Base de Données annotées et Wiki pour la constitution du corpus numérique CARE

Pascale Chevalier, Ludovic Granjon, Eric Leclercq, Arnaud Millereux, Marinette Savonnet, Christian Sapin

Base de Données annotées et Wiki pour la constitution du corpus numérique CARE

Pascale Chevalier[14], Ludovic Granjon[2], Éric Leclercq[3], Arnaud Millereux[3], Marinette Savonnet[3], Christian Sapin[1]

1 Laboratoire ARTeHIS UMR 5594
Université de Bourgogne, Dijon, France
2 MSH Dijon, pôle de géomatique
Université de Bourgogne, Dijon, France
3 Laboratoire Électronique, Informatique et Image UMR 5158
Université de Bourgogne, Dijon, France
4 Université Blaise-Pascal, Clermont-Ferrand, France

Résumé :

L'objet de cet article est la présentation d'une plate-forme collaborative, *WikiBridge*, gérant les connaissances du corpus CARE. 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 par une ontologie. La plate-forme est basée sur une interface wiki associée à une base de données annotées et respecte les recommandations du Web Sémantique (RDF, OWL, SPARQL).

Abstract:

The aim of this paper is the description of a collaborative platform, *WikiBridge*, designed for managing knowledge in the context of the corpus CARE. This platform provides users with tools needed to work on the description of religious buildings and their evolution over the centuries through a specific spatio-temporal model. The knowledge of archaeologists is modeled through an ontology. The platform is based on a wiki interface associated with an annotated database and meets the requirements of the Semantic Web (RDF, OWL, SPARQL).

1. Organisation de la plate-forme numérique du projet CARE

D'un point de vue organisationnel, le projet CARE prend la forme d'un réseau d'experts (des archéologues, des historiens, des historiens de l'art, des dessinateurs topographes) assurant l'alimentation du corpus et collaborant à son exploitation au moyen de travaux de recherche. Cette coopération se double d'un partenariat avec des chercheurs en informatique, qui a pour objectif d'offrir un **corpus numérique** intégrant des informations aussi variées qu'une description textuelle d'objets, une bibliographie, des photographies, etc. Le corpus doit être constitué de façon collaborative puisqu'il s'agit d'une agrégation de connaissances que l'on veut produire, partager, échanger et pouvoir faire évoluer. Le corpus doit pouvoir être consultable sur Internet, être exploité par des utilisateurs experts du domaine mais aussi par des utilisateurs novices. Par conséquent, la plate-forme logicielle qui hébergera le corpus numérique doit fournir 1) une présentation électronique qui doit être le reflet de la version imprimée pour des fins de citation et 2) des outils d'annotation

sémantique pour la représentation des connaissances supplémentaires ou spécialisées. Des caractéristiques propres au projet CARE 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 comprendre, modéliser, annoter les documents ; 3) l'évolution de la connaissance et 4) les compétences des utilisateurs.

Les utilisateurs du projet CARE ont identifié leurs besoins qui se sont traduits au niveau de la conception de la plate-forme par les quatre exigences suivantes : gestion de documents composites, interface utilisateur intuitive, conception collaborative du contenu, support de différentes catégories d'utilisateurs. À partir de l'analyse des besoins et des spécifications techniques, deux axes non fonctionnels ont émergé : d'une part les aspects collaboratifs et d'autre part les aspects sémantiques. Ces derniers peuvent être couverts par une modélisation de la connaissance du domaine sous la forme d'ontologie qui sera exploitée au moyen d'annotations associées aux documents numériques. Ainsi, nous avons développé la plate-forme logicielle, *WikiBridge* (`http://care.u-bourgogne.fr`), s'appuyant sur les technologies standards du Web Sémantique.

D'un point de vue technique, la plate-forme se présente, pour sa partie utilisateur, sous la forme d'un wiki sémantique (section 2) et pour sa partie infrastructure de gestion de données et de connaissances sous la forme d'une base de données annotées couplée avec un triple-store (section3) pour le stockage des annotations et des ontologies.

2. Entre base de données et document : un wiki comme modèle d'interaction

Tout comme Lock [1] qui pense que l'interprétation des données ne doit pas être déterminée par la technologie utilisée mais bien par la discipline étudiée, Bachimont [2] 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 une description textuelle.

2.1 Caractéristiques et utilisations d'un wiki

Wiki est la forme abrégée de "WikiWikiWeb" qui est dérivée du redoublement hawaiien "wiki wiki" signifiant "rapide". Un wiki est un logiciel permettant de gérer un ensemble de pages Web, organisées en articles, reliées par des liens hypertextes. Les différents wikis partagent les caractéristiques suivantes :

- le contenu est modélisé sous la forme d'articles, il est édité et modifié via l'interface d'un navigateur Web, sans avoir à installer de logiciel supplémentaire ;
- le contenu est exprimé dans un format hypertexte simplifié, à l'aide d'une "syntaxe wiki" qui est beaucoup plus facile à utiliser pour des utilisateurs non informaticiens que le langage HTML. Généralement, un éditeur WYSIWYG[1] (fig. 1) permet de s'abstraire de la connaissance de la syntaxe du langage pour gérer la mise en page (texte en italique, en gras, en couleur, réalisation de liste, de tableau, de chapitre, de section, etc.) ;

[1] WYSIWYG : What You See Is What You Get.

- le contenu peut comporter et agréger des données hétérogènes. Plusieurs documents (fichiers, photographies, plans, son, vidéo) peuvent être ajoutés à un article. Après avoir été téléchargé sur la plate-forme, les documents peuvent être visualisés dans l'article en tant que vignettes (fig. 2) ;
- l'utilisation de services externes comme Google Map. Dans le projet CARE, nous avons aussi utilisé des services d'aide à la saisie comme l'extension Maps pour trouver la latitude et la longitude d'un édifice en fonction de son adresse ou le site de l'INSEE donnant le code et le libellé exacts des communes pour l'auto-complétion des listes (fig. 3) ;
- les modifications apportées au contenu d'un article sont sauvegardées chaque fois qu'elles sont publiées, ainsi les versions précédentes de l'article sont conservées. En outre, la plupart des systèmes peuvent comparer deux versions d'un article, ce qui permet d'identifier les changements entre les différentes versions rapidement ;
- l'accès, dans la plupart des wikis, est sans aucune restriction. Tout le monde peut corriger, modifier, compléter ou même supprimer un article ou son contenu. Certains wikis permettent de restreindre l'accès à des utilisateurs et/ou des groupes d'utilisateurs. Pour prendre en compte les différents niveaux d'utilisation nécessaires au projet CARE dans *WikiBridge*, c'est-à-dire à la fois des utilisateurs novices, des utilisateurs expérimentés ou encore des chercheurs, certaines fonctionnalités doivent être désactivées. Un mécanisme d'ACL (Access Control List) est nécessaire pour définir les fonctionnalités accessibles par utilisateur et par groupe. La configuration actuelle (non figée) repose sur trois types d'utilisateurs : 1) le responsable du pays qui a un accès complet à l'ensemble des ressources ; 2) le responsable de région qui a 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 ;

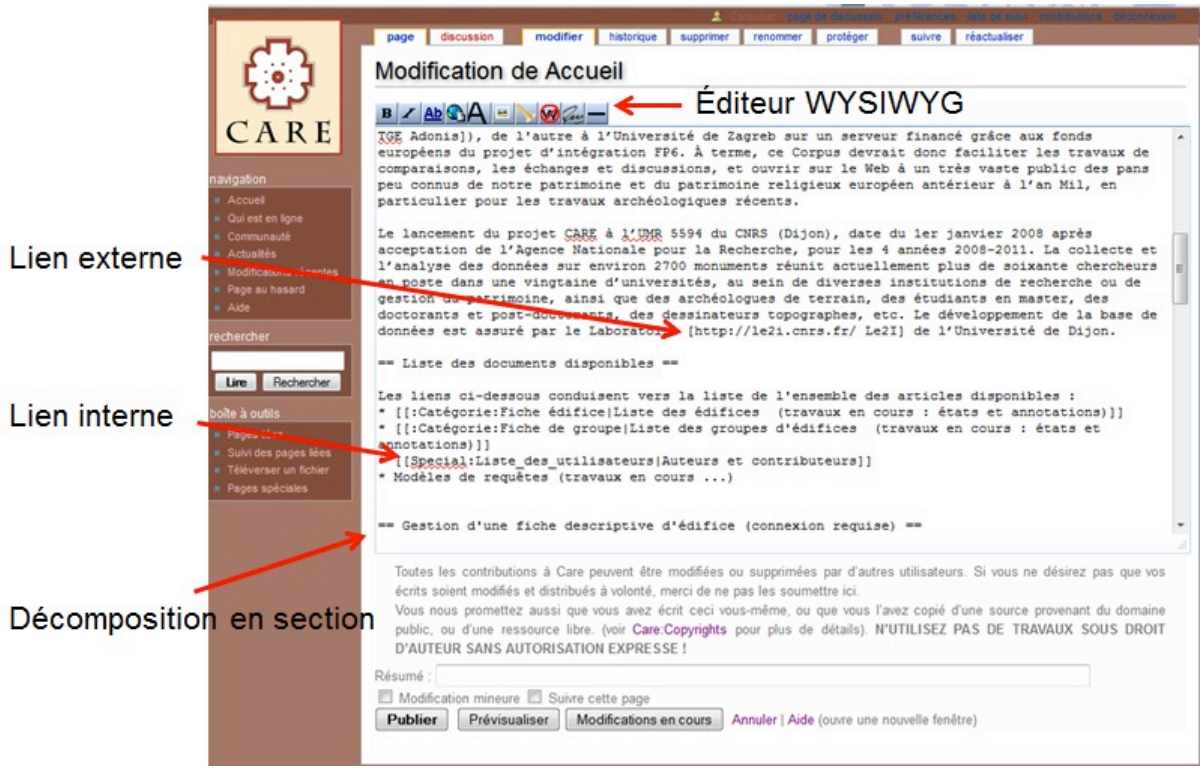

Fig. 1 : Syntaxe wiki

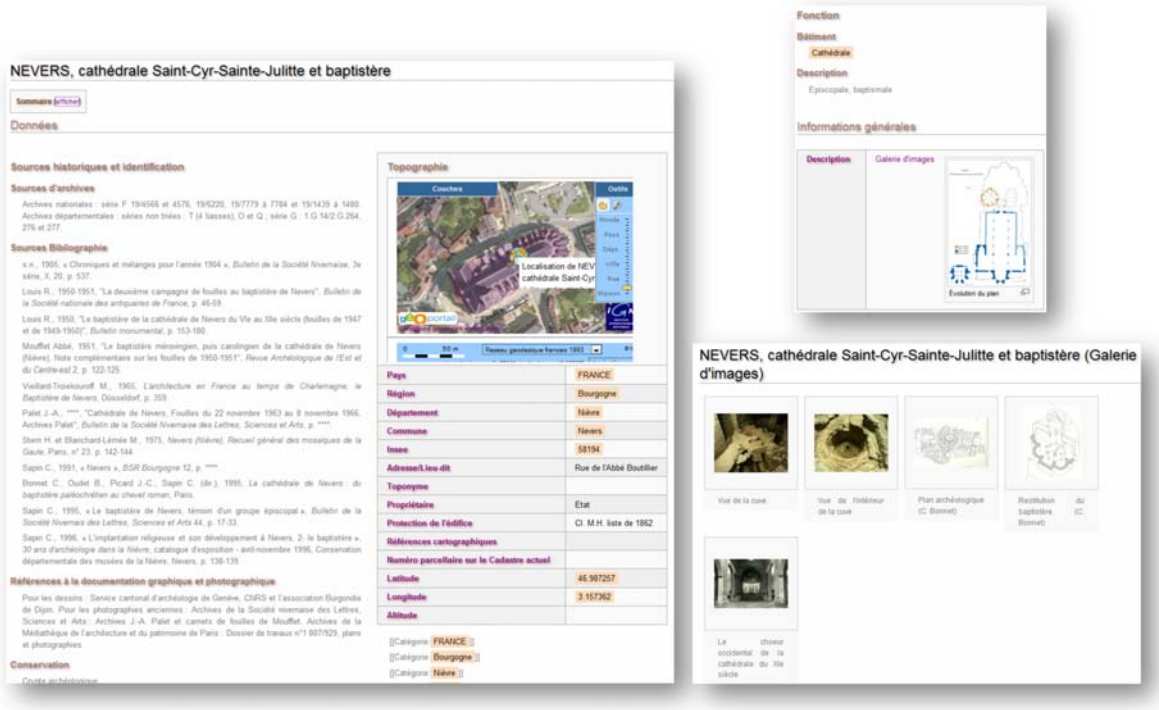

Fig. 2 : Article du wiki, galerie de photographies et plan correspondant à un édifice

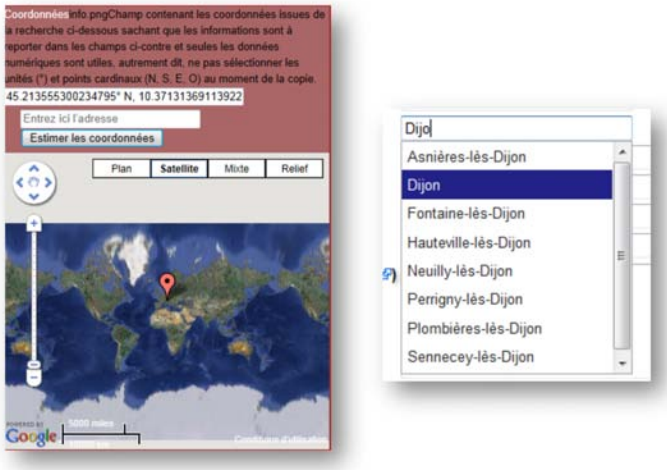

Fig. 3 : Services Web d'aide à la saisie

- la construction collaborative du contenu est supportée par les listes de suivi et les discussions associées aux articles. Cette fonctionnalité est présente dans la majorité des wikis ;
- la recherche d'articles est mise en œuvre par un moteur de recherche textuel. Cet outil facilite la navigation mais il n'est pas suffisant pour permettre l'analyse du corpus. Quelques rares wikis proposent un véritable moteur de requêtes.

En résumé, la couche d'interaction de *WikiBridge* avec les utilisateurs est majoritairement couverte par le moteur de wiki MediaWiki[2]. Il permet *un premier enrichissement* du document papier grâce :

1. aux liens soit internes entre des articles du wiki soit externes vers une page Web extérieure. Les liens internes sont principalement utilisés pour lier un groupe avec ses édifices constituants. Les liens externes permettent de compléter des parties du texte, par exemple on peut donner l'URL d'un musée dans lequel se trouve aujourd'hui un objet faisant partie de l'édifice. Les liens externes peuvent aussi offrir une aide lors de la saisie d'une fiche ;
2. au support du multimédia comme des photographies, des plans, du son ou de la vidéo ;
3. à des capacités d'invocation de services Web externes.

2.2 Structuration de l'information dans *WikiBridge*

Les documents présents dans le wiki sont définis selon trois niveaux (fig. 4) : leur contenu (agrégation de ressources multimédia et textuelles), leur structure, la connaissance qui leur est associée. Le premier niveau est mis en place par le wiki. Les autres niveaux sont, dans la plupart des wikis, des fonctionnalités supplémentaires.

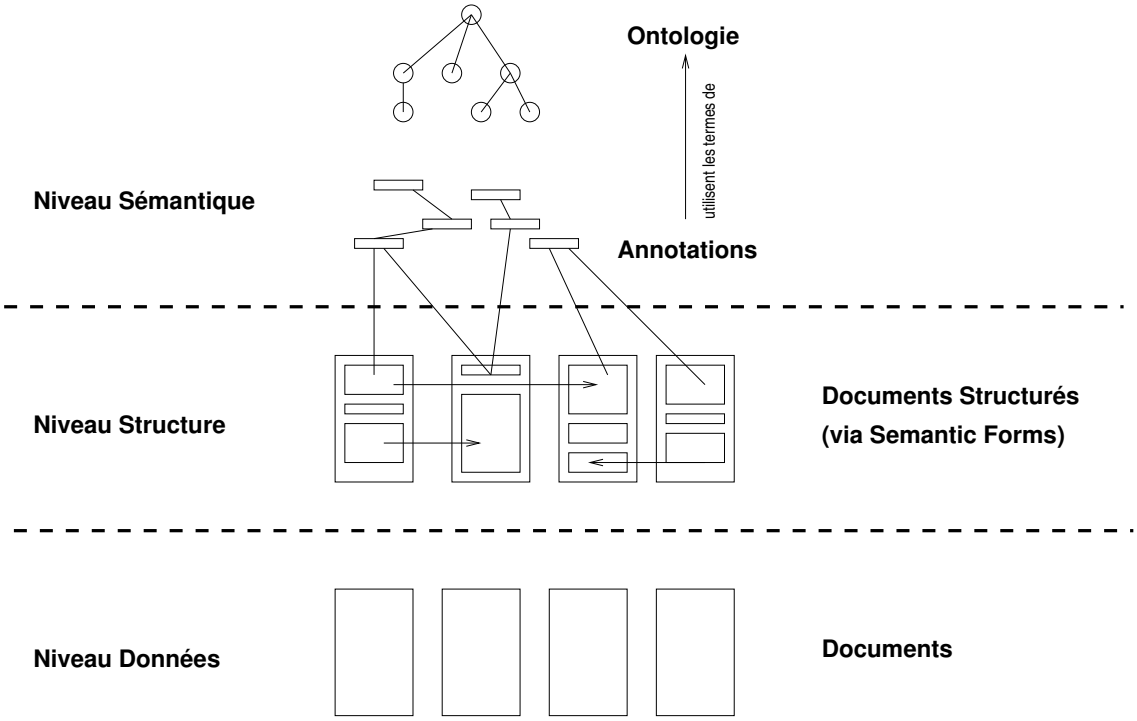

Fig. 4 : Structuration de l'information selon trois niveaux

Chaque document se traduit sous la forme d'un article dans le wiki. La notion de *wiki template* permet aux utilisateurs de définir la structure d'un article [3]. Les templates créationnels sont des pages utilisées comme point de départ pour la création de nouveaux articles ayant la structure définie par le template. Un wiki utilisant des templates

[2] http://www.mediawiki.org Mediawiki est le moteur de wiki choisi par l'encyclopédie libre Wikipédia.

créationnels est qualifié de "*Lightly Constrained Wiki*". Semantic Forms[3] développée pour MediaWiki permet de définir des templates créationnels. L'organisation des documents du projet CARE est modélisée dans *WikiBridge* grâce à cette extension (fig. 5). Les champs à remplir 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.

Fig. 5 : Structuration des documents

Les formulaires, reflétant l'organisation du document papier, permettent à des utilisateurs non-experts du domaine de remplir grâce à un copier-coller les champs à partir des documents papier produits par les archéologues. Un aperçu de l'interface de saisie est donné en fig. 6.

2.3 Exploitation des documents

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 d'un

[3] http://www.mediawiki.org/wiki/Extension:Semantic_Forms

véritable langage de requête comparable à SQL. Par conséquent, 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.

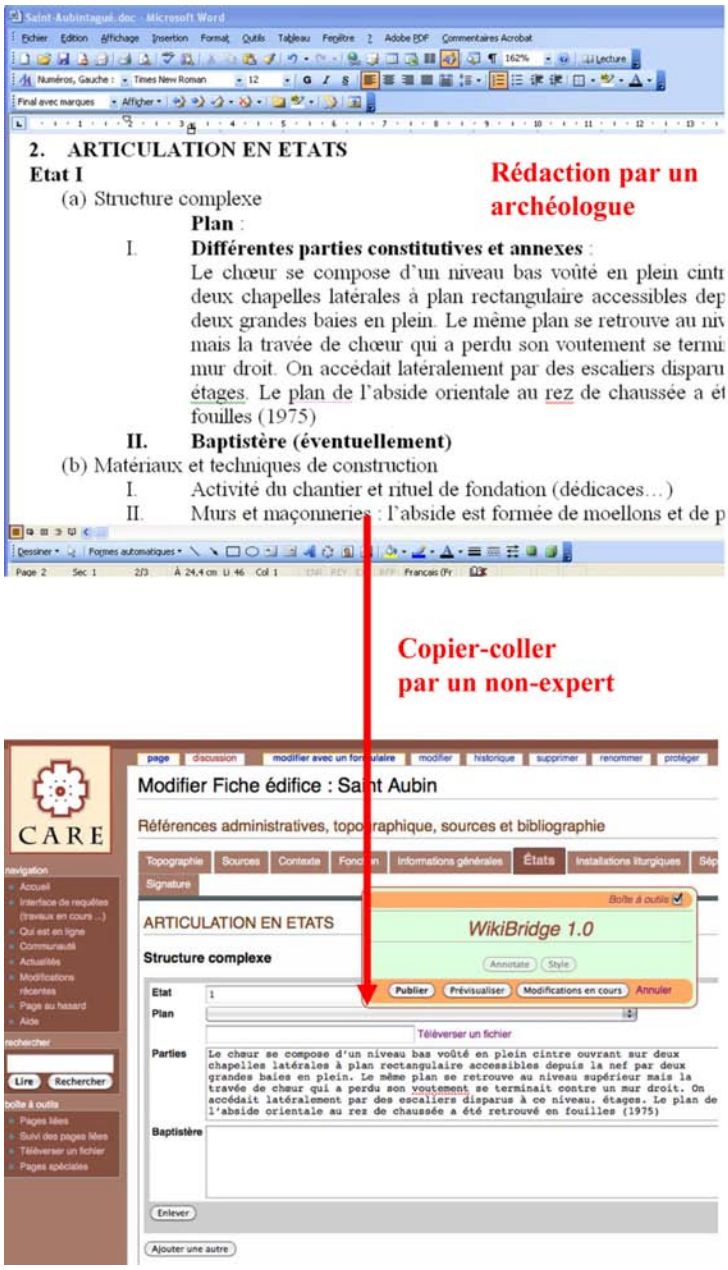

Fig. 6 : Processus de saisie d'un article

3. Base de données annotées

La base de données annotées que nous avons mis en place dans *WikiBridge* complète la base de données des documents héritée de MediaWiki avec une structure pour l'ontologie et une structure pour les annotations.

3.1 Ontologie

Une ontologie, en informatique, est une spécification d'une conceptualisation [4, 5]. Dans le cadre du projet CARE, elle prend la forme d'un ensemble de concepts propres au domaine de l'archéologie et de relations entre ces concepts. Les concepts sont complétés par la notion d'individus, c'est-à-dire des termes qui font référence aux concepts. Un autel est par exemple un individu du concept installation liturgique. Une ontologie est donc une forme évoluée de thésaurus ou de terminologie. La base théorique est la logique de description [6], elle apporte un mécanisme de déduction basée sur la subsomption[4]. Les technologies de représentation standards des ontologies sont issues du W3C[5], il s'agit de la famille de langage OWL[6].

Pour la construction de l'ontologie CARE, notre base de travail a été constituée d'extraits en langage naturel des descriptions des différents éléments d'un édifice. Nous avons travaillé sur une représentation formelle des connaissances des éléments d'architecture en :

- classifiant les éléments architecturaux d'un édifice ;
- définissant les relations entre les éléments. 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. ;
- établissant un 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, par exemple la nef est l'endroit où sont rassemblés les fidèles.

Cette méthodologie nous a permis de dégager un ensemble de concepts connectés à des termes architecturaux organisés par des relations spatiales, méronomiques (décomposition morphologique) et représentationnelles (dimensions, matériaux, etc.). Les concepts ont ensuite été rapprochés de l'ontologie CIDOC-CRM. Le CIDOC (Comité International pour la Documentation) soutenu par l'ICOM (International Council of Museums) a pour objectif d'améliorer la gestion des archives et des produits scientifiques liés au patrimoine artistique et culturel. L'idée de base est de mettre en place un modèle de données standard pour décrire des objets de musées et des informations culturelles comme des œuvres d'art ou des vestiges. Ce modèle doit être exploitable par des systèmes informatiques et être suffisamment riche pour rendre compte de la diversité des analyses et des interprétations. Le modèle appelé CRM (Conceptual Reference Model) élaboré depuis 1994, a été publié en 2006 par l'ISO en tant que norme internationale (ISO 21127:2006) (http://www.cidoc-crm.org). CIDOC-CRM apporte pour le projet CARE la modélisation temporelle sous la forme d'états. Nous avons abouti à une spécialisation de CIDOC-CRM comportant 124 concepts et 715 individus. La fig. 7 montre un extrait de l'ontologie CARE, les concepts avec un suffixe Exx sont des concepts issus de CIDOC-CRM.

---

[4] La subsomption désigne une relation hiérarchique entre des concepts, dans les logiques de description.

[5] Le World Wide Web Consortium (W3C) est un organisme fondé en 1994 chargé de promouvoir les technologies du Web telles que HTML, XHTML, XML, RDF, SPARQL.

[6] Web Ontology Language (OWL) est un langage de représentation des connaissances construit sur le modèle de données de RDF pour définir des ontologies structurées (http://www.w3.org/2004/OWL/).

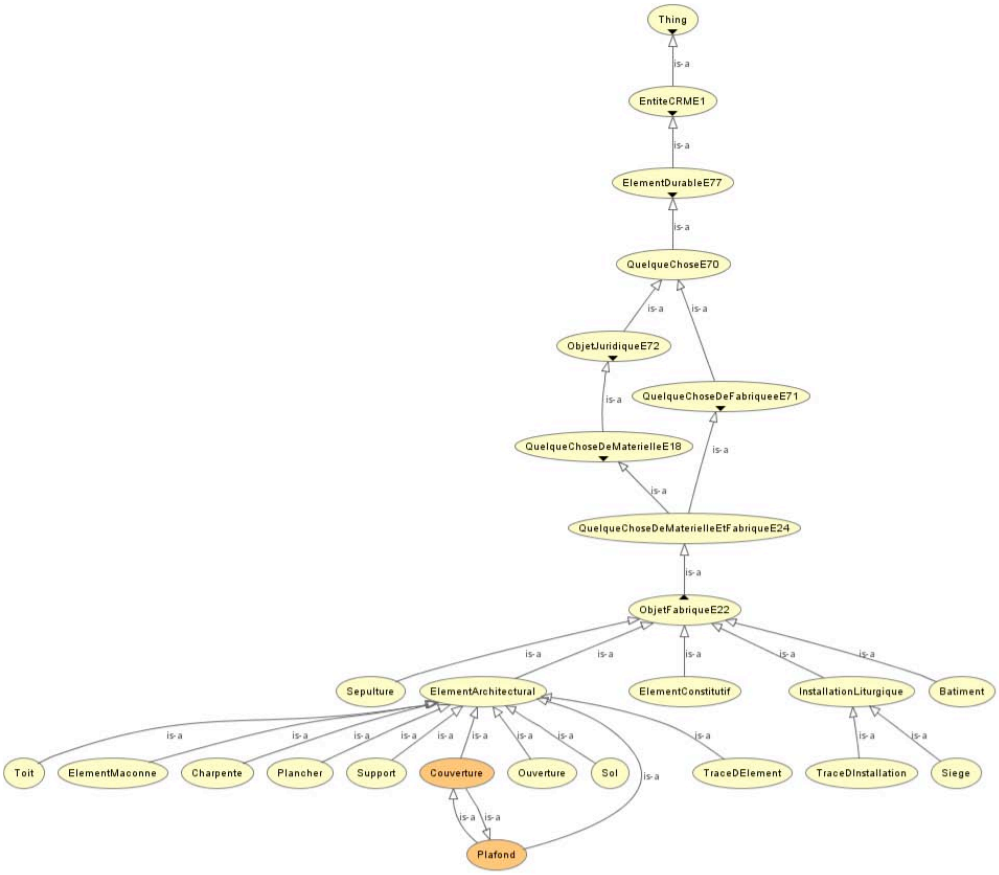

Fig. 7 : Extrait de l'ontologie CARE

3.2 Annotations

Le mécanisme d'annotation associe à des ressources textuelles (article, section, paragraphe, mot) une signification donnée au moyen d'un terme de l'ontologie. La base formelle du mécanisme d'annotation est le graphe conceptuel [7]. Comme les ontologies, les annotations possèdent des langages de représentation, par exemple, RDF[7] permet de représenter des ressources sur le Web de manière uniforme. Afin de décrire ces ressources, RDF se base sur la notion de triplets avec :

- un *sujet*, c'est-à-dire la ressource identifiée par une URI[8] ;
- un *prédicat*, c'est-à-dire la propriété assignée à la ressource, également identifié par une URI faisant référence à un terme de l'ontologie ;
- un *objet*, c'est-à-dire la valeur de la propriété. La valeur peut être de type primitif (numérique, chaîne de caractères), être un terme de l'ontologie ou être une autre ressource. Dans ce cas, l'objet de l'annotation peut à son tour être le sujet d'un nouveau triplet.

[7] Resource Description Framework (RDF) est un langage pour décrire de façon formelle les ressources Web et leurs métadonnées, en vue d'un traitement automatique (http://www.w3.org/RDF/).

[8] Uniform Ressource Identifier.

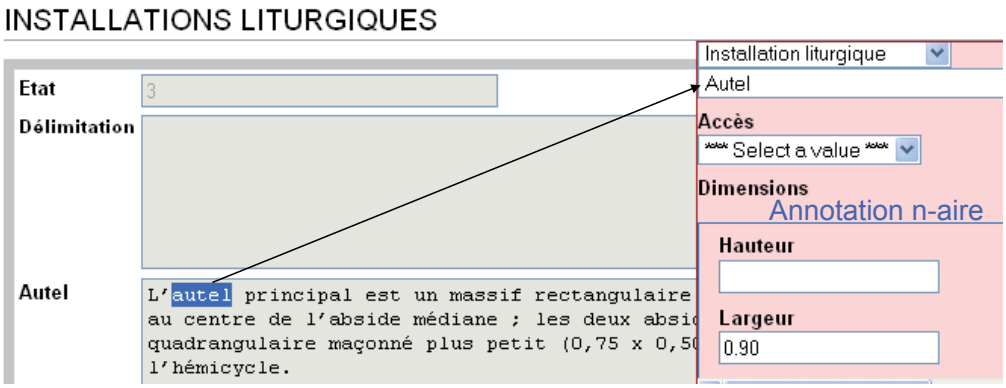

Fig. 8 : Exemple d'annotation

Dans *WikiBridge*, les interfaces de saisie des documents et des annotations sont uniformes. La fig. 8 présente un exemple d'annotation qui se traduit par les triplet suivants (fig. 9) :

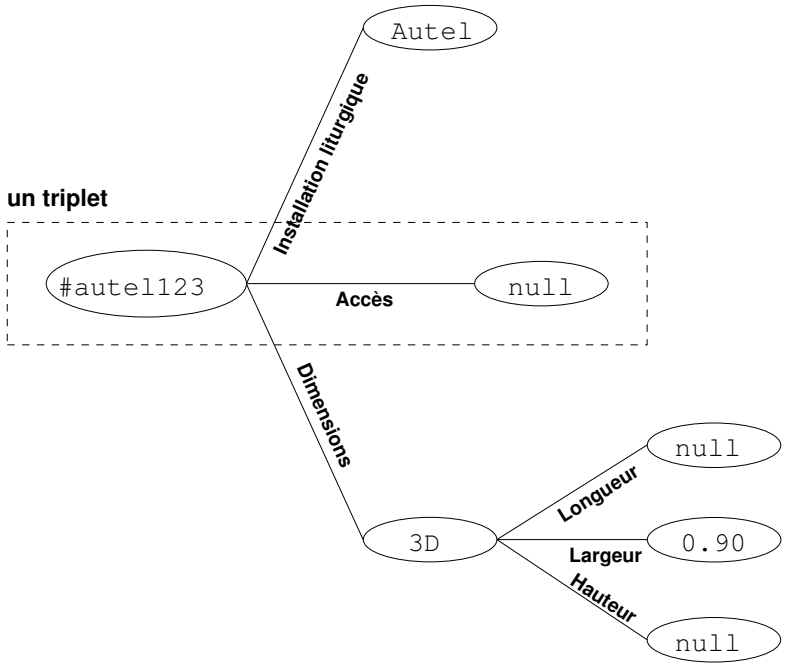

Fig. 9 : Annotation de la figure 8 représentée sous la forme d'un graphe de triplets

Les capacités de déduction des logiques de description permettent de contrôler la validité des annotations offrant ainsi un même niveau de contrôle que celui présent dans les bases de données.

3.3 Exploitation des données et de la connaissance

L'annotation sémantique offre une meilleure qualité dans le processus d'évaluation qu'un moteur de recherche textuel. En effet, le langage SPARQL[9] permet d'interroger la base d'annotations (stockée dans une structure spécifique nommée triple-store) et l'ontologie au moyen de requêtes comparables à celles écrites dans le langage SQL des bases de données.

[9] SPARQL Protocol and RDF Query Language.

Nous utilisons, d'une part SPARQL pour exécuter des requêtes dont les paramètres sont saisis au travers d'un formulaire du wiki (fig. 10) et d'autre part pour remplir des listes de valeurs à partir des individus des concepts de l'ontologie qui seront utilisées lors de saisie des articles. De plus, pour les utilisateurs experts, les requêtes SPARQL peuvent être incluses dans des pages du wiki (in-line queries).

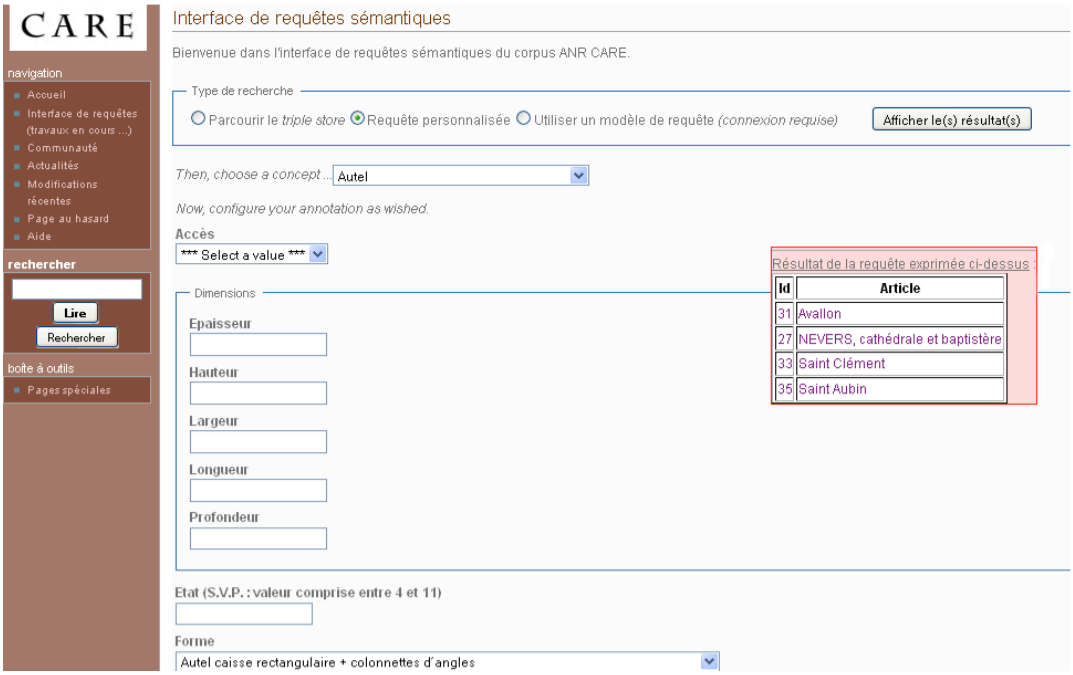

Fig. 10 : Interface de requêtes utilisant un formulaire

Nous avons également intégré le mécanisme de requêtes dans un service Web qui permet l'échange de données avec d'autres applications comme par exemple l'outil d'analyse spatiale développé par la MSH de Dijon (fig. 11). Dans cette figure, les valeurs présentes dans la boîte de dialogue proviennent de la base de données annotées.

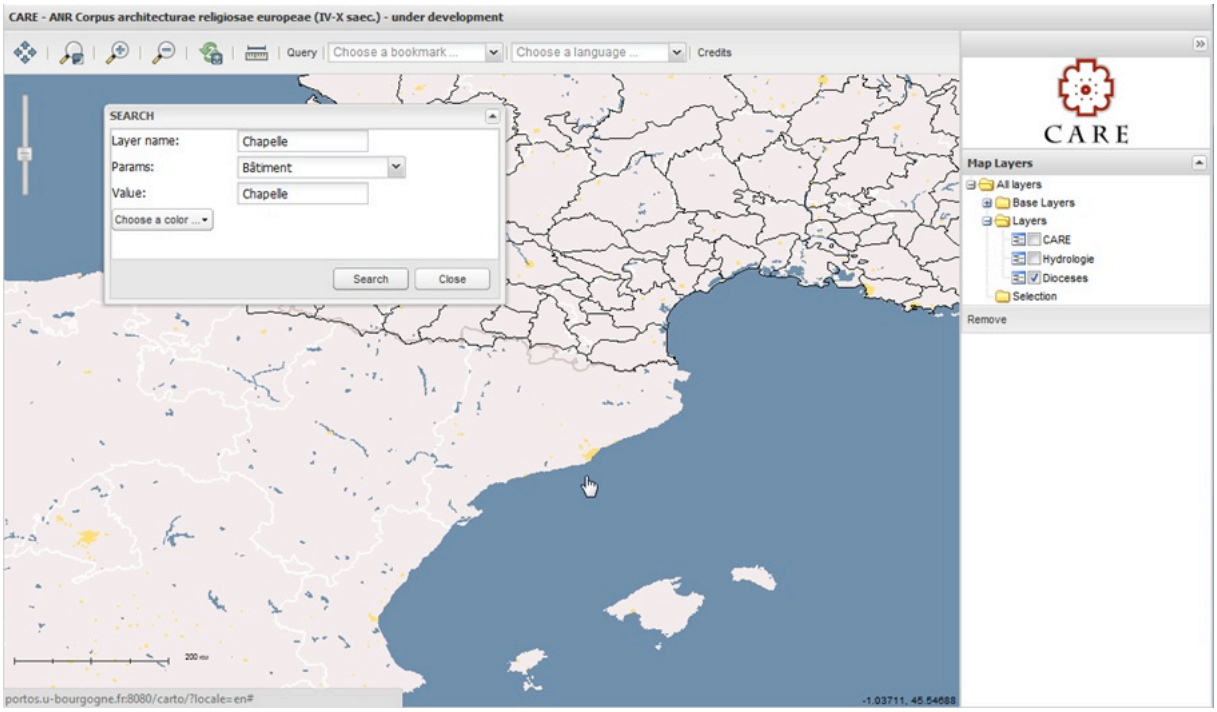

Fig. 11 : Outil d'analyse spatiale

L'utilisation de technologies standards (OWL, RDF, SPARQL) et les services Web permet une interopérabilité sémantique entre *WikiBridge* et d'autres applications (bases de données traditionnelles, SIG, wikis) grâce aux annotations s'appuyant sur une ontologie commune.

3. Conclusion

La création de *WikiBridge* a été pour les informaticiens un terrain d'expérimentation des concepts qu'ils élaborent dans les thématiques d'ingénierie des ontologies et du Web Sémantique [8].

Wikibridge permet de constituer un patrimoine numérique exploitable sur le long terme car la sémantique est rendue explicite par l'utilisation d'une base de données annotées. Le projet CARE fait l'objet d'une intégration sur la grille ADONIS[10] et des services Web sont en cours de développement pour une utilisation avec le système ISIDORE[11] afin d'assurer sa pérennité et d'améliorer sa visibilité.

WikiBridge est conçu de manière à pouvoir être transposé pour répondre aux besoins d'autres domaines car il s'appuie sur les technologies standards du Web Sémantique.

L'utilisation des capacités de raisonnement apportées par les bases formelles des technologies utilisées et le support de processus d'intelligence collective via la collaboration autour de la plate-forme *WikiBridge* doivent permettre : 1) de faire émerger de nouvelles connaissances à partir de celles déjà acquises et des données du wiki ; 2) de contrôler la

[10] http://www.tge-adonis.fr/
[11] http://www.rechercheisidore.fr/

signification des annotations par rapport à leur contexte d'utilisation et 3) d'améliorer la navigation et les recherches.

Bibliographie

[1] G. LOCK, *Archaeological Computing Then and Now : Theory and Practice, Intentions and Tensions*, in *Archeologia e Calcolatori*, 2009, p.75–84.

[2] B. BACHIMONT, L'intelligence artificielle comme écriture dynamique : de la raison graphique à la raison computationnelle. *Au nom du sens*, 2000, p. 290–319.

[3] A. DI IORIO, S. ZACCHIROLI, *Constrained Wiki : an Oxymoron ?* in *Int. Sym. Wikis*, 2006, p. 89–98.

[4] T. GRUBER, *A Translation Approach to Portable Ontology Specifications*, in *Knowledge Acquisition*, 5(2), 1993, p. 199-220.

[5] T. GRUBER, *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2009.

[6] D. NARDI, R. BRACHMAN. *An Introduction to Description Logics,* in *the Description Logic Handbook*, edited by F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, P.F. Patel-Schneider, Cambridge University Press, 2002, p. 5-44.

[7] F. SOWA, *Conceptual graphs for a database interface,* in *IBM Journal of Research and Development*, 20(4), 1976, p. 336-357.

[8] E. LECLERCQ, M. SAVONNET, *Système d'Information pour la production de connaissances : l'approche wiki sémantique,* in *INFORSID*, 2011, p. 233-248.