ÉDITORIAL

Le web de données (LOD - *Linked Open Data*) est une initiative du W3C, qui consiste en un ensemble de bonnes pratiques pour publier et lier des données structurées (sous le format RDF) dans le web. En utilisant des technologies du web sémantique, des applications peuvent partager, extraire, interroger ou raisonner sur les données publiées.

Le LOD a récemment pris une nouvelle dimension avec la publication de grandes quantités de données (le LOD est passé de 500 millions triplets RDF en 2007 à 130 milliards triplets en 2016). Ces données sont encyclopédiques telles que DBpedia, Yago ou encore Google Knowledge Vault et concernent plusieurs domaines d'application comme les sciences du vivant, la culture et les statistiques. Toutefois, si ces données se retrouvent isolées leur utilité reste très limitée. En effet, un des points angulaires du web de données est le fait que les données soient liées entre elles par des liens sémantiques tels que les liens d'identité (owl:sameAs) qui expriment que deux ressources différentes réfèrent à la même entité (p. ex., même personne, même article, même gène). C'est notamment grâce à ces liens qu'il est possible de développer des applications capables de combiner des données provenant de différentes sources et de naviguer à travers le web de données. Le principe est analogue à celui du web de documents qui tient toute sa puissance des liens hypertextes entre les documents.

Un des éléments essentiels du web de données est l'utilisation de différents vocabulaires pour décrire les données publiées. Ces vocabulaires sont représentés dans des ontologies de façon structurée en associant une sémantique logique qui permet de raisonner automatiquement sur les données et les connaissances. L'utilisation des ontologies pour décrire les données engendre diverses problématiques allant du choix et de la réutilisation de vocabulaires existants à l'alignement et l'interopérabilité des ontologies. Enfin, un des défis qui émergent aujourd'hui est celui de la capitalisation du contenu du web de données qui couvre, notamment, les problèmes d'extraction de connaissances à partir des données incertaines et incomplètes, de raisonnement avec des incohérences, de fusion de données et d'évaluation de la qualité des données et des connaissances dans le web de données.

Les recherches actuelles dans le cadre du web de données visent à concevoir des architectures et à définir des méthodes offrant, à des institutions et à des organismes, la possibilité de publier et de lier leurs données avec celles déjà publiées. Les travaux de recherches dans ce domaine visent également à développer des applications extrayant une forte valeur ajoutée des données liées.

Le but de ce numéro spécial est de réunir des travaux traitant des problématiques telles que la publication des données sur le LOD, le liage de données et la capitalisation des connaissances issues du LOD. Six articles ont été sélectionnés pour ce numéro. Le premier article offre un état de l'art sur le liage de données, le deuxième et le troisième s'intéressent au problème de l'évaluation de la qualité des données et des connaissances dans le LOD. Les deux suivants présentent des exemples d'applications sur les données du LOD : un wiki sémantique et un outil de visualisation d'ontologies. Le dernier article montre un exemple de plateforme d'intégration de données du LOD appliquée à l'agronomie. Dans ce qui suit, nous fournissons un résumé de chacun des articles présentés dans ce numéro spécial de la revue ISI – Ingénierie des Systèmes d'Information.

Manel Achichi, Zohra Bellahsene et Konstantin Todorov présentent dans « A Survey on Data Linking » un état de l'art des méthodes et outils traitant du problème de liage de données. Le processus de liage est considéré dans cette étude comme une chaîne de traitement composée de trois phases : 1) pré-traitement, 2) appariement d'instances et 3) post-traitement. Une classification des approches et des outils dans une (pseudo-) taxonomie en fonction des trois grandes étapes du processus est proposée. Cette classification comprend plusieurs catégories ; en fonction des tâches que chaque approche utilise et selon les techniques qui y sont appliquées. Une quatrième catégorie de méthodes appelée *multi-étapes* est considérée. Celle-ci comprend les méthodes agissant sur plus d'une étape du processus de liage. Les auteurs proposent également une analyse comparative des différentes approches et outils existants dans ce domaine.

Dans « Évaluation de la qualité des sources du web de données pour la résolution d'entités nommées », Carmen Brando, Nathalie Abadie et Francesca Frontini présentent une étude empirique réalisée afin d'évaluer la qualité de jeux de données du web de données en tant que bases de connaissances potentielles pour une application de résolution d'entités nommées dans le contexte des humanités numériques. Pour ce faire, les auteurs s'appuient sur des mesures d'évaluation de la qualité des sources de données du web de données de l'état de l'art mis en œuvre du point de vue de l'adéquation des données à un besoin particulier. Ces mesures ont été testées sur des sources de données de deux types : une source de données du web de données généraliste et d'autres portant sur des domaines plus spécifiques. L'objectif visé était de déterminer s'il est possible d'évaluer a priori quelle source de données serait la plus à même de produire de bons résultats de résolution d'entités nommées dans le cas de textes littéraires en français.

L'article « Interopérabilité sémantique entre vocabulaires contrôlés : Évaluation de la qualité des alignements sur des données de standards du diagnostic in vitro » de Melissa Mary, Lina F. Soualmia et Xavier Gansel s'intéresse à l'intégration des connaissances entre SOC (Systèmes d'organisation des connaissances) qui est une problématique largement étudiée dans des domaines plus ou moins spécialisés (la biologie, et la santé). Ils proposent une évaluation de l'alignement de concepts issus du DIV (le diagnostic *in vitro*) présents dans les SOC de référence disponibles en

ligne. Les méthodes proposées reposent sur trois mesures de similarité syntaxiques et un algorithme heuristique. Les résultats obtenus dans cette étude ont montré que les mesures de similarité syntaxiques ne se révèlent pas suffisamment probantes pour se voir appliquées de manière systématique au domaine des tests de laboratoire. En revanche, la qualité des alignements obtenus via l'algorithme heuristique, filtré *a posteriori* en fonction d'une dimension sémantique, conforte les critères de performance établis par les auteurs.

Yaya Traore, Cheikh Talibouya Diop, Fatou Kamara-Sangare, Sadouanouan Malo, Moussa Lo et Stanislas Ouaro proposent dans « Motifs fréquents pour améliorer la catégorisation dans un wiki sémantique » une approche qui permet d'extraire parmi les tags (les mots-clés) annotant les pages wikis, des motifs fréquents qui guident la découverte de nouvelles catégories et qui améliorent la catégorisation du contenu du wiki. Les auteurs utilisent l'ontologie associée au wiki pour bénéficier de plus d'informations structurées afin de guider l'expert dans la création de nouvelles catégories dans le wiki. Les expérimentations réalisées sur un wiki sémantique avec des pages annotées ont montré que la méthode permet d'améliorer la catégorisation du contenu du wiki et la recherche sémantique par catégorie.

Dans « Un outil de visualisation d'ontologies pour le web des données, utilisable par tous », Fatma Ghorbel, Elisabeth Métais, Nebrasse Ellouze et Faiez Gargouri présentent un outil de visualisation d'ontologies nommé MEMO GRAPH qui permet la visualisation de données ouvertes et liées. Cet outil est conçu pour être utilisé par tous ; les experts du domaine et les utilisateurs non connaisseurs des technologies du web sémantique. Il offre une interface qui illustre le concept du « Design For All » ou « design universel » et a été intégré dans la prothèse de mémoire CAPTAIN MEMO afin de visualiser un jeu de données à petite échelle (PersonLink). MEMO GRAPH a, également, été utilisé en tant qu'application autonome, pour visualiser une partie d'un jeu de données ouvertes liées à large échelle (DBpedia). L'efficacité et l'accessibilité de MEMO GRAPH a été évaluée par des experts du domaine et par des patients atteints de la maladie d'Alzheimer. Les premiers résultats ont montré qu'il est efficace et convivial.

Enfin, l'article « AgroLD API : Une architecture orientée services pour l'extraction de connaissances dans la base de données liées AgroLD » de Gildas Tagny Ngompe, Aravind Venkatesan, Nordine El Hassouni, Manuel Ruiz et Pierre Larmande présente le projet Agronomic Linked Data (AgroLD) qui est une base de connaissances du web sémantique conçue pour intégrer des données provenant de diverses sources de données centrées sur des plantes disponibles publiquement. L'objectif de AgroLD est de fournir un portail web pour les bioinformaticiens et les experts du domaine afin d'exploiter les données homogénéisées et permettre de combler les connaissances dans ce domaine.

Nous remercions les auteurs pour leur contribution et les membres du comité de lecture pour leur participation active dans la relecture des articles de ce numéro.

Fayçal HAMDI
CEDRIC, Cnam Paris

Fatiha SAÏS
LRI, Université Paris-Sud