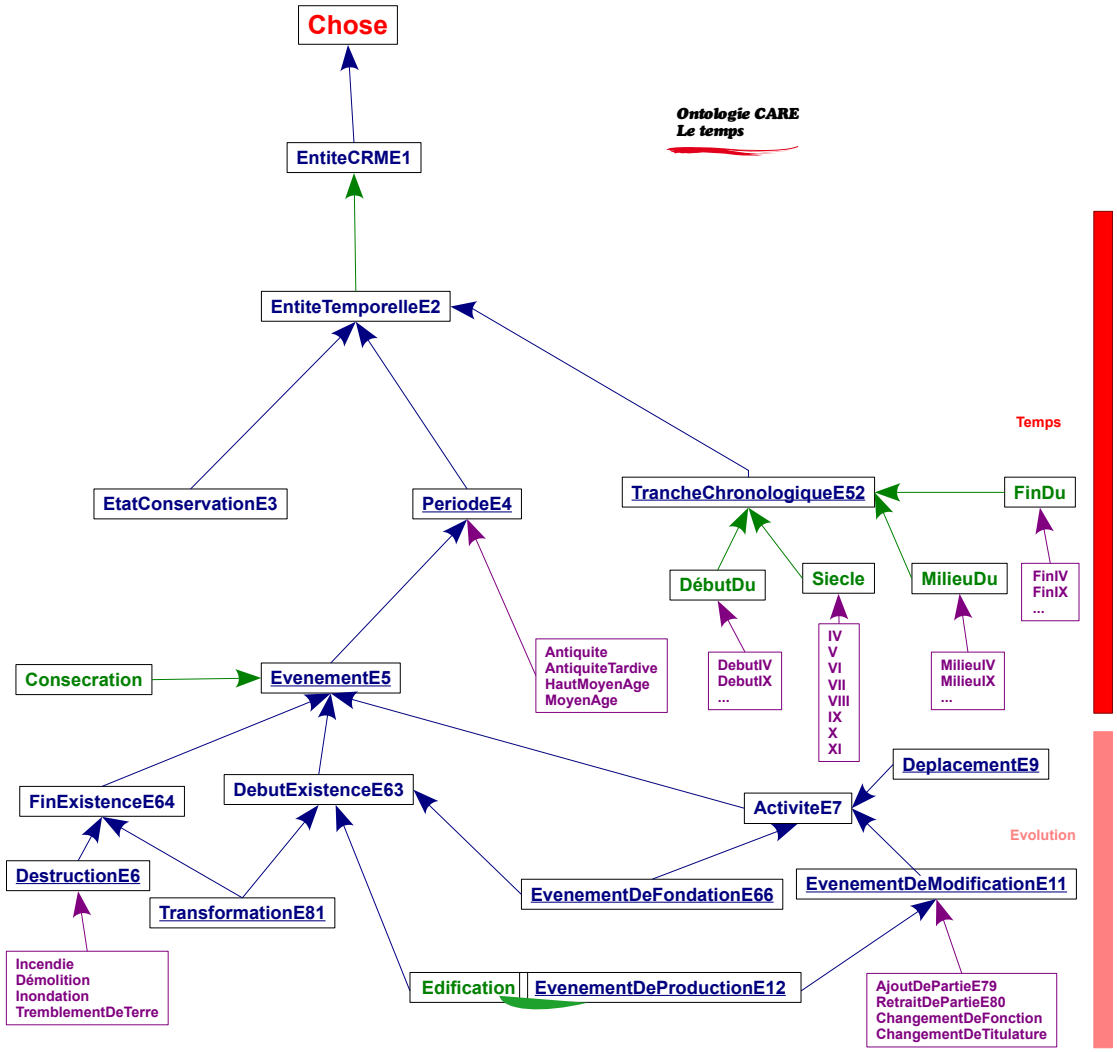


FIGURE A.27 – Partie d'ontologie relative au temps pour le projet CARE

Le principe est le même que celui de l'espace : il n'existe pas de redondance temporelle possible. Il n'existe qu'un seul espace, même si dans le temps plusieurs objets archéologiques s'y succèdent. De la même manière il n'y a qu'un seul temps, même s'il peut contenir plusieurs objets archéologiques situés dans des espaces distincts. Nous remarquons par comparaison que les éléments de représentation du temps sont assimilés à une droite où un point représente un instant et un segment de droite un intervalle de temps. Le tableau A.3 présente une analogie entre l'espace et le temps.

Synthèse

Nous avons construit une ontologie dont les branches correspondent aux aspects spatial, temporel et architectural. L'aspect architectural traite les aspects spécifiques aux édifices religieux chrétiens : les espaces et éléments constitutifs d'un édifice et les objets religieux. Chaque branche issue de `THING` peut être considérée comme une ontologie d'un domaine donné. Pour couvrir une plus large étendue de connaissances, nous avons établi les relations entre les différentes

Concepts d'analyse spatiale	Concepts d'analyse temporelle
Lieu	Date
Distance	Durée
Position relative	Datation relative
Interaction spatiale	Interaction temporelle

Table A.3 – Analogie Espace - Temps

branches. Nous avons ainsi obtenu une ontologie stratifiée où les concepts racine sont disjoints et où les dépendances entre les strates sont explicites.

Annexe B

Définitions des différentes composantes du corpus CARE

Sommaire

Chaque composante d'un édifice a fait l'objet d'une description textuelle précise par un archéologue. Nous avons remis en forme ces descriptions pour faire apparaître les concepts, propriétés et relations. Enfin, des définitions empruntées au vocabulaire religieux seront énumérées.

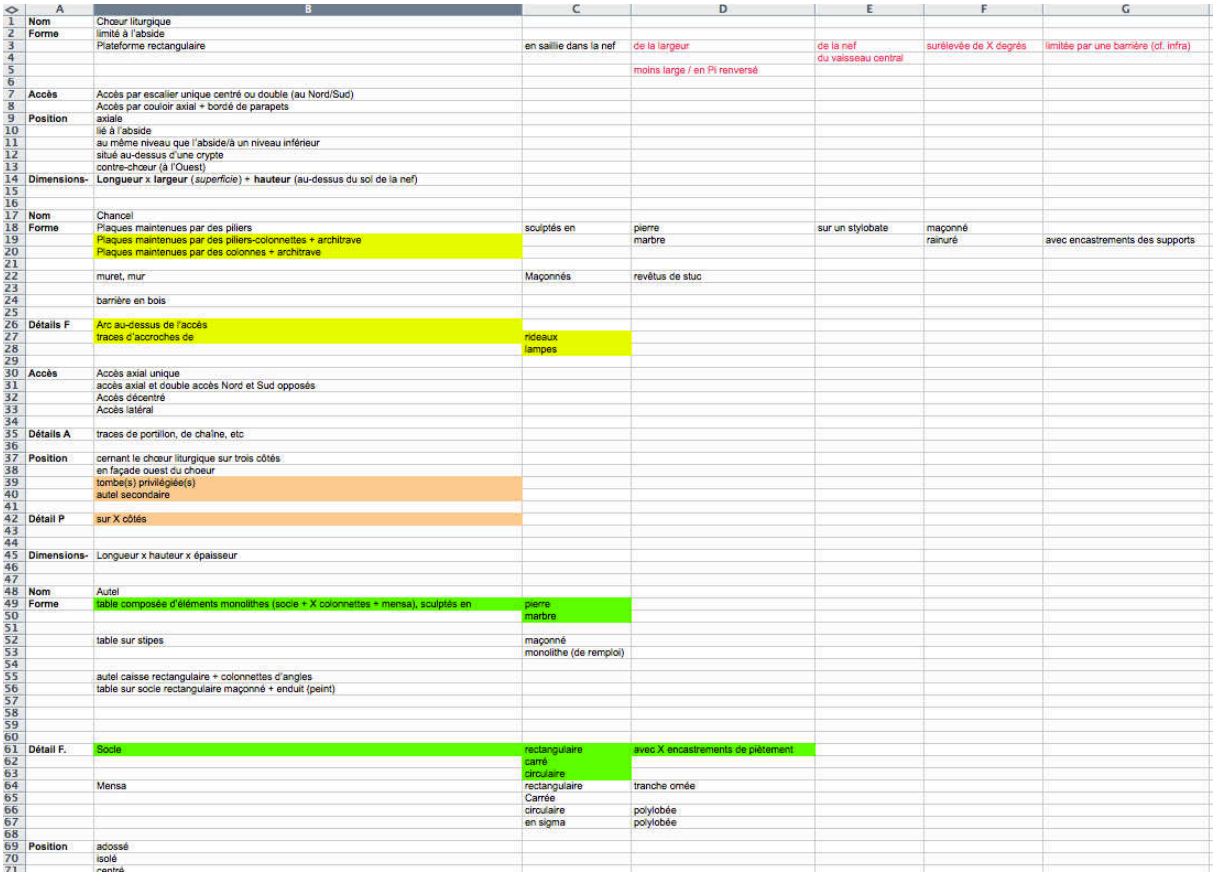
1 Mise en forme des descriptions textuelles des composantes du corpus CARE

Nous avons remis en forme les différentes descriptions dans un fichier excel (figure B.1) en faisant apparaître les caractéristiques : nom, forme, accès, position et dimensions, etc. Les couleurs dans la figure B.1 indiquent que les informations sont liées.
Pour des raisons de lisibilité, nous donnons un exemple :

[Nom] Atrium

[Forme] Quadrangulaire
Quadriportique, le portique Est formant un vestibule
Rectangulaire
Triportique joignant un vestibule fermé
Trapézoïdal
Ne présentant que des portiques latéraux Nord et Sud
[Accès] par porte unique centrée à l'Ouest
par porche saillant unique centré à l'Ouest
plusieurs accès à l'Ouest
accès latéral/au Nord

[Position] A l'Ouest de la nef
Au Nord



	A	B	C	D	E	F	G
1	**Nom**	Chœur liturgique					
2	**Forme**	limité à l'abside					
3		Plateforme rectangulaire	en saillie dans la nef	de la largeur	de la nef du vaisseau central	surélevée de X degrés	limitée par une barrière (cf. infra)
4							
5				moins large / en Pi renversé			
6							
7	**Accès**	Accès par escalier unique centré ou double (au Nord/Sud)					
8		Accès par couloir axial + bordé de parapets					
9	**Position**	axiale					
10		lié à l'abside					
11		au même niveau que l'abside/à un niveau inférieur					
12		situé au-dessus d'une crypte					
13		contre-chœur (à l'Ouest)					
14	**Dimensions-**	**Longueur x largeur** (*superficie*) + **hauteur** (au-dessus du sol de la nef)					
15							
16							
17	**Nom**	Chancel					
18	**Forme**	Plaques maintenues par des piliers	sculptés en	pierre	sur un stylobate	maçonné	
19		Plaques maintenues par des piliers-colonnettes + architrave		marbre		rainuré	avec encastrements des supports
20		Plaques maintenues par des colonnes + architrave					
21							
22		muret, mur	Maçonnés	revêtus de stuc			
23							
24		barrière en bois					
25							
26	**Détails F**	Arc au-dessus de l'accès					
27		traces d'accroches de	rideaux				
28			lampes				
29							
30	**Accès**	Accès axial unique					
31		accès axial et double accès Nord et Sud opposés					
32		Accès décentré					
33		Accès latéral					
34							
35	**Détails A**	traces de portillon, de chaîne, etc					
36							
37	**Position**	cernant le chœur liturgique sur trois côtés					
38		en façade ouest du chœur					
39		tombe(s) privilégiée(s)					
40		autel secondaire					
41							
42	**Détail P**	sur X côtés					
43							
44							
45	**Dimensions-**	Longueur x hauteur x épaisseur					
46							
47							
48	**Nom**	Autel					
49	**Forme**	table composée d'éléments monolithes (socle + X colonnettes + mensa), sculptés en	pierre				
50			marbre				
51							
52		table sur stipes	maçonné				
53			monolithe (de remploi)				
54							
55		autel caisse rectangulaire + colonnettes d'angles					
56		table sur socle rectangulaire maçonné + enduit (peint)					
57							
58							
59							
60							
61	**Détail F.**	Socle	rectangulaire	avec X encastrements de piètement			
62			carré				
63			circulaire				
64		Mensa	rectangulaire	tranche ornée			
65			Carrée				
66			circulaire	polylobée			
67			en sigma	polylobée			
68							
69	**Position**	adossé					
70		isolé					
71		centré					

Figure B.1 – Extrait de fichier excel des descriptions d'éléments

Au Sud

[Dimensions] Longueur x largeur

2 Vocabulaire religieux

Les définitions sont extraites de http://www.eglise.catholique.fr/ressources-annuaires/lexique/lexique.html

Abbaye Monastère dirigé par un Abbé ou une Abbesse. Bâtiments de ce monastère

Abbatiale Église principale d'une abbaye

Abside Extrémité arrondie de la nef principale d'une église généralement tournée vers l'Est, elle termine le chœur. Sa partie extérieure s'appelle le **chevet**

Ambon Podium ou pupitre surélevé, placé à l'entrée du chœur. C'est de l'ambon qu'est proclamée la parole de Dieu

Arc Structure de soutien à profil courbe : arc brisé, arc de plein cintre, etc.

Autel Endroit où l'on célèbre l'eucharistie

Baldaquin Sorte de dais, supporté par des colonnes, surmontant le maître autel ou le trône épiscopal

Basilique Église bâtie sur le plan des basiliques romaines c'est-à-dire rectangulaire divisé en nefs parallèles

Cathédrale Église principale d'un diocèse où se trouve le siège de l'évêché

Cathèdre Siège de l'évêque dans sa cathédrale où il préside les cérémonies

Chapelle rayonnante Petite chapelle qui entoure le chœur

Chapiteau Tête d'une colonne couronnant le fût, elle est habituellement sculptée

Chevet Extrémité de l'église derrière l'autel, on admire le chevet de l'extérieur

Chœur Lieu où se trouve l'autel et où se déroule la liturgie

Clocher-porche Lieu situé généralement à l'entrée principale de l'édifice, soit la façade ouest.

Collégiale Église qui, sans être cathédrale, possède un chapitre de chanoines

Colonnade Rangée ornementale de colonnes sur une façade ou autour d'un édifice

Crypte Chapelle souterraine, généralement placée sous le chœur d'une église

Déambulatoire Galerie faisant le tour du chœur

Dédicace Consécration solennelle d'une église comme lieu de culte et de prière

Massif occidental —en allemand Westwerk ou encore Westbau— Type de façade particulier d'église romane.

Narthex Portique ou vestibule transversal à l'entrée de certaines églises

Nef Partie allongée de l'église ouverte aux fidèles, comprise entre le mur antérieur et l'entrée du chœur ou la croisée du transept. C'est le plus grand ensemble, c'est là que sont rassemblés les fidèles

Pilastre Sorte de pilier décoratif engagé dans un mur

Piscine Bassin du baptistère dans lequel on descendait par des marches pour recevoir le baptême par immersion du corps

Plan basilical Plan rectangulaire avec trois nefs séparées par des colonnes et prolongé par une abside en forme de demi-cercle

Plan centré Plan circulaire, polygonal ou en forme de croix grecque

Primatiale Qui appartient ou ayant un rapport à un primat

Stalles Sièges autour du chœur d'une église pour l'usage du clergé

Tour-porche Ne porte pas de cloches, ce en quoi elle se différencie du clocher

Transept Nef transversale coupant la nef principale donnant ainsi la forme d'une croix latine

Travée Division transversale de la nef comprise entre deux piliers

La figure B.2 matérialise quelques unes de ces définitions.

3 Techniques de construction

Appareil ou opus en latin Terme qui désigne la façon dont les moellons, les pierres de taille ou les briques sont assemblés en maçonnerie

Petit appareil Appareil constitué de moellons, pierre ou brique d'une dimension inférieure à 20 cm

Moyen appareil Appareil constitué d'éléments ayant une dimension de 20 cm à 30 cm

Grand appareil Appareil constitué d'éléments de plus de 30 cm

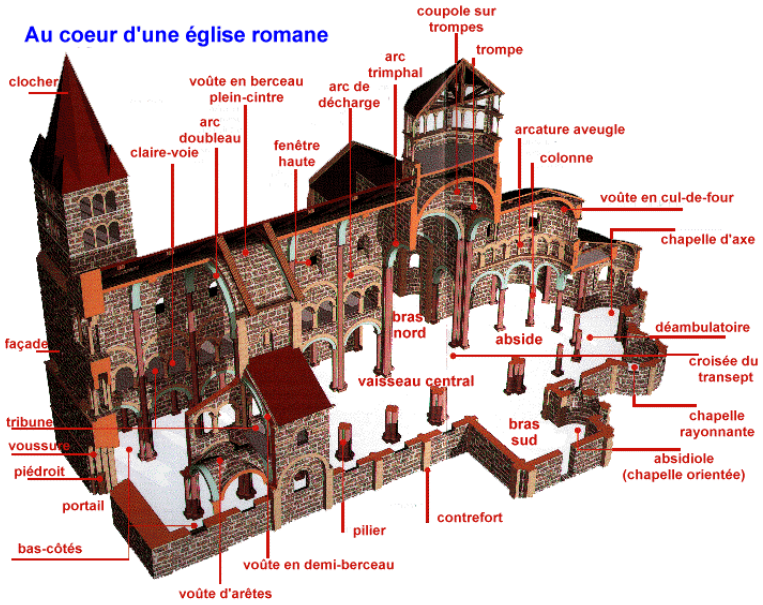


Figure B.2 – Plan 3D d'une église

Opus quadratum Assemblage de pierres taillées sans joint

Opus incertum Sans régularité

Opus mixtum Chaînage de briques alternant avec un appareil de pierres régulier

Opus spicatum Éléments posés sur leur champ, en lignes inclinées alternativement à droite et à gauche, figurant un motif en épi de blé. C'est un appareil de brique utilisé pour les sols

Opus sectile Plaquettes de marbre ou de pierre, parfois de verre coloré, découpées et assemblées de façon à constituer un dessin figuratif, c'est une technique de mosaïque

Opus signinum Mortier romain imperméable réalisé en mélangeant de la chaux, de l'eau, du sable de rivière, de la poudre de tuileaux. Il est employé en revêtement de sol

Terrazzo Sol de ciment pigmenté avec des morceaux de marbre de couleur

Annexe C

Fiche de dépouillement à usage des archéologues pour le corpus CARE

Les données de terrain recueillies par les archéologues, complétées aussi bien par des sources littéraires (encyclopédies, chroniques, annales, mémoires), des sources d'archives (archives d'État, archives ecclésiastiques) que des sources épigraphiques font l'objet d'un dépouillement exhaustif. La fiche de dépouillement créée par les archéologues comporte les rubriques suivantes :

- les informations générales sur l'édifice comme la topographie, les sources historiques, épigraphiques, archéologiques, la titulature, le diocèse, la fonction, le contexte d'implantation
- une description par état :
  - des éléments architecturaux,
  - des installations liturgiques,
  - des sépultures,
  - des inscriptions
- les objets dispersés non rattachables à l'architecture de l'édifice
- les considérations critiques sur les états et la chronologie puis des informations sur la publication (nom du rédacteur, date de rédaction de la fiche et sa fiabilité).

Cette fiche a été adoptée par l'ensemble des participants européens au projet. Elle est donnée dans [BJ12].

1. RÉFÉRENCES ADMINISTRATIVES, TOPOGRAPHIQUES, SOURCES ET BIBLIOGRAPHIE

(a) Topographie
Pays
Région
Département
Commune
N° INSEE
Adresse/Lieu-dit
Toponyme
Propriétaire
Protection de l'édifice : I.S.M.H. ou CL.M.H. ou sans
Références cartographiques
Sources cadastrales
Coordonnées WGS84
Altitude

(b) Sources historiques et identification

Sources indirectes
Sources archéologiques
Sources d'archives
Bibliographie
Sources épigraphiques
Références à la documentation graphique et photographique
Conservation
État actuel
Restauration
Titulature actuelle Titulature historique
Diocèse actuel Diocèse historique

(c) Contexte d'implantation
Brève description du site
Structures antérieures
Habitat contemporain

(d) Fonction
Bâtiment
Description

2. INFORMATIONS GÉNÉRALES

3. ARTICULATION EN ÉTATS
(a) Architecture
(b) Matériaux et techniques de construction
(c) Sépultures
(d) Inscriptions

4. OBJETS DISPERSÉS NON RATTACHABLES À L'ARCHITECTURE DE L'ÉGLISE

5. CONSIDÉRATIONS CRITIQUES SUR LES ÉTATS ET SUR LA CHRONOLOGIE

6. INFORMATIONS SUR LA PUBLICATION
Date
Auteur de la fiche
Statut de la fiche
Qualité de la fiche

Annexe D

Modélisation de la fiche de dépouillement avec l'extension Semantic Forms

Nous avons modélisé la fiche de dépouillement décrite en annexe C avec une extension de MediaWiki, Semantic Forms (http://www.mediawiki.org/wiki/Extension:Semantic_Forms) qui sert à créer des formulaires. Les formulaires sont écrits dans un langage de balisage et stockés dans des fichiers textes. Les fichiers textes sont ensuite analysés à la volée quand un formulaire est appelé. Cette modélisation permet d'intégrer automatiquement chaque édifice ou groupe d'édifices au corpus numérique. Les informations sont saisies directement au moyen d'un formulaire à remplir *via* des cases à cocher, des menus déroulants et des champs de saisie libre.
Nous présentons ci-dessous l'écriture de la fiche de dépouillement d'un édifice modélisée avec Semantic Forms.

```
<noinclude>
Ceci est le formulaire "Fiche d'édifice". Pour ajouter une page avec ce formulaire,
entrez son nom ci-dessous ; si elle existe déjà, vous serez dirigé
vers un formulaire destiné à l'éditer.

{{#forminput:Fiche édifice}}

</noinclude><includeonly>
==Références administratives, topographique, sources et bibliographie==
=Topographie=
===Topographie===
{{{for template|Topographie|label=Topographie}}}
{| class="formtable"
! Pays
| {{{field|pays|input type=dropdown|property=Pays|mandatory|default=France}}}
|-
! Région
| {{{field|region|input type=dropdown|property=Region|mandatory|default=Bourgogne}}}
|-
! Département
| {{{field|departement|input type=dropdown|property=Departement|mandatory}}}
|-
```

```
! Commune
| {{{field|commune|mandatory}}}
|-
! Nř Insee ([http://www.insee.fr/fr/methodes/nomenclatures/cog/ INSEE])
| {{{field|insee|mandatory}}}
|-
! Adresse / Lieu-dit
| {{{field|adresse}}}
|-
! Propriétaire
| {{{field|proprietaire}}}
|-
! Protection de l'édifice
([http://www.culture.gouv.fr/culture/inventai/patrimoine/ Architecture et Patrimoine])
| {{{field|protection|input type=textarea}}}
|-
! Références cartographiques (éventuelles)
| {{{field|references_cartographiques}}}
|-
! Numéro parcellaire sur le Cadastre actuel
| {{{field|num_parc_cad_act}}}
|-
! Coordonnées WGS84 ([http://www.multimap.com/maps/?qs=nevers&countryCode=FR Multimap])
|-
! Latitude
| {{{field|latitude}}}
|-
! Longitude
| {{{field|longitude}}}
|-
! Altitude
| {{{field|altitude}}}
|}
{{{end template}}}

=Sources=
===Sources historiques et identification===
{{{for template|SrcHistId|label=Sources historiques et identification}}}
{| class="formtable"
! Sources épigraphiques
| {{{field|sources_epigraphiques|input type=textarea}}}
|-
! Sources indirectes
| {{{field|sources_indirectes|input type=textarea}}}
|-
! Sources archives
| {{{field|sources_archives|input type=textarea}}}
|-
! Sources archéologiques
```

```
| {{{field|sources_archeologiques|input type=textarea}}}
|-
! Bibliographie
| {{{field|bibliographie|input type=textarea}}}
|-
! Références à la documentation graphique et photographique
| {{{field|references_documentation|input type=textarea}}}
|-
! Conservation
| {{{field|conservation|input type=textarea}}}
|-
! Titulature actuelle
| {{{field|titulature_acteulle|input type=textarea}}}
|-
! Titulature historique
| {{{field|titulature_historique|input type=textarea}}}
|-
! Diocèse actuel
| {{{field|diocese_actuel|input type=textarea}}}
|-
! Diocèse historique
| {{{field|diocese_historique|input type=textarea}}}
|}
{{{end template}}}

=Contexte=
===Contexte d'implantation===
{{{for template|Contexte|label=Contexte d'implantation}}}
{| class="formtable"
! Brève description du site
| {{{field|description|input type=textarea}}}
|-
! Structures antérieures
| {{{field|structures_anterieures|input type=textarea}}}
|-
! Habitat contemporain (du IVe au Xe siècle)
| {{{field|habitat_contemporain|input type=textarea}}}
|}
{{{end template}}}

=Fonction=
===Fonction===
{{{for template|Fonction|label=Fonction}}}
{| class="formtable"
! Description
| {{{field|description|input type=textarea}}}
|}
{{{end template}}}
```

```
=Informations générales=
==INFORMATIONS GENERALES==
{{{for template|Informations générales|label=Informations Générales}}}
{| class="formtable"
! Description
| {{{field|description|input type=textarea}}}
|}
{{{end template}}}

=États=
==ARTICULATION EN ETATS==
===Structure complexe===
{{{for template|Structure|multiple}}}
{| class="formtable"
! Etat
| {{{field|etat|input type=number|mandatory}}}
|-
! Plan
| {{{field|plan|input type=dropdown|property=Plan}}}
|-
!
| {{{field|PlanImage|uploadable}}}
|-
! Parties
| {{{field|parties|input type=textarea}}}
|-
! Baptistère
| {{{field|eventuel_baptistere|input type=textarea}}}
|}
{{{end template}}}

===Matériaux et techniques de construction===
{{{for template|Matech|multiple}}}
{| class="formtable"
! Etat
| {{{field|etat||input type=number|mandatory}}}
|-
! Activité
| {{{field|activite|input type=textarea}}}
|-
! Murs et maçonnerie
| {{{field|maconnerie|input type=textarea}}}
|-
! Sol - pavement
| {{{field|sol|input type=textarea}}}
|-
! Couverture
| {{{field|couverture|input type=textarea}}}
|-
```

```
! Autres éléments structurels et architectoniques
| {{{field|autres|input type=textarea}}}
|-
! Décor appliqué aux murs et maçonneries
| {{{field|decor|input type=textarea}}}
|}
{{{end template}}}

=Installations liturgiques=
==INSTALLATIONS LITURGIQUES==
{{{for template|Installations liturgiques|multiple}}}
{| class="formtable"
! Etat
| {{{field|etat||input type=number|mandatory}}}
|-
! Barrières
| {{{field|delimitation|input type=textarea}}}
|-
! Autel
| {{{field|autel|input type=textarea}}}
|-
! Loculus
| {{{field|loculus|input type=textarea}}}
|-
! Reliquaire
| {{{field|reliquaire|input type=textarea}}}
|-
! Ciborium
| {{{field|ciborium|input type=textarea}}}
|-
! Pupitre
| {{{field|pupitre|input type=textarea}}}
|-
! Ambon
| {{{field|ambon|input type=textarea}}}
|-
! Bénitier
| {{{field|benitier|input type=textarea}}}
|-
! Lavabo
| {{{field|lavabo|input type=textarea}}}
|-
! Armoire
| {{{field|armoire|input type=textarea}}}
|-
! Supports
| {{{field|supports|input type=textarea}}}
|-
! Traces
```

```
| {{{field|traces|input type=textarea}}}
|}
{{{end template}}}

=Sépultures=
==SEPULTURES==
{{{for template|Sépultures|multiple}}}
{| class="formtable"
! Etat
| {{{field|etat||input type=number|mandatory}}}
|-
! Emplacement et relation avec l'édifice
| {{{field|emplacement|input type=textarea}}}
|-
! Structure
| {{{field|structure|input type=textarea}}}
|-
! Usage
| {{{field|usage|input type=textarea}}}
|-
! Mobilier
| {{{field|mobilier|input type=textarea}}}
|-
! Rituel
| {{{field|rituel|input type=textarea}}}
|-
! Anthropologie, paléo-pathologie et paléo-nutrition du défunt
| {{{field|defunt|input type=textarea}}}
|}
{{{end template}}}

=Objets dispersés=
==OBJETS DISPERSES NON RATTACHABLES A L'ARCHITECTURE DE L'EGLISE==
{{{for template|Objets}}}
{| class="formtable"
! Description
| {{{field|description|input type=textarea}}}
|}
{{{end template}}}

=Inscriptions=
==INSCRIPTIONS==
{{{for template|Inscriptions|multiple}}}
{| class="formtable"
! Etat
| {{{field|etat||input type=number|mandatory}}}
|-
! Localisation
| {{{field|localisation|input type=textarea}}}
```

```
|-
! Typologie
| {{{field|typologie|input type=textarea}}}
|-
! Dimensions
| {{{field|dimensions|input type=textarea}}}
|-
! Materiaux
| {{{field|materiaux|input type=textarea}}}
|-
! Chronologie
| {{{field|chronologie|input type=textarea}}}
|-
! Conservation
| {{{field|conservation|input type=textarea}}}
|-
! Bibliographie
| {{{field|bibliographie|input type=textarea}}}
|-
! Transcription
| {{{field|transcription|input type=textarea}}}
|-
! Apparat critique
| {{{field|apparat_critique|input type=textarea}}}
|-
! Apparat textuel
| {{{field|apparat_textuel|input type=textarea}}}
|-
! Traduction
| {{{field|traduction|input type=textarea}}}
|-
! Commentaire
| {{{field|commentaire|input type=textarea}}}
|}
{{{end template}}}

=Critiques=
==CONSIDERATIONS CRITIQUES SUR LES ETATS ET SUR LA CHRONOLOGIE==
{{{for template|Considérations critiques}}}
{| class="formtable"
! Chronologie et arguments de datation
| {{{field|chronologie|input type=textarea}}}
|-
! Interprétation
| {{{field|interprétation|input type=textarea}}}
|-
! Comparaisons
| {{{field|comparaisons|input type=textarea}}}
|}
```

```
{{{end template}}}

=Résumé=
'''Texte libre: pour les données qui n'ont trouvé aucun refuge (dev only) '''
{{{standard input|free text}}}
{{{standard input|summary}}}

=Traçabilité=
{{{standard input|minor edit}}} {{{standard input|watch}}}

=Signature=
==Signature==
{{{for template|System}}}
{| class="formtable"
! Date
| {{{field|date_fiche|input type=date|mandatory}}}
|-
! Auteur
| {{{field|auteur|input type=dropdown|property=Author|mandatory}}}
|}
{{{end template}}}

<headertabs/>

{{{standard input|save}}} {{{standard input|preview}}}
{{{standard input|changes}}} {{{standard input|cancel}}}
</includeonly>
```

Annexe E

Le projet I3-CRB (*Infrastructure Informatique Interopérable pour les Centres de Ressources Biologiques*)

Sommaire

Texte tiré en grande partie du rapport final d'activité envoyé au GIS IBiSA.

Notre projet I3-CRB (*Infrastructure Informatique Interopérable pour les Centres de Ressources Biologiques*) a été sélectionné à la suite d'un appel d'offres lancé par le GIS IBiSA (Infrastructure en Biologie Santé et Agronomie). Ce projet, d'une durée de deux ans (2009-2010), avait pour objectif de fournir un moyen d'information à la communauté scientifique sur les ressources biologiques humaines, microbiennes, animales et végétales, à sa disposition, par un **annuaire**. Cet annuaire des Centres de Ressources Biologiques [1] (CRB) doit améliorer la visibilité des collections biologiques et faciliter les échanges d'échantillons comme plate-forme documentaire communautaire pour l'ensemble des collections biologiques [2] conservées par les différents CRB français.

Les ressources biologiques (par exemple, des organismes vivants, des cellules, des gènes) et les informations qui s'y rapportent sont des éléments essentiels de la recherche en biologie, en santé mais aussi pour les progrès des biotechnologies. Il est donc capital de garantir la qualité de conservation des échantillons comme celle de leur distribution et de leurs données associées. Ce rôle est tenu par les CRB. Les collections d'échantillons de ressources biologiques sont constituées d'échantillons de différentes natures : des cellules, du sang, de l'urine, de l'ADN, etc. Ces collections comportent également les données associées aux échantillons : nature de l'échantillon, conditions de prélèvement et de conservation, informations relatives à l'état de santé du patient,

1. D'après la définition de l'OCDE (http://www.oecd.org/dataoecd/7/11/38777441.pdf), un CRB est une structure détenant des échantillons biologiques et des données associées provenant des différents règnes du vivant : humain, animaux, végétaux et micro-organismes. Le rôle des CRB est de conserver, transformer et distribuer les échantillons biologiques à des fins de recherche-développement.

2. Les collections publiées respecteront les droits des différentes parties sous la responsabilité du CRB d'origine.

voire des traitements réalisés sur l'échantillon, etc. Sans ces informations, les échantillons biologiques sont inexploitables : la qualité et la faisabilité des recherches dépendent largement de la pertinence des données associées aux échantillons ainsi que du nombre d'échantillons disponibles. Pour cela, les CRB disposent de bases de données contenant des informations sur ces collections, et d'outils nécessaires à l'analyse de ces données comme DataBiotec [3], Cresalys [4]. Le développement des CRB est en plein essor depuis dix ans et s'est fortement accéléré récemment, notamment en Europe dans le cadre du projet *Biobanking and Biomolecular Resources Research Infrastructure* (BBMRI) [5].

1 Exigences auxquelles doit répondre l'annuaire

L'annuaire répond à cinq exigences :

1. les quatre règnes du vivant sont représentés dans l'annuaire : humain, animal, végétal, micro-organisme ;
2. l'annuaire est conçu sur un mode participatif : les personnels des CRB enregistrent et mettent à jour eux-mêmes les données qu'ils souhaitent partager de manière complètement autonome après inscription de leur CRB à l'annuaire ;
3. la saisie des données est simple et peu coûteuse en temps, les champs à renseigner sont simplifiés, relativement peu nombreux et portent sur des données non sensibles ;
4. la plupart des champs doivent être renseignés en sélectionnant dans des listes déroulantes un ou plusieurs éléments issus d'ontologies, permettant ainsi de limiter les erreurs de saisie et de faciliter la consultation en homogénéisant les données ;
5. l'annuaire est consultable librement en anglais et en français sur Internet (http://www.i3crb.fr).

2 Architecture de I3-CRB

Nous avons modélisé trois entités et leurs relations pour la réalisation de l'annuaire : les CRB, leurs membres et leurs collections (voir figure E.1). La modélisation des échantillons a été écartée car les données qui décrivent un échantillon sont complexes et nombreuses et dépassent donc le cadre de l'annuaire. La saisie de la plus grande partie des informations est basée sur des ontologies afin d'éviter les erreurs de saisie, d'accroître la cohérence du vocabulaire utilisé pour décrire les données et d'améliorer l'efficacité du système d'interrogation par mots-clefs de l'annuaire. En effet, les recherches sur les collections traitant du cancer sont compliquées quand la maladie est décrite tantôt par "néoplasme", tantôt par "cancer" voire par "tumeur".
La thématique d'une collection est définie par l'association de trois mots-clefs traitant du "Domaine", de la "Maladie" et de l'"Organe touché". Le choix a été arrêté sur la terminologie internationale Medical Subject Headings (MeSH) [6], utilisée pour l'indexation et la recherche d'informations médicales. Les mots-clés de l'annuaire sont mis à jour après la mise à disposition de la version annuelle du thésaurus MeSH par le NCBI.
L'ontologie des espèces contient 421 660 entrées qui correspondent à la base Taxonomy du NCBI [7]. Pour chaque espèce sont décrits son nom scientifique et son lignage, ainsi que, le cas

3. http://www.oriam.com/Produits/DataBiotec.html
4. http://www.alphelys.com/alph01/prod/fr/gestionebio/GestionEchantBio.php
5. http://www.bbmri.eu
6. http://www.nlm.nih.gov/mesh/meshhome.html
7. http://www.ncbi.nlm.gov/Taxonomy

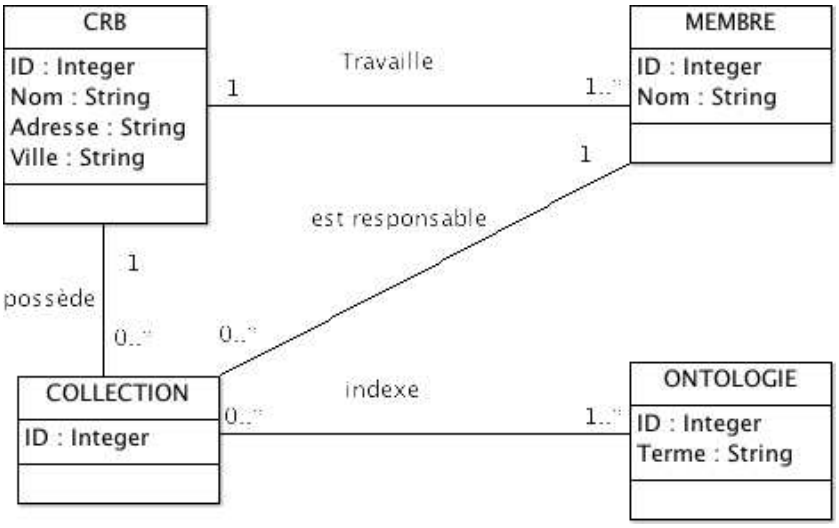


FIGURE E.1 – Schéma général de l'annuaire I3-CRB

échéant, son nom commun anglais et d'éventuels synonymes. La saisie de collections multi-espèces a été autorisée, notamment pour les collections micro-organismes pouvant contenir des échantillons de nombreuses souches différentes. Le nombre d'entrées dans l'ontologie des espèces étant trop élevé pour que la saisie des espèces puisse se faire par un choix dans une liste multiple, il a fallu implémenter un autre système de saisie. Le formulaire de saisie des espèces est donc composé d'un champ de recherche du nom scientifique ou du nom commun et de la sélection dans la liste des espèces correspondant à la recherche en cours. L'ontologie des espèces est mise à jour deux fois par an.
L'annuaire étant bilingue, le portail terminologique TermSciences, développé par l'INIST, a été utilisé pour la traduction des termes en français.

Le champ "Produit(s)" définit le type d'échantillons de la collection disponibles pour un échange entre le CRB et une autre structure. La liste des produits ne contient que des termes génériques comme cellule, liquide biologique ou protéine. Néanmoins, des CRB traitant du règne humain souhaitent une description plus fine en particulier pour le sang et ses produits dérivés. Une liste, tirée de MeSH, de quarante-quatre liquides biologiques a été dressée pour les règnes humain et animal. Nous avons finalement décidé de séparer dans la liste des produits "Sang et produits dérivés" et "Autres liquides biologiques" afin de conserver un même niveau de détail pour chaque règne et éviter de fournir une liste trop volumineuse de termes.

3 Couche d'interaction avec les utilisateurs

La saisie des données dans l'annuaire n'est possible par les membres d'un CRB qu'après l'inscription de celui-ci à l'annuaire. L'inscription est facile et peu contraignante : il suffit au responsable administratif du CRB de compléter le formulaire d'inscription, téléchargeable depuis le site, et de le retourner au responsable informatique de l'annuaire en deux exemplaires : l'un par message électronique, et l'autre par la poste (il tient lieu de demande officielle en comportant la signature originale du responsable et le tampon du CRB). Après réception du formulaire, le CRB est enregistré et un compte d'Administrateur-Annuaire du CRB est créé pour son responsable administratif. Cet administrateur pourra alors à tout moment mettre à jour les données de son CRB, créer, modifier ou supprimer des membres de son CRB, et créer, modifier et supprimer des

collections. Lors de la création d'un nouveau compte membre, l'administrateur du CRB peut choisir de lui donner les droits standards de Membre du CRB ou les droits d'administrateur du CRB. Les membres du CRB peuvent modifier leurs données professionnelles, et mettre à jour les collections dont ils sont responsables.
L'interface de saisie n'est disponible qu'en langue française, mais les ontologies sur lesquelles sont basées la plupart des champs sont bilingues, ce qui assure la traduction en anglais à la volée des données, et permet les interrogations en français et en anglais.

4 Couche d'accès à l'information par l'utilisateur

L'annuaire peut donc être consulté en anglais ou en français, et selon deux modes de consultation différents : parcourir et rechercher.
Le mode parcourir est disponible à partir de la page d'accueil de l'annuaire. Il présente les données de manière synthétique sous la forme de deux tableaux récapitulatifs. Le premier tableau recense les données globales c'est-à-dire le nombre de CRB, de membres, de collections, d'échantillons et d'individus ayant apporté un ou plusieurs échantillons. Le second tableau recense le nombre de ces mêmes objets par règne. Il est possible d'accéder à des données plus détaillées en cliquant sur les boutons contenus dans les cellules des tableaux.

Le second mode de consultation de l'annuaire, se fait par un formulaire de recherche avancée permettant une recherche multi-critères et itérative portant sur les différentes entités enregistrées dans l'annuaire. Les critères disponibles sont basés sur les ontologies (voir figure E.2). Par exemple, pour les collections, son nom, son appartenance à une biothèque, son statut juridique, le règne, le(s) espèce(s), le(s) domaine(s), le(s) maladie(s), le(s) produit(s), le commentaire, et le nombre (minimum, maximum) des individus et des échantillons associés. Les critères sélectionnés aux étapes précédentes d'une recherche restent valides pour l'ensemble de la recherche.

L'annuaire est en ligne depuis février 2009 et les statistiques sont comptabilisées depuis avril 2009. En juillet 2012, l'annuaire comptait 39 CRB, 59 membres, 164 collections et 662 022 échantillons. Le nombre de consultations augmente régulièrement et dépasse les 16 000 consultations.

Publications associées

[1] F. Jadeau, A Bartheleix, A. Burgun-Parenthoine, C. Deleval, P. Gele, C. Libersa, M. Savonnet, E. Tabone, C. Combet, The I3-CRB project and the French biological resources centers directory, *Fundamental & Clinical Pharmacology*, Vol.24, pp. 71-71, 2010

[2] F. Jadeau, A Bartheleix, A. Burgun-Parenthoine, C. Deleval, P. Gele, C. Libersa, M. Savonnet, E. Tabone, C. Combet, L'annuaire des Centres de Ressources Biologiques français, *Médecine/Sciences*, Vol.27, n^{o}10, pp. 895-897, Octobre 2011

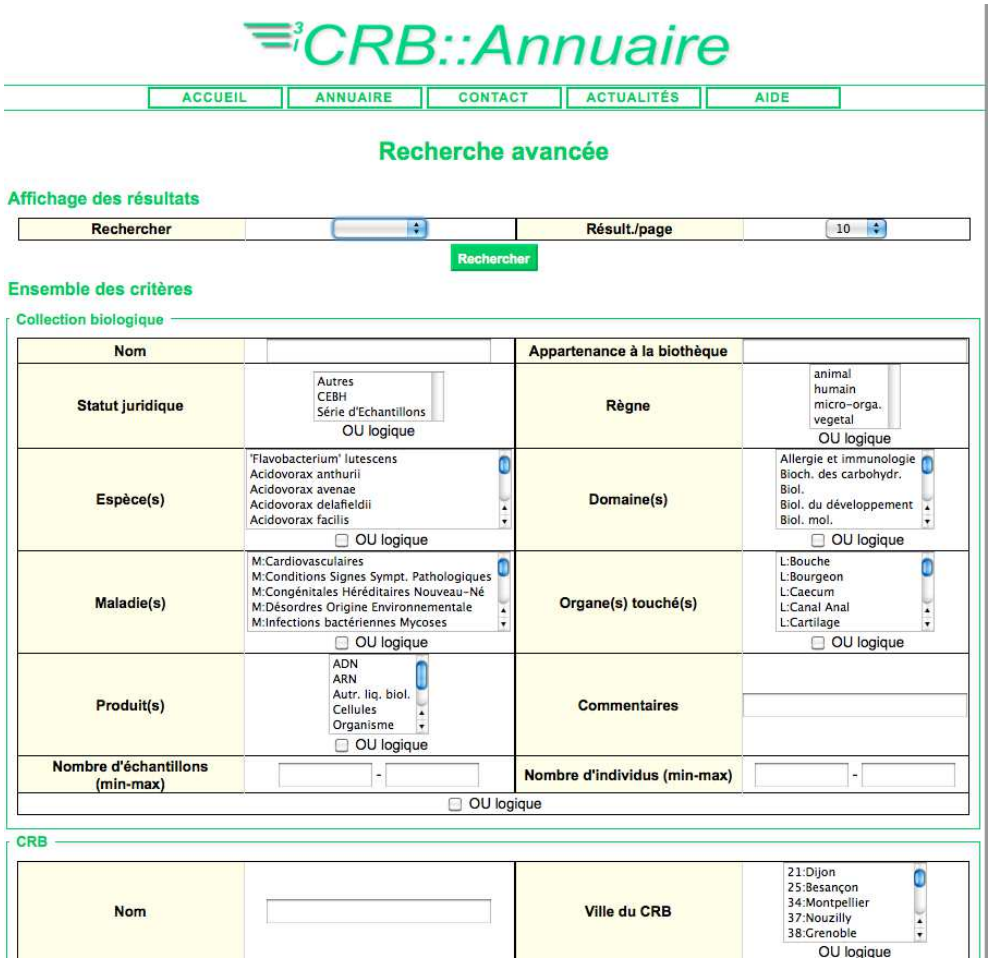


FIGURE E.2 – Capture d'écran de la fonctionnalité recherche avancée du projet I3-CRB

Annexe F

Résumé d'activité

1 Déroulement de carrière

1996 - Maître de Conférences à l'UFR Sciences et Techniques de l'Université de Bourgogne

1994 -1996 ATER à l'UFR Sciences et Techniques de l'Université de Bourgogne

1996 Doctorat en informatique de l'Université de Bourgogne « Fragtique : une méthodologie de fragmentation de Bases de Données Orientée Objets »

1993 DEA d'Informatique Automatique et Productique filière informatique (Université de Bourgogne et de Franche-Comté)

2 Encadrement (thèses, mémoires d'ingénieur, masters)

Co-encadrement de la thèse Région/Entreprise de Pierre Naubourg sous la direction de Marie-Noëlle Terrasse puis de Kokou Yétongnon : « Une approche préventive de fusion et d'évolution des données d'un système d'information : application aux données biomédicales » soutenue en décembre 2011

Co-encadrement de la thèse de Jean-Claude Simon sous la direction de Marie-Noëlle Terrasse : « Une approche liée à la préservation des propriétés transversales pour construire une offre de services web d'un éco-système d'entreprises » soutenue en novembre 2008

Encadrement du **mémoire d'ingénieur CNAM** Ingénierie Intégration Informatique : Systèmes d'Information par Mickaël Choisnard : « Interopérabilité de la gestion des services et des locaux pour une optimisation des ressources humaines et matérielles à l'Université de Bourgogne » soutenu en mars 2008

Encadrement du **mémoire de Master recherche** Informatique et Instrumentation de l'Image (3I) par Laurent Bossu : « Étude des transformations de modèles » soutenu en juillet 2007

Co-encadrement du **mémoire de Master recherche** Informatique et Instrumentation de l'Image par Mounir Haddou, « Méta-modélisation et Ontologies GO et GONG » soutenu en juillet 2007

Co-encadrement du **mémoire de DEA** Instrumentation et Informatique de l'Image par Christophe Allier, « Application des bases de données multidimensionnnelles aux images » soutenu en 2000

3 Activités de recherche

Depuis que j'ai été nommée Maître de Conférences au laboratoire LE2I UMR 6306 (http://le2i.cnrs.fr/), mes activités de recherche s'articulent autour des Systèmes d'Information selon les axes suivants :

1. la poursuite de mes travaux de thèse avec une stratégie de fragmentation et de distribution de Bases de Données Orientées Objets ;
2. l'étude des relations entre modèles, méta-modèles et ontologies pour la construction d'une architecture de modèles dans l'ingénierie de plate-formes ;
3. le contrôle des données par la connaissance et la problématique d'extensibilité dans les Systèmes d'Information Scientifique.

Environnement de conception et de gestion de Bases de Données Orientées Objets distribuées (1996-2001)
Nous avons proposé une stratégie de distribution orientée de telle sorte que la sémantique de la Base de Données soit préservée tout en assurant une autonomie et une spécialisation maximum des sites. La distribution de la Base de Données se déroulait en deux étapes : 1) mise en place d'une version initiale de la distribution sur la base des informations globales ; 2) évaluation de la distribution obtenue en termes de charge et de flux d'information et remise en cause éventuelle, pour l'itération suivante, de certains des choix faits. Dans cette proposition, les différentes évaluations étaient menées par approximations successives afin d'éviter, autant que possible, une descente des calculs jusqu'au niveau des objets. Cette approximation facilitait l'application de notre mécanisme à des données complexes pour lesquelles des calculs exhaustifs deviennent trop lourds. Les travaux menés portaient essentiellement sur le problème d'évaluation de la qualité de la distribution obtenue.

Métamodélisation et ingénierie de plate-formes (2000-2007)
Notre objectif était de construire un Système d'Information pour un domaine complexe généralement constitué de plusieurs sous-domaines. Nous avons proposé une plate-forme générique, dont l'architecture présente les caractéristiques suivantes :
- un système formel de référence basé sur la notion de méta-modèle et d'ontologie permettant de concevoir des applications mais aussi de contrôler leurs évolutions et de garantir les bonnes conditions de leurs interactions ;
- un modèle applicable à la majorité des applications du domaine à modéliser. Il est accompagné d'un guide d'instanciation expliquant les enjeux des différents choix à effectuer pour construire le modèle propre à une application. Un couplage entre cette architecture et le vocabulaire propre au domaine a été réalisé, ce couplage permet entre autres de définir les domaines des attributs du modèle instancié.
L'apport d'une telle plate-forme en termes d'interopérabilité et d'évolution de Systèmes d'Information a été étudié dans des domaines d'application tels que le domaine biomédical, les systèmes d'information géographiques, le e-learning, l'e-administration.
Dans le domaine biomédical, les travaux ont donné lieu à la thèse de Pierre Naubourg que j'ai co-encadrée.

Systèmes d'Information Scientifique (depuis 2008)
Les systèmes d'Information scientifique sont caractérisés par :
– de grandes collections de données ;
– des données hétérogènes provenant de sources multiples et évoluant rapidement ;
– une grande variabilité inter-acteurs et inter-domaines ;
– un objectif de production de connaissance.

Ces systèmes ont été étudiés dans les domaines des Sciences Humaines et Sociales et les Sciences du Vivant.

4 Animation scientifique

Dans le cadre de mes activités de recherche sur les Systèmes d'Information Scientifique, les quatre premiers projets ont permis de développer l'utilisation d'une ontologie comme pivot pour l'accès a des jeux de données et des mécanismes de contrôle d'annotation par une ontologie. Ils ont fait l'objet de financement institutionnel. Dans le cadre de mes activités de recherche sur l'ingénierie de plate-formes, le projet européen OpenTTT a permis d'étudier l'opportunité pour des PME/PMI d'utiliser des briques logicielles libres dans leur Systèmes d'Information avec des problématiques d'interopérabilité, de support et de perennité *via* la mise en place des communauté du Logiciel Libre.

Atlas historique et technique de la pierre à bâtir bourguignonne (2013)
L'objet de notre participation est le développement d'un Système d'Information traitant la nature et l'usage de la pierre bourguignonne à partir des données fournies par le Bureau de Recherches Géologiques et Minières. Ce projet est financé par la Région Bourgogne (CPER) et l'Union Européenne (FEDER).

PEPS HuMaIn : Logiques modales pour le traitement de l'incertitude des données archéologiques (2013)
La modélisation et le raisonnement sur les données incertaines sont des éléments incontournables des SIS du fait de la nature des recherches, de l'implication d'équipes pluridisciplinaires et multiculturelles. L'incertitude dépend du contexte d'acquisition ou d'utilisation des données et doit par conséquent être reliée à la connaissance du domaine. Dans le cadre de ce PEPS, nous souhaitons nous concentrer sur la problématique de la modélisation des données incertaines ainsi que sur les modes de raisonnement associés. L'objectif est d'élaborer une modélisation de l'incertitude au moyen d'annotations afin de prendre en compte la notion d'incertitude spécifique à chaque spécialité. Pour ce faire, nous avons besoin de logiques modales couplées avec une logique probabiliste [DS07].

ANR CARE (2008-2011) *Corpus Architecturae Religiosae Europeae* (ANR-07-CORP-011)
L'ANR CARE réunissait plus de soixante archéologues en poste dans une vingtaine d'universités, des dessinateurs topographe, le pôle de géomatique de la MSH de Dijon et notre équipe. L'objet de notre participation était le développement d'une plate-forme collaborative gérant des connaissances relatives au corpus européen des édifices religieux du IV^e siècle au X^e siècle. Cette plate-forme offre les outils nécessaires au travail de synthèse sur le référencement des édifices religieux et sur leurs évolutions au cours des siècles à travers un modèle spatio-temporel spécifique. La connaissance des archéologues est modélisée à travers une ontologie d'application qui spécialise une ontologie de domaine. L'application est basée sur MediaWiki que nous avons étendu afin d'y intégrer la sémantique des domaines impliqués et la démarche scientifique spécifique à la discipline. Cette plate-forme s'inscrit dans le cadre du Web 2.0 avec l'utilisation des recommandations du Web Sémantique (RDF, OWL, SPARQL). Les aspects contributifs sont couverts par le wiki et les aspects sémantiques sont couverts par les annotations et la modélisation de la connaissance du domaine (http://care.tge-adonis.fr).

Projet national I3-CRB (2009-2010) *Infrastructure Informatique Interopérable pour les Centres de Ressources Biologiques* financé par le GIS IBiSA.

Cette collaboration nationale, entre plusieurs Centres de Ressources Biologiques (CRB), l'Institut de Biologie et Chimie des Protéines (IBCP) et notre équipe, a pour objectif la mise en place d'une infrastructure Web 2.0 à l'échelle nationale avec et pour les CRB. Cette infrastructure, mise en place par l'IBCP et nous, permet la pérennisation, la standardisation, l'intégration, la mise à disposition et l'exploitation des données biologiques acquises par les CRB. La stratégie proposée repose sur une architecture distribuée autour des CRB (http://www.i3crb.fr).

Projet Européen OpenTTT (2006-2008) *Open Transnational Technological Transfer* (SSA-030595 INN7)
OpenTTT était un projet de l'ARIST, créé à l'initiative de la Commission Européenne en partenariat avec les CCI de Bourgogne. L'objectif était de déterminer les processus qui permettent la construction de Logiciels Libres innovants et de qualité. Notre équipe a joué le rôle d'expert scientifique et a été responsable du workpackage « État de l'art sur les offres Open Source ». Ce projet a donné lieu a une publication ainsi qu'un délivrable, coordonné par Éric Leclercq et moi-même, composé d'une synthèse des offres de logiciels libres par domaine ainsi qu'une méthode d'ingénierie spécifique au Logiciel Libre (regroupant entreprises clientes, SSLL et communautés du libre) a été produit.

5 Relation avec le monde industriel ou socio-économique - Transfert de technologie

Rapport d'expertise dans le cadre du diagnostic Welience[1] (Prestation de Conseil Technologique) pour la société Teletech International en 2011. Ce diagnostic portait sur la gestion des connaissances et les réseaux sociaux dans le domaine de la gestion de la relation client pour une société de centre d'appel.

Transferts de technologie pour l'Office de Coopération et Information Muséographique (OCIM) entre 2004 et 2007. Dans ce cadre, j'ai participé à la rédaction des cahiers des charges pour des applications internet/intranet : « Réalisation d'un carnet d'adresses électroniques » et « Refonte de leur site internet ».

Audit du service informatique d'une banque régionale entre mai et juillet 2004

Formation pour la Fondation Transplantation en décembre 2002. Cette formation de six jours traitait de l'administration d'Oracle 9i en environnement distribué. Notre équipe continue de collaborer avec la Fondation Transplantation, en 2010 une thèse CIFRE a été soutenue.

Projet PANDA (Point d'Accès Numérique de Dijon et de son Agglomération) : définition de l'architecture, aide à la mise en place (plusieurs contrats de 2004 à 2006).

6 Collaborations scientifiques

Dans le domaine des Sciences du Vivant et de la recherche médicale :

Centre de Ressources Biologiques Ferdinand Cabanne de Dijon avec le projet I3-CRB
Plateforme de protéomique CLIPP (Clinical & Innovation Proteomic Platform) depuis 2007
Centre des Sciences du Goût et de l'Alimentation depuis 2005
Société ASA (Advanced Solutions Accelerator) http://www.advancedsolutionsaccelerator.com/

1. Welience est une marque d'uB-Filiale, filiale de valorisation de la recherche de l'Université de Bourgogne

Dans le domaine des Sciences Humaines et Sociales :
Laboratoire ARTeHIS (laboratoire Archéologie, Terre, Histoire et Sociétés) UMR 6298 Université de Bourgogne, avec l'ANR CARE

7 Comité de programme

Conférence internationale DEXA (Database and Expert Systems Applications) de 2000 à 2007 et depuis 2009

Conférence internationale IDEAS (Database Engineering & Applications Symposium) en 2011 et 2012

Conférence internationale SITIS (Signal-Image Technology & Internet-Based Systems) :
– Track Open Source OSSDS (Software Development and Solutions) en 2008 et 2009
– Track IBCS (Internet-Based Computing and Systems) en 2010 et 2012
– Workshop on Web Based and Distributed Information Systems (WEBDIS) en 2012

Conférence nationale INFORSID (INFormatique des Organisations et Systèmes d'Information de Décision) en 2003 et 2013.

Rencontres Mondiales du Logiciel Libre (RMLL) en 2005
– Responsable de la session Base de Données

8 Comité d'organisation

Rencontres Mondiales du Logiciel Libre (RMLL) en 2005 (plus de 1000 participants et 150 conférences)

Conférence internationale SITIS (Signal-Image Technology & Internet-Based Systems) en 2011

9 Activité d'enseignement

Activités statutaires

Tout au long de ma carrière, j'ai enseigné l'informatique à tous les niveaux du L1 au M2 professionnel pour environ 255 EqTD par an

Responsabilité pédagogiques

L3 Informatique Responsable de l'U.E Bases de Données depuis 1996

L3 Informatique Responsable de l'U.E. Modélisation Orientée Objets depuis 2009

Licence Professionnelle SIL Internet/Intranet pour l'entreprise Responsable de la partie Bases de données dans l'U.E. Interface de bases de données et web dynamique

Licence Professionnelle SIL Internet/Intranet pour l'entreprise Responsable de la partie UML dans l'U.E. Génie logiciel et modélisation objet

Thématiques abordées

En Bases de données, j'enseigne :
– les aspects fondamentaux des Bases de Données : modèle relationnel, langages de description et de manipulation de données
– le langage PL/SQL
– le lien entre un diagramme de classe UML et une Base de Données Relationnelles.

Le projet noté consiste à évaluer la cohérence entre un diagramme de classes (réalisé durant le projet de l'U.E. Modélisation) et le schéma relationnel de la base de données, à appliquer le principe de la normalisation afin de constituer une Base de Données et de

proposer un jeu de requêtes et d'écrire quelques fonctionnalités en PL/SQL. La mise en œuvre est faite avec Oracle10g.

L'objectif principal en Modélisation Orientée Objets est d'initier les étudiants à l'utilisation d'UML. Au delà de l'aspect purement syntaxique des diagrammes étudiés, l'accent est mis sur deux points qui sont le lien entre la description informelle du domaine à modéliser et la qualité des modèles. Les étudiants doivent relever dans le texte descriptif proposé les éléments qui relèvent des différents diagrammes (essentiellement les diagrammes Use Case, de classe, statechart et séquence complétés par des contraintes OCL) et choisir comment seront modélisés ces éléments. Une présentation des réseaux sémantiques et des frames qui mettent en perspective les concepts actuels de la modélisation objet est réalisée.

Le projet noté consiste à rechercher de la documentation sur le sujet proposé (par exemple modélisation d'Espace Public Numérique, d'une cuisine centrale, d'un Centre de Ressources Biologiques, d'une entreprise de transport routier, etc.) afin d'aboutir à une bibliographie. Celle-ci doit ensuite être exploitée pour aboutir à une description des fonctionnalités, de la structure et du comportement du système. Une partie cohérente du projet doit être choisie en vue d'une implémentation sous Oracle.

J'assure les Travaux Pratiques de l'U.E. Systèmes d'Information Réparties : Bases de Données Distribuées avec Oracle, Base de Données NoSQL avec Neo4j.

TICE

De février 2001 à 2006, réponse à différents appels d'offres faits par l'Université de Bourgogne. Ces réponses ont donné lieu à la mise en place d'un serveur pédagogique et à l'utilisation de la plate-forme d'e-learning Ganesha de la société Anemalab.

Culture scientifique

Trésorière du Centre de Culture Scientifique Technique et Industrielle de Bourgogne (CCSTIB) depuis 2009

10 Administration

Membre du conseil de l'UFR Sciences et Techniques depuis janvier 2007

Élue au bureau de l'UFR Sciences et Techniques de novembre 2009 à mars 2013

Membre du conseil de direction du Laboratoire LE2I de 2005 à 2009

Chef de projet fonctionnel « Logiciel Services » à l'Université de Bourgogne depuis 2003. Ce logiciel a pour objectif l'aide à la construction du service des différents intervenants de l'Université (enseignants sur poste comme vacataires). Il automatise aussi les demandes de mise en paiement des heures complémentaires. Dans le cadre du contrat quadriennal 2012-2015, participation à l'élaboration du cahier des charges du logiciel EVALENS. EVALENS permet une évaluation de la charge d'enseignement des formations de l'Université des Bourgogne. Ces deux logiciels sont des outils de pilotage pour l'Université.

Membre du Comité Hygiène Sécurité Environnement de l'Université de Bourgogne (1999- 2005)

Membre de la Commission de Spécialistes $27^{ème}$ section en tant que membre titulaire pour la période 2001-2004 et ensuite en tant que suppléante

Membre du comité de sélection sur le poste n^o 1269 de l'UFR Sciences et Techniques de l'Université de Bourgogne en 2009

Bibliographie

[ABC03] Tiphaine Accary, Aurélien Bénel et Sylvie Calabretto : Modélisation de connaissances temporelles en Archéologie. *In Actes des Journées francophones d'Extraction et de Gestion des Connaissances [EGC'2003]*, volume 17 de *Revue des Sciences et Technologies de l'Information (RSTI)*, pages 503–508. Hermès-Lavoisier, 2003.

[ABK+10] Yanif Ahmad, Randal C. Burns, Michael M. Kazhdan, Charles Meneveau, Alexander S. Szalay et Andreas Terzis : Scientific data management at the Johns Hopkins institute for data intensive engineering and science. *SIGMOD Record*, 39(3):18–23, 2010.

[AC04] Tiphaine Accary et Sylvie Calabretto : La temporalité des corpus archéologiques. *In Document Numérique*, volume 8, pages 111–124. Hermès-Lavoisier, 2004.

[ACP10] Waseem Akhtar, Álvaro Cortés Calabuig et Jan Paredaens : Constraints in RDF. *In Semantics in Data and Knowledge Bases (SDKB)*, pages 23–39, 2010.

[AG05] Nathalie Aussenac-Gilles : *Méthodes ascendantes pour l'ingénierie des connaissances*. Thèse de doctorat, Université de Toulouse III - Paul Sabatier, 2005.

[AG06] Nathalie Aussenac-Gilles : Ontology or meta-model for Retrieving Scientific Reasoning in Documents : the Arkeotek project. *In Workshop on Exploring the limits of global models for integration and use of historical and scientific information*, 2006.

[AKD10] Anastasia Ailamaki, Verena Kantere et Debabrata Dash : Managing Scientific Data. *Communications of ACM*, 53(6):68–78, 2010.

[AKTV08] Anupriya Ankolekar, Markus Krötzsch, Thanh Tran et Denny Vrandecic : The two cultures : Mashing up Web 2.0 and the Semantic Web. *Journal of Web Semantics : Science, Services and Agents on the World Wide Web*, 6(1):70–75, 2008.

[AL07] Sören Auer et Jens Lehmann : What Have Innsbruck and Leipzig in Common ? Extracting Semantics from Wiki Content. *In 4th European Semantic Web Conference (ESWC)*, pages 503–517, 2007.

[AL10] Elie Abi-Lahoud : *Composition dynamique de services : application à la conception et au développement de systèmes d'information dans un environnement distribué*. Thèse de doctorat, Université de Bourgogne, 2010.

[All83] James F. Allen : Maintaining Knowledge about Temporal Intervals. *Communications of ACM*, 26(11):832–843, 1983.

[Ama07] Florence Amardheil : *Web sémantique et informatique linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle*. Thèse de doctorat, Université Paris X - Nanterre, 2007.

[aNDO05] Gábor Nagypál, Richard Deswarte et Jan Oosthoek : Applying the Semantic Web : the VICODI Experience in Creating Visual Contextualization for History. *Literary and Linguistic Computing*, 20(3):327–349, 2005.

[ARF11] Mauricio Almeida, Souza Renato et Fonseca Fred : Semantics in the Semantic Web : a critical evaluation. *Knowledge Organization Journal*, 38(3):187–203, 2011.

[AWH92] Alexander Aiken, Jennifer Widom et Joseph M. Hellerstein : Behavior of Database Production Rules : Termination, Confluence, and Observable Determinism. *In SIGMOD Conference*, pages 59–68, 1992.

[Bac00] Bruno Bachimont : L'intelligence artificielle comme écriture dynamique : de la raison graphique à la raison computationnelle. *Au nom du sens*, pages 290–319, 2000.

[Bac04] Bruno Bachimont : *Arts et Sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle*. Thèse de doctorat, Université de Technologie de Compiègne, 2004.

[BB03] Alexander Borgida et Ronald J. Brachman : Conceptual Modeling with Description Logics. *In The Description Logic Handbook, Theory, Implementation and Applications*, pages 349–372, 2003.

[BBC04] Sandra Bringay, Catherine Barry et Jean Charlet : Les documents et les annotations du dossier patient hospitalier. *Information - Interaction - Intelligence*, 4(1), 2004.

[BBDH08] Olivier Biton, Sarah Cohen Boulakia, Susan B. Davidson et Carmem S. Hara : Querying and Managing Provenance through User Views in Scientific Workflows. *In 24th International Conference on Data Engineering (ICDE)*, pages 1072–1081, 2008.

[BBP+08] Uldis Bojars, John G. Breslin, Vassilios Peristeras, Giovanni Tummarello et Stefan Decker : Interlinking the Social Web with Semantics. *IEEE Intelligent Systems*, 23(3):29–40, 2008.

[BCG+10] Jean-François Baget, Madalina Croitoru, Alain Gutierrez, Michel Leclère et Marie-Laure Mugnier : Translations between RDF(S) and Conceptual Graphs. *In* Madalina Croitoru, Sébastien Ferré et Dickson Lukose, éditeurs : *Conceptual Structures : From Information to Intelligence, 18th International Conference on Conceptual Structures (ICCS)*, volume 6208 de *Lecture Notes in Computer Science*, pages 28–41. Springer, 2010.

[BCM+03] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi et Peter F. Patel-Schneider, éditeurs. *The Description Logic Handbook : Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[BCTV05] Deepavali Bhagwat, Laura Chiticariu, Wang Chiew Tan et Gaurav Vijayvargiya : An annotation management system for relational databases. *VLDB Journal*, 14(4):373–396, 2005.

[BDI+08] Randal C. Burns, Susan B. Davidson, Yannis E. Ioannidis, Miron Livny et Jignesh M. Patel : Scientific Data Management : An Orphan in the Database Community ? *In 24th International Conference on Data Engineering (ICDE)*, page 9, 2008.

[BDL+98] Pierre Bessière, Éric Dedieu, Olivier Lebeltel, Emmanuel Mazer et Kamel Mekhnacha : Interprétation ou Description (I) : Proposition pour une théorie probabiliste des systèmes sensori-moteurs. *Intellectica*, 1-2(26-27):257–311, 1998.

[BFZK10] Michel Buffa, Catherine Faron-Zucker et Anna Kolomoyskaya : Gestion sémantique des droits d'accès au contenu : l'ontologie AMO. *In Extraction et gestion des connaissances (EGC)*, pages 471–482, 2010.

[BGE+08] Michel Buffa, Fabien L. Gandon, Guillaume Erétéo, Peter Sander et Catherine Faron : SweetWiki : A semantic wiki. *Journal of Web Semantics*, 6(1):84–97, 2008.

[Bin10] Ceri Binding : Implementing Archaeological Time Periods Using CIDOC CRM and SKOS. *In Extended Semantic Web Conference, Part I, LNCS 6088 (ESWC)*, pages 273–287, 2010.

[BJ12] Gianpietro Brogiolo et Miljenko Jurkovicn : Corpus Architecturae Religiosae Europae (IV-X Saec.) Introduction. *HORTUS ARTIUM MEDIEVALIUM*, 18(1): 7–26, 2012.

[BJR07] Neerja Bhatnagar, Benjoe Juliano et Renee Renner : Data Annotation Models and Annotation Query Language. *In International Conference on Knowledge Mining (ICKM)*, pages 440–445, 2007.

[BKvH02] Jeen Broekstra, Arjohn Kampman et Frank van Harmelen : Sesame : A Generic Architecture for Storing and Querying RDF and RDF Schema. *In International Semantic Web Conference*, pages 54–68, 2002.

[BLHL01] Tim Berners-Lee, James Hendler et Ora Lassila : The Semantic Web. *Scientific American Magazine*, 284(5):34–43, 2001.

[BLS+00] Djamal Benslimane, Eric Leclercq, Marinette Savonnet, Marie-Noëlle Terrasse et Kokou Yétongnon : Ont le Definition of generic multilayered ontologies for urban applications. *Computers, Environment and Urban Systems*, 24:191–214, 2000.

[BMPV08] Andrea Bonomi, Alessandro Mosca, Matteo Palmonari et Giuseppe Vizzari : Integrating a Wiki in an Ontology Driven Web Site : Approach, Architecture and Application in the Archaeological Domain. *In Third Workshop on Semantic Wikis - The Wiki Way of Semantics (SemWiki)*, 2008.

[BMV06] Andrea Bonomi, Glauco Mantegari et Giuseppe Vizzari : A Framework for Ontological Description of Archaeological Scientific Publications. *In Semantic Web Applications and Perspectives (SWAP)*, 2006.

[Bor97] Willem Nico Borst : *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. Thèse de doctorat, Twente University, 1997.

[Bor98] Andrée Borillo : *L'espace et son expression en français*. L'essentiel français, Ophrys, 1998.

[BS09] Cristina Bogdanschi et Simone Santini : An annotation database for multimodal scientific data. *In Proc. SPIE, Electronic Imaging Science and Technology Symposium*, volume 7255, 2009.

[BT10] Nicole Bidoit-Tollu : Types and Constraints : From Relational to XML Data. *In 4th International Workshops Semantics in Data and Knowledge Bases (SDKB)*, pages 40–53, 2010.

[Bun77] Mario Bunge : Treatise on Basic Philosophy. Ontology I : The Furniture of the World. Rapport technique, Riedel, 1977.

[Bun00] Horst Bunke : Graph Matching : Theoretical Foundations, Algorithms, and Applications. *Vision Interface*, pages 82–88, 2000.

[Bun09] Peter Buneman : Curated databases. *In European Conference on Digital Libraries (ECDL)*, page 2, 2009.

[BWF+00] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov et Philip E. Bourne : The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[Büt11] Stéphane Büttner : Atlas historique et technique de la pierre à bâtir bourguignonne. plateforme numérique mutualiste sur la nature et l'usage de la pierre. *Bulletin du centre d'études médiévales d'Auxerre*, 15(1):351–355, 2011.

[CC02] CIDOC-CRM : Définition du Modèle Conceptuel de Référence du CIDOC(CRM) - Version 3.4. Rapport technique, International Council of Museum, 2002.

[CC03] Jake Yue Chen et John V. Carlis : Genomic data modeling. *Journal Information Systems - Special issue : Data management in bioinformatics*, 28(4):287–310, 2003.

[CGQ+08] Dolors Costal, Cristina Gómez, Anna Queralt, Ruth Raventós et Ernest Teniente : Improving the definition of general constraints in UML. *Software and System Modeling*, 7(4):469–486, 2008.

[CHLBS98] J. M. Cardot, T. Hulot, C. Le Bricon et A. Stockis : LIMS : from theory to practice. *European journal of drug metabolism and pharmacokinetics*, 23(2):207–212, 1998.

[CID94] CIDOC : The International Committee for Documentation of the International Council of Museum (ICOM-CIDOC). Rapport technique, International Council of Museum, 1994.

[CM09] Michel Chein et Marie-Laure Mugnier : *Graph-based Knowledge Representation : Computational Foundations of Conceptual Graphs*. Springer, 2009.

[CS08] Pascale Chevalier et Christian Sapin : ANR Corpus Architecturae Religiosae Europeae [CARE], saec. IV-X. Rapport technique, Centre d'études médiévales d'Auxerre, 2008.

[CS12] Pascale Chevalier et Christian Sapin : Les avancées du corpus CARE en France (2008-2011). *HORTUS ARTIUM MEDIEVALIUM*, 18(1):85–96, 2012.

[CTV05] Laura Chiticariu, Wang Chiew Tan et Gaurav Vijayvargiya : DBNotes : a Post-It System for Relational Databases based on Provenance. *In ACM SIGMOD International Conference on Management of Data*, pages 942–944, 2005.

[Dal09] Costis Dallas : From Artefact Typologies to Cultural Heritage Ontologies : or, an Account of the Lasting Impact of Archaeological Computing. *Archeologia e Calcolatori*, 20:205–221, 2009.

[DBB09] Véronique Dupierris, Damien Barthe et Christophe Bruley : ePIMS : un LIMS pour la gestion des données de spectrométrie de masse. *Spectra Analyse*, 38(269):36–40, 2009.

[DHM+08] Allen Dreibelbis, Eberhard Hechler, Ivan Milman, Martin Oberhofer, Paul van Run et Dan Wolfson : *Enterprise Master Data Management : An SOA Approach to Managing Core Information*. IBM Press, 1ère édition, 2008.

[DKS04] Martin Doerr, Athina Kritsotaki et Stephen Stead : Which Period is it ? A Methodology to Create Thesauri of Historical Periods. *In Computer Applications and quantitative methods in Archaeology (CAA)*, 2004.

[DKS05] Martin Doerr, Athina Kritsotaki et Stephen Stead : Thesauri of Historical Periods - A Proposal for Standardization. *In CIDOC Conference*, 2005.

[DLNS98] Francesco M. Donini, Maurizio Lenzerini, Daniele Nardi et Andrea Schaerf : AL-log : Integrating Datalog and Description Logics. *Journal of Intelligent Information Systems*, 10(3):227–252, 1998.

[DOB95] Susan Davidson, Chris Overton et Peter Buneman : Challenges in Integrating Biological Data Sources. *Journal of Computational Biology*, 2(4):557–572, 1995.

[Doe03] Martin Doerr : The CIDOC CRM - An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, 24:75–92, 2003.

[DOS05] Michael Daconta, Leo Obrst et Kevin Smith : *The Semantic Web : A guide to the Future of XML, Web Services and Knowledge Management*. Willey, 2005.

[DPKB04] Martin Doerr, Dimitris Plexousakis, Katerina Kopaka et Chryssoula Bekiari : Supporting Chronological Reasoning in Archaeology. *In Computer Applications and quantitative methods in Archaeology (CAA)*, 2004.

[DS07] Nilesh N. Dalvi et Dan Suciu : Management of probabilistic data : foundations and challenges. *In Proceedings of the Twenty-Sixth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 1–12, 2007.

[DSSS09] Renata Queiroz Dividino, Sergej Sizov, Steffen Staab et Bernhard Schueler : Querying for provenance, trust, uncertainty and other meta knowledge in RDF. *Journal of Web Semantics*, 7(3):204–219, 2009.

[EAE+09] Mohamed Y. Eltabakh, Walid G. Aref, Ahmed K. Elmagarmid, Mourad Ouzzani et Yasin N. Silva : Supporting Annotations on Relations. *In 12th International Conference on Extending Database Technology (EDBT)*, pages 379–390, 2009.

[EHBN06] Brian Elvesæter, Axel Hahn, Arne-Jørgen Berre et Tor Neple : Towards an Interoperability Framework for Model-Driven Development of Software Systems. *In* Dimitri Konstantas, Jean-Paul Bourrières, Michel Léonard et Nacer Boudjlida, éditeurs : *Interoperability of Enterprise Software and Applications*, pages 409–420. Springer London, 2006.

[EOA07] Mohamed Y. Eltabakh, Mourad Ouzzani et Walid G. Aref : bdbms - A Database Management System for Biological Data. *In Third Biennial Conference on Innovative Data Systems Research (CIDR)*, pages 196–206, 2007.

[EOA+08] Mohamed Y. Eltabakh, Mourad Ouzzani, Walid G. Aref, Ahmed K. Elmagarmid, Yasin Laura-Silva, Muhammad U. Arshad, David Salt et Ivan Baxter : Managing Biological Data using bdbms. *In 24th International Conference on Data Engineering (ICDE)*, pages 1600–1603, 2008.

[Eva04] Eric Evans : *Domain-Driven Design : Tackling Complexity in the Heart of Software*. Addison Wesley Professional, 2004.

[EW05] Joerg Evermann et Yair Wand : Ontology based object-oriented domain modelling : fundamental concepts. *Requirements Engineering*, 10(2):146–160, 2005.

[EZPMP00] Elöd Egyed-Zsigmond, Yannick Prié, Alain Mille et Jean-Marie Pinon : A graph based audio-visual document annotation and browsing system. *In 6th International Conference Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications - RIAO)*, pages 1381–1289, 2000.

[FDHM+05] Hélène Ferry-Dumazet, Gwenn Houel, Pierre Montalent, Luc Moreau, Olivier Langella, Luc Negroni, Delphine Vincent, Céline Lalanne, Antoine de Daruvar, Christophe Plomion, Michel Zivy et Johann Joets : PROTICdb : A web-based application to store, track, query, and compare plant proteome data. *Proteomics*, 5(8):2069–2081, 2005.

[FJP90] James French, Anita Jones et John Pfaltz : Scientific Database Management. Rapport technique, Université de Virginie, États-Unis, 1990.

[Fra95] Frederick K. Frantz : A Taxonomy of Model Abstraction Techniques. *In Winter Simulation Conference*, pages 1413–1420, 1995.

[Für02] Frédéric Fürst : L'ingénierie ontologique. Rapport technique, Institut de recherche en informatique de Nantes, 2002.

[Gan06] Fabien Gandon : Le web sémantique n'est pas antisocial. *In Actes d'IC*, pages 131–140, 2006.

[GDE+07] Yolanda Gil, Ewa Deelman, Mark H. Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole A. Goble, Miron Livny, Luc Moreau et Jim Myers : Examining the Challenges of Scientific Workflows. *IEEE Computer*, 40(12):24–32, 2007.

[GFS12] Michael Y. Galperin et Xosé M. Fernández-Suarez : The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*, 40(Database-Issue):1–8, 2012.

[Gho10] Debasish Ghosh : Multiparadigm Data Storage for Enterprise Applications. *IEEE Software*, 27:57–60, 2010.

[GHV07] Claudio Gutierrez, Carlos A. Hurtado et Alejandro A. Vaisman : Introducing Time into RDF. *IEEE Transactions on Knowledge and Data Engineering*, 19(2):207–218, 2007.

[GKT07] Todd J. Green, Gregory Karvounarakis et Val Tannen : Provenance semirings. *In PODS*, pages 31–40, 2007.

[GLNS+05] Jim Gray, David T. Liu, Maria Nieto-Santisteban, Alex Szalay, David J. DeWitt et Gerd Heber : Scientific data management in the coming decade. *SIGMOD Record*, 34(4):34–41, 2005.

[Gou99] Jean-Pierre Goulette : Sémantique formelle de l'espace. une application au raisonnement spatial qualitatif en architecture. *Intellectica*, 29(2):9–34, février 1999.

[Gra12] Ludovic Granjon : Application cartographique en ligne du projet CARE : Principes et fonctionnement. *HORTUS ARTIUM MEDIEVALIUM*, 18(1):37–43, 2012.

[Gru93] Thomas Gruber : A Translation Approach to Portable Ontologies. *Knowledge Acquisition*, 5(2):199–220, 1993.

[Gru08] Tom Gruber : Collective knowledge systems : Where the Social Web meets the Semantic Web. *Journal of Web Semantics : Science, Services and Agents on the World Wide Web*, 6(1):4–13, 2008.

[GS04] Pierre Grenon et Barry Smith : SNAP and SPAN : Towards Dynamic Spatial Ontology. *Spatial Cognition and Computation*, 4(1):69–104, 2004.

[HFK10] Frederik Hogenboom, Flavius Frasincar et Uzay Kaymak : A Review of Approaches for Representing RCC8 in OWL. *In ACM Symposium on Applied Computing (SAC)*, pages 1444–1445, 2010.

[HH91] Max Hegenhofer et John Herring : Categorizing Binary Topological Relations Between Regions, Lines and Points in Geographic Databases. Rapport technique, National Center for Geographic Information and Analysis, CA, 1991.

[HLS05] Anja Haake, Stephan Lukosch et Till Schümmer : Wiki-templates : adding structure support to wikis on demand. *In Int. Sym. Wikis*, pages 41–51, 2005.

[HM08] Tom Heath et Enrico Motta : Ease of interaction plus ease of integration : Combining Web2.0 and the Semantic Web in a reviewing site. *Journal of Web Semantics : Science, Services and Agents on the World Wide Web*, 6(1):76–83, 2008.

[HO09] Jon Holmen et Christian-Emil Ore : Deducing Event Chronology in a Cultural Heritage Documentation System. *In Computer Applications and quantitative methods in Archaeology (CAA)*, 2009.

[HPS04] Ian Horrocks et Peter F. Patel-Schneider : A proposal for an OWL rules language. *In 13th international World Wide Web Conference (WWW)*, pages 723–731, New York, NY, USA, 2004.

[HPSBT05] Ian Horrocks, Peter F. Patel-Schneider, Sean Bechhofer et Dmitry Tsarkov : OWL rules : A proposal and prototype implementation. *Web Semantics : Science, Services and Agents on the World Wide Web*, 3:23–40, 2005.

[HT03] Tony Hey et Anne Trefethen : The Data Deluge : An e-Science Perspective. pages 809–824, 2003.

[HTT09] Tony Hey, Stewart Tansley et Kristin Tolle, éditeurs. *The Fourth Paradigm : Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009.

[Huv09] Isto Huvila : Steps towards a participatory digital library and data archive for archaeological information. *In 10th Libraries in the Digital Age (LIDA) Conference*, pages 149–159, 2009.

[Huv12] Isto Huvila : Being Formal and Flexible : Semantic Wiki as an Archaeological e-Science Infrastructure. *In* Mingquan Zhou, Iza Romanowska, Zhongke Wu, Pengfei Xu et Philip Verhagen, éditeurs : *Revive the Past : Proceeding of the 39th Conference on Computer Applications and Quantitative Methods in Archaeology, Beijing, 12-16 April 2011*, pages 186–197. Amsterdam University Press, 2012.

[HWL12] Li-Yung Ho, Jan-Jan Wu et Pangfeng Liu : Distributed Graph Database for Large-Scale Social Computing. *In IEEE Fifth International Conference on Cloud Computing (CLOUD)*, pages 455–462, 2012.

[IAJA11] Stratos Idreos, Ioannis Alagiannis, Ryan Johnson et Anastasia Ailamaki : Here are my Data Files. Here are my Queries. Where are my Results ? *In Fifth Biennial Conference on Innovative Data Systems Research (CIDR)*, pages 57–68, 2011.

[IGN+12] Stratos Idreos, Fabian Groffen, Niels Nes, Stefan Manegold, K. Sjoerd Mullender et Martin L. Kersten : MonetDB : Two Decades of Research in Column-oriented Database Architectures. *IEEE Data Engineering Bulletin*, 35(1):40–45, 2012.

[IKM12] M. Ivanova, M. L. Kersten et S. Manegold : Data Vaults : A Symbiosis Between Database Technology And Scientific File Repositories. *In International Conference on Scientific and Statistical Database Management*, pages 1233–1236, 2012.

[INGK07] Milena Ivanova, Niels Nes, Romulo Goncalves et Martin L. Kersten : MonetDB/SQL Meets SkyServer : the Challenges of a Scientific Database. *In 19th International Conference on Scientific and Statistical Database Management (SSDBM)*, page 13. IEEE Computer Society, 2007.

[IVZ08] Angelo Di Iorio, Fabio Vitali et Stefano Zacchiroli : Wiki Content Templating. *In 17th International Conference on World Wide Web (WWW)*, pages 615–624, 2008.

[IZ06] Angelo Di Iorio et Stefano Zacchiroli : Constrained Wiki : an Oxymoron ? *In Int. Sym. Wikis*, pages 89–98, 2006.

[JHX+07] Stéphane Jean, Dehainsala Hondjack, Dung Nguyen Xuan, Guy Pierra, Ladjel Bellatreche et Yamine Aït Ameur : OntoDB : It Is Time to Embed Your Domain Ontology in Your Database. *In Conference on Database Systems for Advanced Applications (DASFAA)*, pages 1119–1122, 2007.

[JMA+07] A.R. Jones, M. Miller, R. Aebersold, R. Apweiler et al. : The Functional Genomics Experiment model (FuGE) : an extensible framework for standards in functional genomics. *Nature Biotechnology*, 25(10):1127–1133, 2007.

[JO04] H. V. Jagadish et Frank Olken : Database Management for Life Sciences Research. *SIGMOD Record*, 33(2):15–20, 2004.

[JZM08] Ke Jiang, Lei Zhang et Shigeru Miyake : Using OCL in Executable UML. *Electronic Communications of the EASST- ECEASST*, 9, 2008.

[KBA02] Ivan Kurtev, Jean Bézivin et Mehmet Aksit : Technological Spaces : an Initial Appraisal. *In International Symposium on Distributed Objects and Applications (DOA), CoopIS Federated Conferences, Industrial track*, 2002.

[KG12] Grigoris Karvounarakis et Todd J. Green : Semiring-annotated data : queries and provenance ? *SIGMOD Record*, 41(3):5–14, 2012.

[KGSS12] Matthias Konrath, Thomas Gottron, Steffen Staab et Ansgar Scherp : SchemEX - Efficient Construction of a Data Catalogue by Stream-based Indexing of Linked Data. *Web Semantics : Science, Services and Agents on the World Wide Web*, 0(0), 2012.

[KPT+04] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov et Damyan Ognyanoff : Semantic annotation, indexing, and retrieval. *Web Semantics : Science, Services and Agents on the World Wide Web*, 2(1):49 – 79, 2004.

[KRH08] Markus Krötzsch, Sebastian Rudolph et Pascal Hitzler : ELP : Tractable Rules for OWL 2. *In* A. P. Sheth, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy Finin et Krishnaprasad Thirunarayan, éditeurs : *The Semantic Web - ISWC 2008*, volume 5318 de *Lecture Notes in Computer Science*, pages 649–664. Springer Heidelberg, 2008.

[KRS10] Markus Krötzsch, Sebastian Rudolph et Peter H. Schmitt : On the Semantic Relationship between Datalog and Description Logics. *In Fourth International Conference Web Reasoning and Rule Systems (RR)*, volume 6333 de *Lecture Notes in Computer Science*, pages 88–102. Springer, 2010.

[KVV06] Markus Krötzsch, Denny Vrandecic et Max Völkel : Semantic MediaWiki. *In International Semantic Web Conference*, pages 935–942, 2006.

[Lin92] Marc Linster : Viewing Knowledge Engineering as a Symbiosis of *Modeling to Make Sense* and *Modeling to Implement Systems*. *In 6th European Knowledge Acquisition Workshop (GWAI)*, pages 87–99, 1992.

[LLC08] Qinglan Li, Alexandros Labrinidis et Panos K. Chrysanthis : User-Centric Annotation Management for Biological Data. *In Second International Provenance and Annotation Workshop (IPAW)*, volume 5272 de *Lecture Notes in Computer Science*, pages 54–61, 2008.

[LMS08] Georg Lausen, Michael Meier et Michael Schmidt : SPARQLing constraints for RDF. *In 11th international conference on Extending database technology : Advances in database technology*, EDBT, pages 499–509, 2008.

[Loc09] Gary Lock : Archaeological Computing Then and Now : Theory and Practice, Intentions and Tensions. *Archeologia e Calcolatori*, 20:75–84, 2009.

[LP05] Shan Lu et Jeffrey Parsons : Enforcing Ontological Rules in UML-Based Conceptual Modeling : Principles and Implementation. *In CAISE Workshop*, pages 451–462, 2005.

[LPSZ10] Nuno Lopes, Axel Polleres, Umberto Straccia et Antoine Zimmermann : AnQL : SPARQLing Up Annotated RDFS. *In 9th International Semantic Web Conference (ISWC)*, volume 6496 de *Lecture Notes in Computer Science*, pages 518–533, 2010.

[MAB+07] Harald Mischak, Rolf Apweiler, Rosamonde E. Banks, Mark Conaway, Joshua Coon, Anna Dominiczak, Jochen H. H. Ehrich, Danilo Fliser, Mark Girolami, Henning Hermjakob, Denis Hochstrasser, Joachim Jankowski, Bruce A. Julian, Walter Kolch, Ziad A. Massy, Christian Neusuess, Jan Novak, Karlheinz Peter, Kasper Rossing, Joost Schanstra, O. John Semmes, Dan Theodorescu, Visith Thongboonkerd, Eva M. Weissinger, Jennifer E. Van Eyk et Tadashi Yamamoto : Clinical proteomics : A need to define the field and to begin to set adequate standards. *PROTEOMICS - Clinical Applications*, 1(2):148–156, 2007.

[Mar08] Loraine Marchaix : Conception d'une ontologie à partir d'un thésaurus spécialisé dans le domaine de l'archéologie et des sciences de l'antiquité. Rapport technique, Mémoire de Master II professionnel de Gestion de l'Information et du Document, Spécialité Gestion des connaissances, Université Paris 8, 2008.

[MB02] Stephen J. Mellor et Marc Balcer : *Executable UML : a foundation for Model-Driven Architecture*. Addison Wesley, 2002.

[MHS07] Boris Motik, Ian Horrocks et Ulrike Sattler : Bridging the gap between OWL and relational databases. *In 16th International Conference on World Wide Web (WWW)*, pages 807–816, 2007.

[Mir09] Alina Dia Miron : *Découverte d'associations sémantiques pour le Web Sémantique Géospatial : le framework ONTOAST*. Thèse de doctorat, Université Joseph Fourier - Grenoble, 2009.

[MJLP11] Thomas Meilender, Nicolas Jay, Jean Lieber et Fabien Palomares : Les moteurs de wikis sémantiques : un état de l'art. *In Extraction et Gestion de Connaissances (EGC)*, pages 575–580, 2011.

[MN05] Elita Miliauskaite et Lina Nemuraite : Representation of integrity constraints in conceptual models. *Information Technology And Control*, 34(4): 355–365, 2005.

[MPV10] Glauco Mantegari, Matteo Palmonari et Giuseppe Vizzari : Rapid Prototyping a Semantic Web Application for Cultural Heritage : The Case of MANTIC. *In 7th Extended Semantic Web Conference (ESWC)*, pages 406–410, 2010.

[MR10] Boris Motik et Riccardo Rosati : Reconciling Description Logics and Rules. *Journal of the ACM*, 57(5):1–62, 2010.

[MRvdA10] Jan Mendling, Hajo A. Reijers et Wil M. P. van der Aalst : Seven process modeling guidelines (7PMG). *Information and Software Technology*, 52(2):127–136, 2010.

[MSS05] Boris Motik, Ulrike Sattler et Rudi Studer : Query Answering for OWL DL with rules. *Web Semantics*, 3(1):41–60, 2005.

[MVM95] Riichiro Mizogouchi, Johan Vanwelkenhuysen et Ikeda Mitsuru : Task Ontology for Reuse of Problem Solving Knowledge. *2nd International Conference on Very large-Scale Knowledge Bases*, pages 46–59, 1995.

[Nau11] Pierre Naubourg : *Une approche préventive de fusion et d'évolution des données d'un système d'information : application aux données biomédicales.* Thèse de doctorat, Université de Bourgogne, 2011.

[NEO03] Harvey B. Newman, Mark H. Ellisman et John A. Orcutt : Data-intensive e-science frontier research. *Communications of the ACM*, 46(11):68–77, 2003.

[NOVS11] Elena Nardini, Andrea Omicini, Mirko Viroli et Michael I. Schumacher : Coordinating e-health systems with TuCSoN semantic tuple centres. *ACM SIGAPP Applied Computing Review*, 11(2):43–53, 2011.

[NPG07] Shamkant B. Navathe, Upen Patil et Wei Guan : Genomic and Proteomic Databases : Foundations, Current Status and Future Applications. *Journal of Computing Science and Engineering (JCSE)*, 1(1):1–30, 2007.

[NSLY11] Pierre Naubourg, Marinette Savonnet, Éric Leclercq et Kokou Yétongnon : Approche préventive de la qualité des données dans le contexte de la protéomique clinique. *Revue des Nouvelles technologies de l'Information*, E.22:189–234, 2011.

[NV06] Roberto Navigli et Paola Velardi : Enriching a Formal Ontology with a Thesaurus : an application int the Cultural Heritage Domain. *In 2nd Workshop on Ontology Learning and Population (OLP)*, 2006.

[OAF+04] Thomas M. Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, R. Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat et Peter Li : Taverna : a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, 2004.

[OMS+06] Eyal Oren, Knud Möller, Simon Scerri, Siegfried Handschuh et Michael Sintek : What are Semantic Annotations ? Rapport technique, DERI Galway, 2006.

[PBB+03] F.M. Pearl, C.F. Bennett, J.E. Bray, A.P. Harrison, N. Martin, A.Shepherd, I. Sillitoe, J. Thornton et C.A. Orengo : The CATH database : an extended protein family resource for structural and functional genomics. *Nucleic Acids Research*, 31(1), 2003.

[PCB07] Patrick Poullet, Sabrina Carpentier et Emmanuel Barillot : myProMS, a web server for management and validation of mass spectrometry-based proteomic data. *Proteomics*, 7(15):2553–2556, 2007.

[Pfa07] John L. Pfaltz : What constitutes a scientific database ? *In 19th International Conference on Scientific and Statistical Database Management (SSDBM)*, pages 2–11. IEEE Computer Society, 2007.

[Pig08] Christine Piggee : LIMS and the art of MS proteomics. *Analytical Chemistry*, 80(13):4801–4806, 2008.

[PRV09] Alain Plantec, Vincent Ribaud et Vasudeva Varma : Building a Semantic Virtual Museum : from Wiki to Semantic Wiki using Named Entity Recognition. *In Companion to the 24th Annual ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA)*, pages 769–770, 2009.

[RBP08] Jochen Reutelshoefer, Joachim Baumeister et Frank Puppe : Ad-hoc knowledge engineering with semantic knowledge wikis. *In SemWiki*, 2008.

[RC06] Community Research et Development Information Service (CORDIS) : Enterprise interoperability – research roadmap – version 4.0. Rapport technique, European Community, Information Society Technologies, 2006. http://cordis.europa.eu/ist/ict-ent-net/ei-roadmap_en.htm.

[RCC92] David A. Randell, Zhan Cui et Anthony G. Cohn : A Spatial Logic based on Regions and Connection. *In 3rd International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pages 165–176, 1992.

[Ren00] Agnar Renolen : Modelling the Real World : Conceptual Modelling in Spatio-temporal Information System Design. *Transactions in GIS*, 4(1):23–42, 2000.

[Ric09] Julian D. Richards : From Anarchy to Good Practice : the Evolution of Standards in Archeological Computing. *Archeologia e Calcolatori*, 20:27–35, 2009.

[Riv97] Paul Rivett : Conceptual data modelling in an archaeological GIS, 1997.

[RLB+10] Jochen Reutelshoefer, Florian Lemmerich, Joachim Baumeister, Jorit Wintjes et Lorenz Haas : Taking OWL to Athens - Semantic Web technology takes Ancient Greek history to students. *In 7th Extended Semantic Web Conference (ESWC)*, pages 333–347, 2010.

[SAA+08] Guus Schreiber, Alia K. Amin, Lora Aroyo, Mark van Assem, Viktor de Boer, Lynda Hardman, Michiel Hildebrand, Borys Omelayenko, Jacco van Ossenbruggen, Anna Tordai, Jan Wielemaker et Bob J. Wielinga : Semantic annotation and search of cultural-heritage collections : The MultimediaN E-Culture demonstrator. *Journal of Web Semantics*, 6(4):243–249, 2008.

[SAR+07] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel et Suzanna Lewis : The OBO foundry : coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, 2007.

[Sch00] Guido Schimm : Generic Linear Business Process Modeling. *In Workshops on Conceptual Modeling Approaches for E-Business and The World Wide Web and Conceptual Modeling*, pages 31–39, London, UK, 2000. Springer-Verlag.

[Sch07] Satu Elisa Schaeffer : Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[Sim08] Jean-Claude Simon : *Une approche liée à la préservation des propriétés transversales pour construire une offre de services web d'un éco-système d'entreprises.* Thèse de doctorat, Université de Bourgogne, 2008.

[SKB11] Lefteris Sidirourgos, Martin L. Kersten et Peter A. Boncz : SciBORQ : Scientific data management with Bounds On Runtime and Quality. *In Fifth Biennial Conference on Innovative Data Systems Research (CIDR)*, pages 296–301, 2011.

[SKT+10] Christian Stephan, Michael Kohl, Michael Turewicz, Katharina Podwojski, Helmut E. Meyer et Martin Eisenacher : Using Laboratory Information Management Systems as central part of a proteomics data workflow. *Proteomics*, 10(6):1230–1249, 2010.

[Smi05] Barry Smith : Relations in biomedical ontologies. *Genome Biology*, 6(5):46, 2005.

[SMJ02] Peter Spyns, Robert Meersman et Mustafa Jarrar : Data modelling versus ontology engineering. *SIGMOD Record*, 31(4):12–17, 2002.

[SOW84a] Arie Shoshani, Frank Olken et Harry K. T. Wong : Characteristics of Scientific Databases. *In 10th International Conference on Very Large Data Bases*, pages 147–160. Morgan Kaufmann Publishers Inc., 1984.

[Sow84b] J. F. Sowa : *Conceptual structures : information processing in mind and machine*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1984.

[Spe06] Andrew D. Spear : Ontology for the Twenty First Century : An Introduction with Recommendations. Rapport technique, INFOMIS, Sarrbrück, Germany, 2006.

[SSW08] Evren Sirin, Michael Smith et Evan Wallace : Opening, Closing Worlds - On Integrity Constraints. *In Fifth OWLED Workshop on OWL : Experiences and Directions (OWLED)*, 2008.

[Sto12] Michael Stonebraker : SciDB : An Open-Source DBMS for Scientific Data. *ERCIM News*, 2012(89), 2012.

[Str09] Umberto Straccia : A Minimal Deductive System for General Fuzzy RDF. *In Third International Conference Web Reasoning and Rule Systems*, pages 166–181, 2009.

[Str11] Cristof Strauch : NoSQL Databases. Rapport technique, Hochschule der Medien, Stuttgart, Germany, 2011.

[TDC+10] Anne Tireau, Caroline Domerg, Olivier Corby, Juliette Fabre, CatherineFaron-Zucker, Émilie Gennari, Alexandre Granier, Isabelle Mirbel, Vincent Negre et Pascal Neveu : Using Annotations for R Functions Management. *In MOQA : Méta-données et Ontologies pour la Qualité des Annotations*, 2010.

[TGDS04] Kerstin Thurow, Bernd Göde, Uwe Dingerdissen et Norbert Stoll : Laboratory Information Management Systems for Life Science Applications. *Organic Process Research & Development*, 8(6):970–982, nov 2004.

[TGH10] Herbert Thiele, Jörg Glandorf et Peter Hufnagel : Bioinformatics Strategies in Life Sciences : From Data Processing and Data Warehousing to Biological Knowledge Extraction. *Journal of Integrative Bioinformatics*, 7(1):1–1, 2010.

[THa06] C. F. Taylor, H. Hermjakob et al. : The work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI). *OMICS*, 10(2): 145–151, 2006.

[TPL+07] Chris F. Taylor, Norman W. Paton, Kathryn S. Lilley, Pierre-Alain A. Binz, Randall K. Julian, Andrew R. Jones, Weimin Zhu, Rolf Apweiler,

Ruedi Aebersold, Eric W. Deutsch, Michael J. Dunn, Albert J. Heck, Alexander Leitner, Marcus Macht, Matthias Mann, Lennart Martens, Thomas A. Neubert, Scott D. Patterson, Peipei Ping, Sean L. Seymour, Puneet Souda, Akira Tsugita, Joel Vandekerckhove, Thomas M. Vondriska, Julian P. Whitelegge, Marc R. Wilkins, Ioannnis Xenarios, John R. Yates et Henning Hermjakob : The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology*, 25(8):887–893, 2007.

[UCI+06] Victoria Uren, Philipp Cimiano, Jose Iria, Siegfried Handschuh, Maria Vargas-Vera, Enrico Motta et Fabio Ciravegna : Semantic Annotation for Knowledge Management : Requirements and a Survey of the State of the Art. *Web Semantics : Science, Services and Agents on the World Wide Web*, 4(1):14–28, 2006.

[URS10] Octavian Udrea, Diego Reforgiato Recupera et V.S. Subrahmanian : Annotated RDF. *ACM Transaction on Computational Logic*, 11(2):2–41, 2010.

[vdAtHKB03] Wil M. P. van der Aalst, Arthur H. M. ter Hofstede, Bartek Kiepuszewski et Alistair P. Barros : Workflow Patterns. *Distributed and Parallel Databases*, 14(1):5–51, 2003.

[Vie91] Laure Vieu : *Sémantique des relations spatiales et inférences spatio-temporelles : une contribution à l'étude des structures formelles de l'espace en Langage Naturel*. Thèse de doctorat, Université Paul Sabatier - Toulouse, 1991.

[Vie97] Laure Vieu : Spatial Representation and Reasoning in Artificial Intelligence. *In Spatial and Temporal Reasoning*, pages 5–41, 1997.

[VMZ+10] Chad Vicknair, Michael Macias, Zhendong Zhao, Xiaofei Nan, Yixin Chen et Dawn Wilkins : A comparison of a graph database and a relational database : a data provenance perspective. *In Proceedings of the 48th Annual ACM Southeast Regional Conference*, page 42, 2010.

[Vol04] Raphael Volz : *Web ontology reasoning with logic databases*. Thèse de doctorat, Université Karlsruhe (Allemagne), 2004.

[Whi] Raymond Whitlow : A Spatial Ontology for Archaeology. Rapport technique, University at Buffalo.

[Whi04] SA. White : Process Modeling Notations and Workflow Patterns, 2004.

[Wil98] S.J. Willson : Measuring Inconsistency in Phylogenetic Trees. *Journal of Theoretical Biology*, 190(1):15–36, 1998.

[WKKL10] René Witte, Ralf Krestel, Thomas Kappler et Peter C. Lockemann : Converting a Historical Architecture Encyclopedia into a Semantic Knowledge Base. *IEEE Intelligent Systems*, 25(1):58–67, 2010.

[Woo07] Simon Wood : Comprehensive Laboratory Informatics : A Multilayer Approach. *American Laboratory*, 39(16):20–23, 2007.

[WW02] Yair Wand et Ron Weber : Research Commentary : Information Systems and Conceptual Modeling - A Research Agenda. *Information Systems Research*, 13(4):363–376, 2002.

[ZHJ04] Tewfik Ziadi, Loïc Hélouët et Jean-Marc Jézéquel : Towards a UML Profile for Software Product Lines. *In* Frank van der Linden, éditeur : *Software Product-Family Engineering*, volume 3014 de *Lecture Notes in Computer Science*, pages 129–139. Springer Berlin / Heidelberg, 2004.

[ZKIN11] Ying Zhang, Martin L. Kersten, Milena Ivanova et Niels Nes : SciQL : Bridging the Gap between Science and Relational DBMS. *In 5th International Database Engineering and Applications Symposium (IDEAS)*, pages 124–133, 2011.

[ZLPS12] Antoine Zimmermann, Nuno Lopes, Axel Polleres et Umberto Straccia : A general framework for representing, reasoning and querying with annotated semantic web data. *Journal of Web Semantics, Elsevier*, 11:72–95, 2012.

Résumé

Les Systèmes d'Information Scientifique (SIS) sont des Systèmes d'Information (SI) dont le but est de produire de la connaissance et non pas de gérer ou contrôler une activité de production de biens ou de services comme les SI d'entreprise. Les SIS se caractérisent par des domaines de recherche fortement collaboratifs impliquant des équipes pluridisciplinaires et le plus souvent géographiquement éloignées, ils manipulent des données aux structures très variables dans le temps qui vont au-delà de la simple hétérogénéité : nuages de points issus de scanner 3D, modèles numériques de terrain, cartographie, publications, données issues de spectromètre de masse ou de technique de thermoluminescence, données attributaires en très grand volume, etc. Ainsi, contrairement aux bases de données d'entreprise qui sont modélisées avec des structures établies par l'activité qu'elles supportent, les données scientifiques ne peuvent pas se contenter de schémas de données pré-definis puisque la structure des données évolue rapidement de concert avec l'évolution de la connaissance. La gestion de données scientifiques nécessite une architecture de SIS ayant un niveau d'extensibilité plus élevé que dans un SI d'entreprise.

Afin de supporter l'extensibilité tout en contrôlant la qualité des données mais aussi l'interopérabilité, nous proposons une architecture de SIS reposant sur :
– des données référentielles fortement structurées, identifiables lors de la phase d'analyse et amenées à évoluer rarement ;
– des données complémentaires multi-modèles (matricielles, cartographiques, nuages de points 3D, documentaires, etc.).
Pour établir les liens entre les données complémentaires et les données référentielles, nous avons utilisé un **unique paradigme, l'annotation sémantique**. Nous avons proposé un modèle formel d'annotation à base ontologique pour construire des annotations sémantiques dont la cohérence et la consistance peuvent être contrôlées par une ontologie et des règles. Dans ce cadre, les annotations offrent ainsi une contextualisation des données qui permet de vérifier leur cohérence, par rapport à la connaissance du domaine. Nous avons dressé les grandes lignes d'une sémantique du processus d'annotation par analogie avec la sémantique des langages de programmation.

Nous avons validé notre proposition, à travers deux collaborations pluridisciplinaires :
– le projet ANR CARE (*Corpus Architecturae Religiosae Europeae - IV-X saec.* ANR-07-CORP-011) dans le domaine de l'archéologie. Son objectif était de développer un corpus numérique de documents multimédia sur l'évolution des monuments religieux du IV^{e} au XI^{e} siècle (http://care.tge-adonis.fr). Un assistant d'annotation a été développé pour assurer la qualité des annotations par rapport à la connaissance représentée dans l'ontologie. Ce projet a donné lieu au développement d'une extension sémantique pour MediaWiki ;
– le projet *eClims* dans le domaine de la protéomique clinique. *eClims* est un composant clinique d'un LIMS (*Laboratory Information Management System*) développé pour la plate-forme de protéomique CLIPP. *eClims* met en œuvre un outil d'intégration basé sur le couplage entre des modèles représentant les sources et le système protéomique, et des ontologies utilisées comme médiatrices entre ces derniers. Les différents contrôles que nous mettons en place garantissent la validité des domaines de valeurs, la complétude, la consistance des données et leur cohérence. Le stockage des annotations est assuré par une Base de Données orientées colonnes associée à une Base de Données relationnelles.

Abstract

Scientific Information Systems (SIS) aim to produce or improve knowledge on a subject through activities of research and development. Management of scientific data requires a high level of extensibility which is generally much higher than in enterprise Information System (IS). In general, the scope and the complexity of scientific activities are such that it is necessary to cope with a context of multi-disciplinary research teams geographically dispersed, producing and using different kinds of data. Functionalities in an enterprise IS are all directed towards the support of business processes, thus pre-established and comprehensive error procedures are developed to deal with all the possible exceptions. SIS are generally organized according to studies i.e., a scientific research project that addresses a specific subject. It is difficult to model extensively a study both in the data and event models, because the nature of research may change after some data have been collected and analyzed, new questions can arise and can generate new studies.

We propose an architecture that supports extensibility, data quality and interoperability relying on:

– Master data are information that are essential to support a specific business. They are well defined and evolve rarely but they are often scattered in various parts/applications of the information system;
– Complementary multi-model data (sequences, graphs, 3D structure, images, documents, data sets, etc.).

To establish links between master and complementary data, we use **a unique paradigm: semantic annotation**.

We present two applications that validate our architecture:

– The CARE project (*Corpus Architecturae Religiosae Europeae - IV-X saec.* ANR-07-CORP-011) for archaeological domain. The aim of the CARE project is the setting up of a corpus describing Christian edifices in Europe. Each edifice is described in a document that focuses on the definition of states of evolutions from the 4th century to the 11th century (http://care.tge-adonis.fr). The requirements of a web platform with a collaborative component and the need of document management led us to develop a solution based on a wiki rather than a database. So, we develop *WikiBridge* a semantic wiki for the CARE project by extending MediaWiki with some structural DBMS capabilities and semantic tools: form based acquisition interface, annotation interface, annotation validation, semantic rules and a semantic query engine.
– The *eClims* project for clinical proteomic domain. Tracking samples of clinical proteomics needs the establishment of a rigorous management of data which requires the use of a LIMS (Laboratory Information Management System) for controlling data before and during the experiments as well as validating derived data obtained after statistical analysis. Many actors provide data of variable quality, however once imported in the SIS, these data must conform to the same quality as the existing data. To ensure data quality in a biomedical SIS, two mechanisms must be diligently controlled: 1) importation which inserts new data in the IS and, 2) annotation which allows to extend descriptions of existing data with data that were not originally modeled. When importing data sets, some pieces of data exactly match with the existing structures (i.e., master data) and are directly integrated in the database, others data are stored as annotations, on master data's tuples, using RDF triples.

So, we notice that the annotation is a universal structure allowing to develop generic components taking care of the necessary variability in SIS.