

Création et utilisation d'ontologies : de l'ingénierie à la protéomique

Stéphane Jean, Patrick Girard et Ladjel Bellatreche

LIAS/ISAE-ENSMA et Université de Poitiers, Futuroscope, France
{jean,girard,bellatreche}@ensma.fr

Résumé

Proposition pour l'Axe 2 : Analyse de grands réseaux biologiques et intégration des connaissances « omics »

Un des centres d'intérêt de l'équipe ingénierie des données du laboratoire LIAS est la spécification, la représentation et l'intégration des données fortement structurées, telles qu'on les rencontre en particulier dans le domaine technique (par exemple, l'ingénierie automobile ou aéronautique). Ces travaux s'appuient sur des descriptions formelles, des ontologies, qui sont utilisées pour automatiser traitements et conversions de données. Parmi les thématiques évoquées dans l'atelier Prospectom, les problématiques suivantes semblent proches de celles abordées dans nos travaux :

- *Description de données selon plusieurs points de vues.* L'analyse d'un protéome implique plusieurs communautés ayant différents points de vue (par exemple, génomique, protéomique ou métabolomique) et objectifs concernant les données manipulées. Cette problématique se retrouve également en ingénierie puisque les composants industriels sont manipulés par différentes disciplines nécessitant de les décrire selon plusieurs points de vue (par exemple, une vue solide ou une vue comportement). Cette particularité, nous a conduit à définir un modèle d'ontologies spécifique au domaine de l'ingénierie (modèle PLIB, norme ISO 13584) qui permet en particulier la représentation multi-points de vues au sein d'une ontologie.
- *Intégration de données.* L'analyse d'un protéome nécessite de confronter et intégrer plusieurs sources de données contenant le résultat de précédentes analyses ou des données de référence. Dans le domaine de l'ingénierie, plusieurs fournisseurs de composants sont généralement impliqués dans la construction de systèmes. Chacun disposant de sa propre source de données, il est nécessaire de réaliser l'intégration de ces sources. Nous avons proposé et mis en place une approche d'intégration automatique basée sur des services et des ontologies normalisées qui permet à chaque fournisseur de spécialiser cette ontologie pour décrire ses composants d'une manière particulière. Cette approche repose sur des hypothèses fortes valides en ingénierie (par exemple, la structuration forte des données et la présence d'ontologies normalisées).
- *Masse de données.* L'ensemble des protéines du vivant est très vaste et ainsi leur représentation informatique nécessite de grandes bases de données et de connaissances (par exemple, Uniprot est de l'ordre de 220 GO). Dans le domaine de l'ingénierie également le volume de données à traiter est important dû aux très grands nombres de composants industriels impliqués dans la construction de gros systèmes. De tels volumes de données ne pouvant être traités en mémoire centrale nous avons conçu une base de données nommée OntoDB qui permet de stocker des données décrites par des ontologies. Cette base de données est équipée du langage d'interrogation OntoQL qui permet d'accéder aux données au niveau connaissance.

L'exposé que nous proposons aura ainsi deux principaux objectifs : (1) présenter notre expérience sur la construction d'ontologies, l'intégration de données basées sur des ontologies et la gestion de gros volumes de données décrites par des ontologies au sein de bases de données dans le domaine de l'ingénierie et (2) engager la discussion sur les nouvelles problématiques et hypothèses inhérentes au domaine de la protéomique.