Intégration de données hétérogènes dans une base de données graphe de propriétés

Flores R, Confais J, Francillonne N, Rimbert H, Tamby JP, Kreplak J, Imbert B, Toumert Y, Destin J, Bogoin J, Michotey C, Pommier C, Alfama F, Quesneville H, Alaux M

Les recherches menées au sein de l'Institut incluent de plus en plus une formalisation des données sous forme de graphes, notamment pour intégrer des objets biologiques hétérogènes (gènes, produits de gènes, réactions enzymatiques, éléments répétés, phénotypes, connaissances issues de la littérature...). Ces graphes peuvent comporter de très nombreux éléments et être de différents types tels que les graphes de connaissance, les réseaux métaboliques ou les réseaux de régulation. L'avantage d'une telle représentation est qu'elle permet d'échanger entre plusieurs communautés sur des réseaux biologiques divers grâce à un format commun (un ensemble de nœuds et de relations) qu'il est facile d'appréhender. Par ailleurs, cette représentation permet d'y appliquer les algorithmes issus de la théorie des graphes.

L'intégration de données hétérogènes constitue autant un défi scientifique que d'ingénierie. Les questions sur le type de graphe (graphe RDF, graphe de propriétés, *ad hoc*), la modélisation des connaissances, l'usage de vocabulaires contrôlés (ontologies, thésaurus, etc.) ou la provenance des connaissances sont cruciales pour favoriser la compréhension, l'évolution, la réutilisation et la diffusion des résultats au reste de la communauté.

Certaines activités menées dans le cadre du CATI GREP visent à explorer et exploiter l'usage d'une base de données graphe de propriétés (Neo4J) afin de pouvoir adresser des questions complexes en connectant les données entre elles et ainsi répondre à de nouveaux cas d'utilisation. Pour ce faire, la description des concepts du graphe repose en partie sur l'utilisation d'ontologies spécialisées. Cependant des limites ont été constatées lorsque les ontologies existantes ne décrivent pas clairement les nœuds et relations entre les concepts manipulés. La question demeure d'étendre les ontologies existantes, ou d'en proposer de nouvelles, « provisoires », dans l'optique que les communautés s'en emparent ou encore de décider arbitrairement un nommage sans envisager l'intégration dans une ontologie.

Par ailleurs, la centralisation des données hétérogènes permet de contourner les difficultés rencontrées lorsque celles-ci sont distribuées sur différents jeux de données (graphes RDF distribués). Le formalisme graphe de propriétés centralisé en facilite l'exploration par les bio-informaticiens tout au long du processus d'intégration des données, en fournissant un moyen de visualisation et d'interrogation plutôt intuitif. Il permet également la réorganisation du graphe au moyen des algorithmes de la théorie des graphes implémentés au sein de Neo4J et la conservation de la provenance des données.

Si une partie des données publiques est déjà disponible au format RDF, beaucoup de métadonnées omiques restent sous des formats qu'il faut collecter, formater, annoter et lier aux données de référence.

Enfin, les graphes RDF sont des graphes additifs, dans le sens où les triplets ne sont généralement pas modifiés, alors que les graphes de propriétés permettent de manipuler les nœuds et relations et d'y ajouter des propriétés. En revanche, l'un des avantages du RDF étant sa capacité à raisonner sur les données, on atteint avec les graphes de propriété des limites à ce sujet, d'où l'importance de pouvoir *in fine* exporter les graphes construits dans un formalisme RDF.

Nous proposons de faire un état d'avancement des travaux réalisés dans notre groupe de travail et de les confronter aux retours des membres du réseau IN-OVIVE afin d'enrichir notre approche.