

OntoNeuroLOG : une ontologie modulaire et multi-niveaux pour gérer l'hétérogénéité sémantique des métadonnées[†]

Lynda Temal¹, Pascal Lando², Michel Dojat³, Frédéric Fürst², Bernard Gibaud¹, Gilles Kassel² et Anne Lapujade²

¹ Projet Visages U746, INSERM/INRIA/CNRS/Université de Rennes 1, Rennes
{lynda.temal, bernard.gibaud}@irisa.fr

² LaRIA FRE 2733, CNRS/Université de Picardie Jules Verne, Amiens
{pascal.lando, frederic.furst, gilles.kassel, anne.lapujade}@u-picardie.fr

³ GIN U836, INSERM/CEA/CHU de Grenoble, Université Joseph Fourier, Grenoble
Michel.Dojat@ujf-grenoble.fr

Résumé : Dans le cadre du projet NeuroLOG visant à construire un environnement logiciel en milieu ouvert pour la mutualisation de ressources en neuroimagerie, une approche médiateur est mise en place afin d'intégrer des sources de données distribuées et hétérogènes. Une ontologie – OntoNeuroLOG – est ainsi conçue pour permettre la recherche centralisée et unifiée de ressources décrites au moyen de métadonnées. La mise en place d'un tel environnement soulève plusieurs problèmes liés notamment à la construction et l'exploitation de l'ontologie. Lors de la journée, nous comptons adresser principalement les premiers. Nous mettrons en évidence le fait que l'hétérogénéité sémantique importante des métadonnées rencontrée dans le projet NeuroLOG nécessite de concevoir une ontologie couvrant des domaines très divers, ce qui pose des problèmes de « passage à l'échelle » sur le plan de la modélisation des connaissances. Nous présenterons l'approche retenue consistant à développer une ontologie modulaire et multi-niveaux, structurée au moyen d'une ontologie de haut niveau et intégrant des ontologies génériques de plusieurs domaines.

Mots-clés : Méthode de construction d'ontologies, Ontologies modulaires et multi-niveaux, Ontologies formelles et de haut niveau (foundational ontologies), Ontologies noyaux de domaines (core domain ontologies).

1 Problématique

Le projet NeuroLOG a pour objectif de construire un environnement logiciel en milieu ouvert pour l'intégration de ressources en neuroimagerie (des images et des modules d'analyse et de traitement de ces images) et de confronter cet environnement à des applications cibles orientées vers le suivi de pathologies cérébrales de type tumeur, accident vasculaire et sclérose en plaques. L'existence d'un tel

[†] Ce travail est financé dans le cadre du projet NeuroLOG (ANR-06-TLOG-024) du programme Technologies Logicielles de l'Agence Nationale de la Recherche : <http://neurolog.polytech.unice.fr>.

environnement doit faciliter la réalisation de travaux de recherche multicentriques dans le domaine des neurosciences cliniques, en permettant la création de cohortes de sujets virtuelles de taille importante, et en facilitant l'accès à des outils de traitement d'origine diverse, utilisables, soit à des fins méthodologiques pour quantifier leur robustesse, soit pour composer de nouvelles chaînes de traitement optimales.

NeuroLOG entend adresser, notamment, les problèmes relatifs à la gestion et l'accès aux données semi-structurées, hétérogènes et distribuées en environnement ouvert. Pour fédérer ces sources de données, une sémantique de référence est élaborée et spécifiée sous forme d'une ontologie (OntoNeuroLOG). Les primitives conceptuelles de l'ontologie sont utilisées pour décrire les sources sous la forme de métadonnées et permettre ainsi un accès unifié au contenu de ces sources.

La mise en place d'un tel environnement soulève de nombreux problèmes liés notamment à la construction de l'ontologie et à la définition d'une architecture logicielle permettant une recherche sémantique et multi-critères des données. Compte tenu de l'état d'avancement du projet (qui a commencé début 2007), nous adresserons principalement le premier problème en présentant la méthode de construction de l'ontologie et la structuration qui en résulte.

2 Approche ontologique

La construction d'OntoNeuroLOG place ses concepteurs d'emblée devant un problème de taille, lié au nombre de primitives conceptuelles devant être intégrées dans l'ontologie. La modélisation des données (images et outils de traitement d'images) fait en effet apparaître des entités de nature très variée.

Pour les images, il convient de prendre en compte les matériels d'acquisition (ex : radiographie, IRM), les protocoles d'acquisition (ex : images de diffusion, de perfusion), les formats d'encodage (ex : DICOM, GIS, Analyze), la dimensionnalité (ex : 2D, 3D), la résolution (spatiale, temporelle, radiométrique), les régions d'intérêt délimitées par différents annotateurs ainsi que les entités explorées (structures anatomiques, fonctions physiologiques).

Pour les outils de traitement d'image, il convient de prendre en compte la nature logicielle de l'outil (ex : programme isolé, pipe-line), le type de traitement réalisé (ex : segmentation, calage, visualisation), le type d'image accepté en entrée et produit en résultat (ex : masque, champ de déformation) ainsi que les conditions d'utilisation (ex : paramétrage, ressources logicielles nécessaires).

Pour rendre compte de cette hétérogénéité importante d'entités, en termes de statuts ontologiques, l'approche retenue consiste à concevoir une ontologie modulaire et multi-niveaux (Temal *et al.*, 2006), dont la structure est présentée sur la Figure 1. L'approche repose notamment sur deux principes importants de conception :

- La réutilisation de l'ontologie de haut-niveau « foundational » DOLCE (Masolo *et al.*, 2003) et la réutilisation et/ou la conception d'ontologies noyau « core ontologies » (Gangemi & Borgo, 2004) de domaines comme ceux des documents (I&DA (Fortier & Kassel, 2004)) et des programmes informatiques (COPS (Lando *et al.*, 2007)). Ces ontologies générales servent à structurer les

domaines plus spécifiques des images médicales et des outils de traitement d'images.

- Une gestion simultanée de trois manifestations de l'ontologie, correspondant à l'usage de formalismes offrant des puissances d'expression différentes : OntoNeuroLOG est ainsi spécifiée au moyen du langage semi-informel de la méthode OntoSpec (Kassel, 2005), en OWL et sous la forme d'un modèle relationnel de base de données. Ces deux dernières manifestations « opérationnelles » sont exploitées informatiquement à des fins de stockage et de recherche sémantique des données.

Une telle démarche de conception vise à maîtriser plusieurs sources de complexité :

- Une complexité de *modélisation conceptuelle*, en permettant de modéliser des objets complexes comme le sont les images et les outils de traitement d'images, à différents niveaux d'abstraction. Ainsi, une *image médicale* est-elle un *document* auquel il correspond des *inscriptions* (sur fichier, papier ou écran), une *expression* dans un certain *format* et un *contenu* conceptuel. De même, un *outil de traitement d'image* est un *programme* spécifié dans un certain *langage de programmation* et permettant de réaliser un certain *traitement de l'information*.
- Une complexité de *conception*, en permettant de réutiliser des modules ontologiques déjà conçus et évalués dans le cadre d'autres ontologies d'application et, également, de travailler de façon distribuée à la conception de nouveaux modules.
- Une complexité de *maintenance*, en conservant une ontologie semi-informelle plus riche sur le plan sémantique et servant de documentation pour les ontologies opérationnelles.

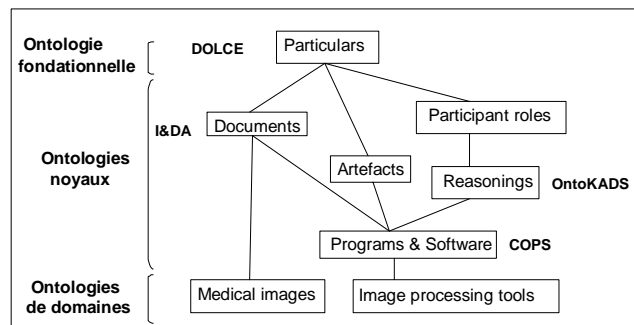


Fig. 1 – Structure de l'ontologie OntoNeuroLOG

3 Architecture logicielle

Une vue d'ensemble de l'architecture prévue¹ pour l'environnement informatique est présentée en Figure 2. L'environnement offrira aux utilisateurs deux fonctionnalités principales : 1) la recherche sémantique multi-critères d'images ; 2) le

¹ Cette architecture est en cours de spécification.

traitement d'images à partir d'outils sélectionnés, utilisant les services d'une infrastructure de grille. Ci-dessous, nous fournissons quelques informations sur les outils logiciels retenus pour implanter une recherche sémantique.

Ces outils, au nombre de quatre, s'organisent selon le principe général suivant : une base de données relationnelle est exploitée pour l'intégration des métadonnées tandis que l'ontologie en OWL est utilisée à des fins d'inférences. La coopération prévue entre les outils est la suivante (voir Figure 2 ci-dessous) :

- L'**interface de requête** permet à l'utilisateur de formuler de manière simple et conviviale ses recherches d'images médicales. Elle convertit ces recherches en requêtes SPARQL².
- Les requêtes SPARQL sont exécutées par **CORESE**³ qui raisonne sur un ensemble de métadonnées décrivant les images au format RDF. CORESE (Conceptual Resource Search Engine) est un moteur de recherche sémantique, qui implante (dans le formalisme des graphes conceptuels) le formalisme RDF(S) et une partie de OWL Lite. Il apporte en complément un langage de règles.
- À l'autre bout de la chaîne, les images médicales que l'utilisateur souhaite visualiser, et leurs métadonnées associées, sont stockées dans des sources de données hétérogènes et réparties (bases de données relationnelles, architectures de répertoires,...). L'outil **Data Federator**⁴ est utilisé pour intégrer toutes ces données au travers d'un schéma relationnel unique.
- Enfin, l'outil **METAmorphoses**⁵ est utilisé pour faire le lien entre le schéma relationnel unique de Data Federator et les méta-données au format RDF. METAmorphoses permet ainsi de peupler l'ontologie RDF manipulée par CORESE à partir des méta-données pointées par Data Federator.

Cette chaîne logicielle fait actuellement l'objet de tests.

² SPARQL est un langage de requêtes RDF conçu par le W3C Data Access Working Group (<http://www.w3.org/TR/rdf-sparql-query/>).

³ CORESE est développé par le projet ACACIA à l'INRIA Sophia-Antipolis (<http://www-sop.inria.fr/acacia/soft/corese/>).

⁴ Data Federator est un produit commercial de la société BusinessObjects (<http://www.businessobjects.com/products/dataintegration/datafederator.asp>).

⁵ METAmorphoses est développé par le Webing Research Group de la Czech Technical University à Prague (<http://metamorphoses.sourceforge.net/>).

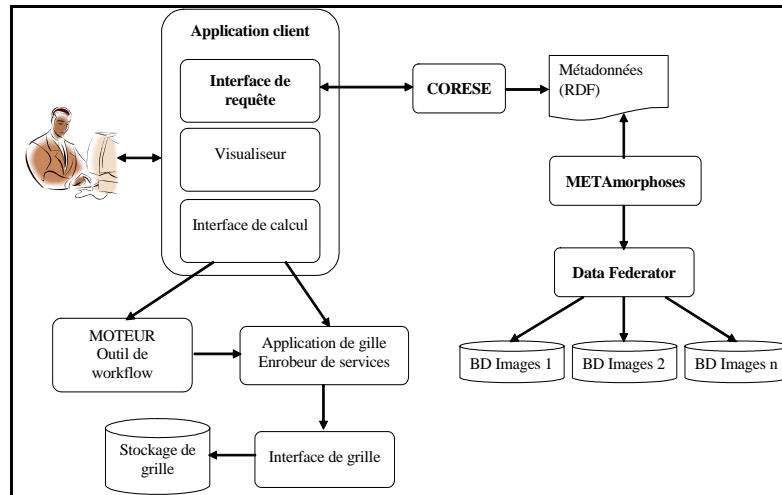


Fig. 2 – Architecture logicielle

Au cours de la journée, nous nous proposons de préciser la démarche de conception d'OntoNeuroLOG, de fournir un retour d'expérience de conception d'une première version de l'ontologie et, également, d'esquisser l'architecture logicielle mise en place pour permettre une recherche sémantique de données distribuées avec l'ontologie.

Références

- FORTIER J.-Y. & KASSEL G. (2004). Managing Knowledge at the Information Level: an Ontological Approach. In *Proceedings of the ECAI'2004 Workshop on Knowledge Management and Organizational Memories*, Valencia (Spain), August 2004, p. 39-45.
- GANGEMI A. & BORGIO S. (eds.) (2004). *Proceedings of the EKAW'04 Workshop on Core Ontologies in Ontology Engineering*, Northamptonshire (UK), October 2004. <http://ceur-ws.org> (Vol-118).
- KASSEL G. (2005). Integration of the DOLCE top-level ontology into the OntoSpec methodology. LaRIA Research Report 2005-08, 2005. Disponible à : <http://hal.ccsd.cnrs.fr/ccsd-00012203>.
- LANDO P., FÜRST F., KASSEL G. & LAPUJADE A. (2007). Premiers pas vers une ontologie générale des programmes informatiques. In *Actes des 18ièmes Journées Francophones d'Ingénierie des Connaissances : IC 2007*, Grenoble, juillet 2007.
- MASOLO C., BORGIO S., GANGEMI A., GUARINO N., OLTRAMARI A. & SCHNEIDER L. (2003). The WonderWeb Library of Foundational Ontologies and the DOLCE ontology. *WonderWeb Deliverable D18, final report* (vr. 1.0, 31-12-2003).
- TEMAL L., LANDO P., GIBAUD B., DOJAT M., KASSEL G. & LAPUJADE A. (2006). OntoNeuroBase: a Multi-Layered Application Ontology in Neuroimaging. In *Proceedings of the 2nd Workshop: Formal Ontologies Meet Industry: FOMI 2006*, Trento (Italy), December 2006.