Linked Open Data : un tour d'horizon

Béatrice MARKHOFF

Web et connaissances : le Linked Open Data (LOD)

Arnaud GIACOMETTI

Le Linked Open Data (LOD) : une source de données pour l'extraction de nouvelles connaissances ?

LIFAT EA 6300
Laboratoire d'Informatique Fondamentale et Appliquée de Tours

université de TOURS

Béatrice MARKHOFF

Web et connaissances : le Linked Open Data (LOD)

Rappelons les objectifs du web à l'origine : relier des documents entre eux et utiliser l'internet pour étendre ces liens entre des ordinateurs distants, permettant le partage, l'échange, la diffusion à grande échelle d'informations. A l'heure actuelle, chacun sait que le web comprend énormément de connaissances utilisables par toute personne disposant d'un navigateur et d'un moteur de recherche. Ce qui est moins connu, c'est que ces connaissances sont en cours d'encodage pour que des logiciels puissent les exploiter automatiquement. Le sens des données affichées dans les pages web (leur *sémantique*) est ainsi représenté puis utilisé, par exemple par les moteurs de recherche qui vous affichent dans un cadre une véritable réponse, en plus de la liste de liens correspondant aux mots donnés.

Le web est également le support des données ouvertes, concept qui désigne des informations publiques librement accessibles. De nombreux gouvernements, institutions, associations et particuliers contribuent au partage et à la diffusion d'information en publiant sur le web des jeux de données. L'analyse et l'exploitation de ces jeux de données reste un challenge, d'une part parce que le sens de ces données n'est pas représenté de façon à les rendre directement utilisables par des logiciels, d'autre part parce que ces jeux de données sont des silos d'informations étanches les uns aux autres.

Je présenterai la réponse à ces verrous, encore une fois apportée par le web : le *Linked Open Data* (LOD) ou « web des données ouvertes et liées », qui est un ensemble de jeux de données dont le sens est formellement représenté, donc interprétable par des programmes, et qui plus est, ces données sont reliées les unes aux autres. Initié en 2007 avec 12 jeux de données, il comprend maintenant 1234 jeux de données reliés par plus de 16000 liens et il ne cesse de s'accroître. Il couvre tous les domaines et certains de ses jeux de données représentent à eux seuls des millions de connaissances.

Béatrice Markhoff *est maître de conférences en informatique, habilitée à diriger des recherches, membre de l'équipe BDTLN du Laboratoire LIFAT et du DI de la faculté des sciences et techniques de Tours. Elle a obtenu son doctorat à l'Université de Franche-Comté en 1995 (sur une implémentation parallèle d'un langage fonctionnel de flux de données) et son Habilitation à Diriger des Recherches à l'Université de Tours (sur la gestion des données du web et l'interopérabilité) en 2013. Ses recherches portent sur la gestion des données du web, la représentation de connaissances, les systèmes sémantiques d'intégration de données, l'interrogation et la fouille de graphes de connaissances. Elle s'investit depuis plusieurs années dans des collaborations avec différentes équipes de recherche en humanités, CITERES-Laboratoire Archéologie et Territoires dans le consortium MASA de la TGIR Huma Num, CITERES-EMAM pour le programme BIBLIMOS et le CESR pour le projet ARVIVA.*

Arnaud GIACOMETTI

Le Linked Open Data (LOD) : une source de données pour l'extraction de nouvelles connaissances ?

Le *Linked Open Data* (LOD) ou « web des données ouvertes et liées » (qui sera introduit dans la présentation précédente par Béatrice Markhoff) constitue aujourd'hui un gisement de données et connaissances en pleine expansion ; il regroupe actuellement plus d'un millier de sources de données (dont DBPedia, Wikidata, Yago, …) comprenant plus d'un milliard de triplets, la plus petite unité de données représentant des association entre sujet, propriété et objet ; par exemple, le triplet `(Alberto Giacometti, is-a, Artist)` représentera qu'Alberto Giacometti est un artiste.

A partir de ce gisement de données, accessible à tout un chacun, de nombreuses enquêtes et analyses de données peuvent être conduites, par exemple, pour étudier pays par pays comment les populations se répartissent dans des villes de plus ou moins grandes tailles. Mais il est alors important de souligner que les sources de données disponibles, très évolutives, construites de manière collaborative, sont souvent encore très incomplètes. Ainsi, de nombreuses informations sont encore manquantes, et le fait que la population

d'une ville ne soit pas renseignée ne permet pas de conclure qu'elle n'a pas d'habitants. Dans ce contexte, il est donc primordial, quand une requête ou analyse est posée sur une base du web des données, de pouvoir évaluer la pertinence et qualité de la réponse apportée.

Après avoir exposé les différents problèmes posés par l'incomplétude des données, nous présenterons quelques travaux menés au sein de l'équipe BdTln du laboratoire LIFAT pour évaluer par exemple, si des données du web sont représentatives ou pas de la réalité (afin d'éviter de réaliser des analyses erronées ou biaisées), si des propriétés du monde réel peuvent être induites des données stockées dans le web des données (par exemple, que tout individu a en général deux parents, une seule date de naissance, …), etc.

***Arnaud Giacometti**, Professeur des universités en informatique, est membre de l'équipe Bases de données et traitement des langues naturelles (BDTLN) du Laboratoire d'Informatique Fondamentale et Appliquée de Tours (LIFAT). Après avoir été co-responsable de l'équipe BdTln (depuis 2010), il est aujourd'hui directeur adjoint du LIFAT avec Jean-Yves Ramel (depuis septembre 2018). Depuis 1995, il est également membre du Département d'Informatique de la faculté des sciences et techniques de Tours. Il a obtenu son doctorat en 1992 à l'École Nationale Supérieure des Télécommunications (aujourd'hui Télécom ParisTech) et son Habilitation à Diriger des Recherches en 2004 à l'Université de Tours (sur les bases de données inductives). Ses recherches portent principalement sur la fouille de données, et plus spécifiquement l'extraction automatique de propriétés dans de grands volumes de données. Suivant le contexte applicatif, les méthodes développées peuvent être utilisées pour extraire des préférences utilisateurs, des biomarqueurs pour le dépistage et le diagnostic médical, des propriétés de connaissances représentées dans le web des données, etc. Plus récemment, ses recherches visent à développer des méthodes de fouille de données centrées-utilisateurs et interactives, prenant au plus tôt en compte ses attentes et préférences.*