Titre :

Modélisation et alignement de modèles de données biologiques avec les technologies du web sémantique

Information :

Encadrant :	Clement Jonquet (LIRMM, UM2) – jonquet@lirmm.fr
Spécialités :	DECOL, AIGLE
Nombre d'étudiants :	2-3
Contexte:	Projet SIFR (Semantic Indexing of French Biomedical Data Resources)
	Projet CR2i Diagnostic Santé, groupe de travail MDR
Ou:	LIRMM, SMILE & TEXTE research team
Quand:	2nd semestre 2012-2013
Contexte :	

Mots clés :

Données biologiques, modèles standards, eCore, UML, OWL, ontologies, représentation de connaissances, web sémantique, intégration de données

Résumé :

L'interopérabilité sémantique des données biologiques et primordiale pour permettre de nouvelles découvertes scientifiques par croisement et analyses des données. Dans le cadre d'un projet de création d'une plateforme d'hébergement de données multi-omiques, nous nous intéressons au rapprochement des modèles FuGE et ISA-Tab qui sont très largement utilisés par les biologistes. L'objectif du TER est de définir formellement en OWL les ontologies correspondantes a ces modèles et de formaliser leur alignements. Nous nous intéresserons également aux liens avec d'autres ontologies biomédicales existantes et aux processus pour représenter les données qui suivent ces modèles au format du web sémantique.

Contexte :

Les différentes approches « omiques »[i] produisent de grandes quantités de données biologiques qui font l'objet de multiples traitements informatique : intégration avec des données externes, analyses statistiques, analyses sémantique, annotation fonctionnelle, représentation dans des bases de connaissances, fouille, etc. L'objectif de ces traitements est la découverte de nouveaux faits scientifiques. Découvertes rendues possibles par le croisement de divers jeux de données représentées dans des formats standards et interopérables.

Ainsi, le système de collecte et de préservation des données constitue l'épine dorsale de cette démarche scientifique avec pour objectif principal l'interopérabilité des données. Les terminologies et ontologies biomédicales jouent un rôle clé dans l'interopérabilité sémantique des données des sciences du vivant en servant de dénominateur commun. L'utilisation d'ontologies pour indexer et intégrer les ressources de données est un moyen de valoriser la connaissance d'un domaine en facilitant la recherche d'information et la fouille de données (translational bioinformatics research). En outre, une fois ces données indexées et annotées sémantiquement les ontologies facilitent leur valorisation grâce aux méthodes formelles de raisonnement ou de classification.

La plateforme SIDR (http://sidr-dr.inist.fr) [5] est un « entrepôt de données » qui permet de colliger les données produites par différentes plateformes multi-omiques sous des formats standards adoptés par la communauté internationale pour en assurer l'interopérabilité avec d'autres données, notamment publiques [4]. Dans la perspective de l'évolution de cette plateforme nous nous intéressons au

rapprochement des modèles FuGE [1] et ISA-Tab [3] qui sont très largement utilisés par les biologistes respectivement dans le domaine de la génomique fonctionnelle et des expérimentations multi-omiques.

Présentation du sujet :

Bien que largement utilisés, les modèles FuGE et ISA-Tab ne sont pas encore décrits dans des langages formels de modélisation. Leurs spécifications sont décrites dans les documents suivants :

http://fuge.sourceforge.net/dev/index.php#v1Final

http://isatab.sourceforge.net/format.html

En collaboration avec le LGI2P (Pr. Vincent Chapurlat), nous avons récemment attaqué une représentation de ces modèles à l'aide du meta-modèle eCore (http://www.eclipse.org/modeling/emf/).

Le travail principal de ce TER consistera à formaliser les ontologies de ces deux modèles. Une fois ces deux ontologies définies nous nous intéresserons aux alignements possibles entre ces ontologies et au portage de jeux de données sous forme d'instances de ces ontologies. Il s'agira donc de :

1. Définir les ontologies des modèles FuGE et ISA-Tab en OWL (Web Ontology Language) qui est le langage de référence pour la définition d'ontologie sur le web sémantique. Ce travail sera fait directement à partir des spécifications textuelles fournies, mais aussi en utilisant la version eCore que nous aurons à disposition. L'utilisation du logiciel Protégé (http://protege.stanford.edu/) sera fortement indiquée.

2. Proposer des alignements formels entre ces ontologies ainsi que d'autres ontologies du domaine. Nous nous intéresserons éventuellement à définir un modèle de haut niveau simplifié (utilisant d'autres ontologies du domaine) qui intègrerai et relierai les deux ontologies. Pour ce travail nous utiliserons la plateforme BioPortal (http://bioportal.bioontology.org) [2]. Le travail de liaison/alignement entre les ontologies pourra être semi-automatisé.

3. Ecrire les scripts qui permettront de passer des données XML aux formats FuGE et ISA-Tab vers le format RDF/OWL utilisé à la fois pour représenter les modèles et les instances. Nous fournirons un jeu de données pour valider les scripts et il s'agira de transformer ce jeu de données en base de connaissances (ontologie + instances).

4. Développer un prototype d'un plugin pour Protégé qui servira d'interface graphique de saisie pour les biologistes qui souhaiteront utiliser les formats FuGe et ISA-Tab.

5. S'intéresser à l'interconnexion des bases de connaissances ainsi crées pour les jeux de données FuGE et ISA-Tab avec des jeux de données déjà présents dans le Web de données « Linked Open Data » (http://linkeddata.org)

Remarque sur l'historique des formats FuGE et ISA-Tab : La spécification FuGE est déduite de MAGE-OM, validé par l'OMG en 2003 : http://www.omg.org/cgi-bin/doc?formal/03-02-03. Le MAGE-OM initial a été transformé dans un format tabulaire (MAGE-TAB) plus facile à lire par la communauté des biologistes (http://www.mged.org/mage-tab). A partir de MAGE-TAB, un framework plus général a été extrapolé : ISA-Tab.

Références :

[1] A. R. Jones, A. L. Lister, P. W. Leandro Hermida, M. Eisenacher, K. Belhajjame, F. Gibson, P. Lord, M. Pocock, H. Rosenfelder, J. Santoyo-Lopez, A. Wipat, and N. W. Paton. Modeling and Managing Experimental Data Using FuGE. *OMICS: A Journal of Integrative Biology*, 13(3):239–251, June 2009.

[2] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. B. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, and M. A. Musen. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37((web server)):170–173, May 2009.

[3] P. Rocca-Serra, M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, S. Neumann, P. Sterk, W. Tong, and S.-A. Sansone. ISA software suite:

supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26(18):2354–2356, August 2010.

[4] S.-A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler, K. Begley, T. Booth, L. Bougueleret, G. Burns, B. Chapman, T. Clark, L.-A. Coleman, J. Copeland, S. Das, A. de Daruvar, P. de Matos, I. Dix, S. Edmunds, C. T. Evelo, M. J. Forster, P. Gaudet, J. Gilbert, C. Goble, J. L. Griffin, D. Jacob, J. Kleinjans, L. Harland, K. Haug, H. Hermjakob, S. J. H. Sui, A. Laederach, S. Liang, S. Marshall, A. McGrath, E. Merrill, D. Reilly, M. Roux, C. E. Shamu, C. A. Shang, C. Steinbeck, A. Trefethen, B. Williams-Jones, K. Wolstencroft, I. Xenarios, and W. Hide. Toward interoperable bioscience data. *Nature Genetics*, 44:121–126, January 2012.

[5] A. Zasadzinski, M.-C. Jacquemot, F. Mazur, D. Fleury, Y. Berchi, M. Mechref, C. Niederlander, and M. Roux. SIDR, a Public Data Repository for Multi-assay Experiments: Issues on Metadata Biocuration. In *Journées Ouvertes en Biologie, Informatique et Mathématiques, JOBIM'11*, Paris, France, Juin 2011.

[i] Omique : branche de la biologie qui étudie l'ensemble des éléments moléculaires d'un type donné (par exemple, la génomique étudie l'ensemble du génome, la protéomique étudie l'ensemble des protéines, etc