

Projet scientifique : Sido

Sémantique et Intermédiation de Données

Animateur : Jean-Yves LAFAYE
Anim. adjoint : Frédéric BERTRAND

Points d'entrée du projet scientifique :

Le projet scientifique « Sémantique et Intermédiation de Données » s'organise en deux thèmes dont les principaux centres d'intérêt sont énumérés ci-dessous. Une présentation plus détaillée est fournie au paragraphe « Problématique scientifique »

Métier, organisations, ontologies :

- Génération d'ontologies à partir de modèles métier et d'organisations :
 - Extraction d'informations à partir de modèles métier et de métadonnées normalisées
 - Définition d'annotations formelles de modèles
- Structuration et persistance de données ontologiques
- Audit et métrologie de l'usage des ontologies :
 - Définition de métriques
 - Evolution des ontologies
- Applications spécifiques : analyse d'images, de documents ...
 - Traitement de l'information et ontologies
 - Modèles et méthodes symboliques

Aspects dynamiques

- Prise en compte de la dynamique des données pour leur structuration
- Gestion de flux de données spatio-temporelles
- Intermédiation de données pour la mobilité

Mots clés du projet scientifique :

Ontologies ; ingénierie dirigée par les modèles ; systèmes d'intermédiation ; métadonnées ; qualité des ontologies, bases de données ontologiques ; interrogation d'ontologies ; modèles symboliques ; mobilité.

Acteurs du projet:

Pr. : Jean-Yves LAFAYE, Georges LOUIS
MCF HDR : Alain BOUJU
MCF : Frédéric BERTRAND, Christophe DEMKO, Marie-Christine LAFAYE, Jamal MALKI
PAST : Yanys WESSELS

Doctorants :

Thèse en cours : Guillaume HILLAIRET, Salvijus LAUCIUS, Manoranjan MUNIANDI, Supparang RUANGVANICH, Antoine WIEDEMANN,

Thèse soutenue : Jean-Michel FOLLIN - 2005 (MCF ESGT, Le Mans)

Chercheurs associés à ce projet:

PR : Michel AUGERAUD (Imagin)
MCF HDR : Mike EBOUEYA (Imagin)
MCF : Karell BERTET (i-MeDoc), David SARRAMIA (Imagin), François Role (Imagin)

Problématique scientifique :

I - Présentation générale

Le projet SIDO s'organise autour de deux thèmes de recherche principaux, qui abordent sous différents aspects la question générale de **l'expression rigoureuse de la sémantique des données**. La rigueur est celle des modèles de référence, des méthodes et de la démarche qui les met en œuvre pour **extraire, synthétiser, exprimer et transformer l'information** d'un domaine, d'un système d'information ou d'un corpus de données. La sémantique inclut à la fois les **concepts pertinents dans le contexte de l'utilisateur et les contraintes d'intégrité et propriétés invariantes** intrinsèques au domaine étudié.

Les applications envisagées sont principalement de deux ordres. Dans un contexte d'utilisateurs appartenant à des organisations réparties et hétérogènes, les recherches concernent **l'intermédiation de données**. Dans le cadre d'un domaine homogène particulier réunissant des utilisateurs aux préoccupations communes, elles concernent **la production, la validation et l'utilisation de structures d'ontologies ou de thesaurus** pour aider à structurer les connaissances du domaine en question.

II – Spécificités locales

Telles que décrites ci-dessus, les préoccupations affichées sont partagées par de nombreux autres laboratoires dont en particulier les laboratoires partenaires du projet SIDO (GEOMER Brest, LINA Nantes, LIRMM Montpellier, LISI Poitiers). SIDO développe cependant plusieurs spécificités qui assurent l'originalité des recherches menées.

II.1 – Exploiter les modèles du domaine et les méta-données

La première spécificité correspond au thème « Métiers, Organisation, Ontologies », affiché dans le projet. Il s'agit de la **volonté d'assister l'extraction de la sémantique et plus particulièrement la génération d'ontologies en s'appuyant sur les connaissances métier** telles qu'elles existent déjà en partie dans les systèmes d'information des organisations considérées. En effet, ces informations contextuelles sont capturées par **analyse de modèles UML du domaine annotés et par l'exploitation de bases de métadonnées normalisées** : (norme ISO 19100, DublinCore). Une telle pratique nécessite à l'évidence de compléter les informations natives présentes dans les organisations concernées par des informations plus riches ou plus précises. Les modèles du domaine et les modèles de métadonnées sont alors enrichis par annotation de diagrammes UML pour constituer des profils spécifiques fondés sur des stéréotypes adaptés.

II.2 – Ingénierie dirigée par les modèles et qualité des produits

L'essentiel de nos travaux est développé dans le contexte de l'ingénierie dirigée par les modèles et donc de la modélisation, méta modélisation et transformation de modèles. Nous nous référons à différents standards (UML, CWM, OWL, RDF, Topic Maps), différents langages (ATL, Kermeta) et différentes logiques (des prédicats, modales, de description). **Nous traitons de chaque étape du cycle de conception des systèmes d'information et des ontologies** ce qui inclut non seulement les raffinements successifs de modèles abstraits, mais aussi l'étude de l'adéquation des langages de requêtes aux structures de données persistantes retenues. Dans toutes ces étapes successives, nous privilégions la traçabilité, la vérification formelle des contraintes et la validation des solutions proposées.

Les processus de traitement sont clairement séparés en une partie automatique fondée sur la transformation explicite de modèles ce qui de fait en assure la correction, et une partie assistée dans laquelle l'utilisateur intervient suivant des protocoles rigoureusement spécifiés et contrôlés ce qui garantit le respect des contraintes d'intégrité du système.

Pour mettre en œuvre une approche qualité intégrée, nous travaillons également à spécifier des métriques pour évaluer nos solutions de manière rigoureuse. La qualité n'est pas une propriété statique, elle concerne également l'adaptation à un contexte de connaissances au sein d'organisations qui évoluent et s'enrichissent. La prise en compte de la dynamique est explicitement évoquée – plus largement - comme deuxième centre d'intérêt dans les thèmes de SIDO.

II.3 – Aspects dynamiques : Invariants locaux, flux de données

Sémantique et ontologies évoquent *a priori* l'idée d'informations stables et peu évolutives. Les aspects dynamiques sont pourtant – à juste titre - très présents dans nos travaux.

L'exploitation des diagrammes de classes du domaine et des métadonnées conduit à exprimer des propriétés statiques. Conjointement, l'exploitation des autres diagrammes UML, par exemple les diagrammes d'états, induit **l'expression de propriétés locales liées aux états** des objets considérés. Ces contraintes sont fondamentales pour la garantie de l'intégrité des modèles abstraits et tout autant pour celle de leurs raffinements successifs jusqu'aux modèles physiques. La littérature atteste de la rareté des travaux sur ces questions que nous privilégions. Dans le même esprit, nous nous intéressons à d'autres formes de représentation de la sémantique, en particulier par des recherches fondamentales sur les structures combinatoires, les données symboliques et les connaissances imprécises.

L'intérêt pour la dynamique nous porte à **nous intéresser également aux flux de données** intervenant lors de la mise à jour et de l'interrogation des systèmes d'information. Sur ce sujet, les questions abordées sont de deux ordres.

Le premier concerne la gestion efficace de flux de données importants et continus, en concurrence avec des flux de requêtes. Les réponses mettent en jeu l'optimisation des protocoles et des processus ainsi que celle des architectures techniques et logicielles.

Le second concerne la nature des flux qui dans un contexte de mobilité et d'informatique embarquée, sont soumis à des contraintes (bande passante, temps réel) et présentent des caractéristiques (données éphémères, requêtes imprécises et réponses incertaines) qui s'avèrent incompatibles avec les paradigmes à partir desquels ont été développés tant les systèmes de base de données que les techniques statistiques classiques.

II.4 – Applications et interactions avec les autres projets L3i

Nous appliquons naturellement nos propositions à la fois sur des cas d'école (*benchmarks*) mais aussi sur des cas réels. En particulier, les interactions avec les deux autres projets L3i traitant de l'analyse d'images « ImageIN » et de documents « IDMN » sont des facteurs de synergie importants. Plus précisément, **l'interaction entre la production de systèmes de reconnaissance de formes et celle de systèmes d'interprétation à base ontologique est un domaine que nous souhaitons clairement afficher**. Ces interactions posent par exemple la question d'unifier grâce à des correspondances explicites les pratiques et modèles de connaissance propres au domaine de la reconnaissance des formes (classification, segmentation, apprentissage, modèles stochastiques, possibilistes, géométriques, symboliques, intelligence artificielle,...) avec les modèles de connaissance structurels, logiques, formels ou semi-formels propres aux approches ontologiques.

Faits marquants de l'activité :

Le projet réunit des chercheurs qui se trouvaient auparavant dispersés au sein de diverses thématiques du L3i dans son ancienne organisation. La nouvelle structure du L3i et la création de SIDO prend acte d'une convergence d'intérêts scientifiques précis autour des aspects « sémantique des données » et « intermédiation ». La complémentarité des compétences et la diversité des expériences des chercheurs impliqués permettent une approche novatrice et pluridisciplinaire des problèmes, qui se veut caractéristique du projet SIDO. Plus précisément, les expériences ainsi mises à profit relèvent des technologies logicielles (ingénierie dirigée par les modèles, spécification formelle, qualité du logiciel, gestion de projet, outils du Web sémantique...), des systèmes d'information géographique (spatio-temporel) et des architectures pour la mobilité...

La participation active au pôle de recherche PRIDES crée un terrain propice aux collaborations inter-laboratoires, et à ce titre des échanges réguliers sont à l'œuvre avec le laboratoire LISI (Poitiers) dans le domaine des ontologies de données techniques industrielles et des bases de données ontologiques techniques

En 2005, un partenariat a explicitement été développé avec le laboratoire GEOMER (UBO, Université de Bretagne Occidentale) pour étudier la conception d'un système d'information réparti entre l'IUEM (Brest) et l'institut du littoral (ILE) La Rochelle. Le projet impliquait également l'IFREMER

(L'Houmeau), l'observatoire régional de l'environnement ORE et bénéficiait du conseil des membres du projet SYSCOLAG (LIRMM)

Actuellement, des liens scientifiques forts, des convergences de points de vue et des réunions régulières nous attachent au laboratoire LINA (Nantes) autour des techniques de transformation de modèles. Les contributions de SIDO, actuellement prioritairement orientées vers des modèles UML, peuvent largement être généralisées à des contextes plus larges où la connaissance a priori est exprimée dans des langages ad hoc (DSI). La convergence des intérêts nantais et rochelais sur le sujet se comprennent dans la mesure où SIDO, apporte des exemples concrets et complets (analyse / conception / implémentation) dans le contexte UML, alors que l'équipe ATLAS de J. Bézivin propose des outils, une infrastructure et des concepts généralisant nos travaux dans le cadre de l'extraction, de la transformation et du tissage de modèles.

Bilan de l'activité :

1) Rayonnement scientifique :

L'ensemble des chercheurs participe aux activités des GDR : I3, SIGMA et ALP.

F. Bertrand et A. Bouju, membres du projet SIDO sont animateurs du groupe de travail « Services Localisés » du GDR SIGMA

J. Malki est co-animateur avec M. Boughanem de l'axe « Recherche d'information dans les documents multimédia » du thème 4 du GDR I3 « Données hétérogènes, distribuées et massives ».

Plusieurs membres sont impliqués dans l'animation du pôle de recherche PRIDES en particulier comme acteur ou responsable des groupes de travail « données » et « Transports / Mobilité ».

M.C.lafaye est membre du comité exécutif de l'association INFORSID depuis 1994 et a participé aux comités de programme 2002, 2004,2007

2) Projets labellisés :

Projet RIAM « RelaxMultimédia » 2004-2006 : *Création d'un environnement de production, de stockage et de consultation pour des dépêches d'agence de presse*

Projet PREDIT « Mobivip » 2004-2006 : *Service intégré de mobilité en milieu urbain pour véhicules individuels publics*

3) Projets : local, national et international

Programme interdisciplinaire CNRS « Sociétés de l'Information » 2001-2005 :

- Contribution à la conception d'un système d'information « mémoire environnementale Iroise » en collaboration avec le laboratoire GEOMER de l'UBO
- Conception d'un système pourvoyeur de services à des clients mobiles géoréférencés en collaboration avec le laboratoire OTELO (Géographie) de l'Université de La Rochelle.

4) Animation scientifique ,

Principalement au sein de la fédération PRIDES et des GDR I3 et ALP.

Le travail réalisé sur l'extraction d'ontologies à partir de modèles métiers UML fait l'objet d'une démarche *open-source* en déposant le code réalisé dans la base de modèles de *Eclipse Modeling Project* (<http://www.eclipse.org/m2m/at1/at1Transformations/UML2OWL/UML2OWL.zip>). Cette démarche vise à créer des contacts en vue de travaux collaboratifs (projets, échanges).

Bibliographie

	2003	2004	2005	2006
HDR		1		
Thèses		1		
Ouvrages				
Chapitres de livres				
Revue Internationale	2		3	
Revue Nationale			1	
Conférences Internationales avec actes	2	1	1	
Conférences Nationales avec actes		2	1	3
Conférences Invitées			2	1
Autres conférences				
Rapport de contrat				