HAL
archives-ouvertes.fr

DBpédia.fr : retour sur la publication de données de la culture française

Fabien Gandon, Raphaël Boyer, Alexandre Monnin

▶ **To cite this version:**

Fabien Gandon, Raphaël Boyer, Alexandre Monnin. DBpédia.fr : retour sur la publication de données de la culture française. I2D – Information, données & documents, A.D.B.S., 2016, Web de données et création de valeurs : le champ des possibles, 53 (2016/2), pp.84. <http://www.adbs.fr/i2d-n-2-juin-2016-dossier-web-de-donnees-et-creation-de-valeurs-le-champ-des-possibles-156675.htm?RH=1426693578415>. <hal-01342757>

HAL Id: hal-01342757
https://hal.inria.fr/hal-01342757

Submitted on 6 Jul 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons ttribution - NonCommercial - ShareAlike ttribution - NonCommercial - ShareAlike 4.0 International License

Les projets DBpédia et SemanticPedia illustrent la possibilité de réutilisation des données dans de nombreuses applications grâce à des langages et à des schémas de descriptions qui sont ici expliqués.

Fabien Gandon
Raphaël Boyer
Alexandre Monnin

DBpédia.fr : retour sur la publication de données de la culture française

L'essor du web sémantique et du web des données démultiplie les services disponibles sur le Web pour tous les domaines de la vie quotidienne. Les données jouent désormais un rôle essentiel dans le rayonnement mondial des cultures : qu'une culture ou une langue vienne à être absente ou mal représentée sur le web de données, elle serait mécaniquement absente ou mal représentée à l'échelle des nombreuses applications qui le mobilisent et, par conséquent, marginalisée du point de vue des usages, ces derniers reposant en fin de compte sur ces applications et leurs données. Dans ce contexte, la publication de données devient un axe majeur des politiques culturelles. Aussi, le 19 novembre 2012, M^{me} Aurélie Filippetti, alors ministre de la Culture et de la Communication, a procédé au lancement de la convention Sémanticpédia, entre le ministère de la Culture et de la Communication (MCC), Wikimédia France et Inria. Cette convention a dès lors fourni un cadre collaboratif favorisant l'enrichissement significatif de l'écosystème francophone du web des données.

L'enjeu d'un chapitre francophone de DBpédia

DBpedia.org est un effort communautaire destiné à extraire des informations à partir des pages de Wikipédia afin de les transformer en données disponibles sur le Web dans des formats structurés, les rendant aisément utilisables par des logiciels ou services en ligne. À l'origine, DBpedia.org se concentrait sur les pages de l'encyclopédie Wikipédia anglophone et laissait par conséquent de côté nombre d'informations présentes sur les pages rédigées dans d'autres langues, et tout particulièrement les pages ne disposant pas d'équivalents en anglais. La signature de la convention Sémanticpedia a ainsi permis, à travers les développements auxquels elle a donné lieu, la réalisation d'une version francophone de DBpédia, qui extrait et expose quantité de données de la Wikipédia en français aux formats du web sémantique. Ce projet, le premier de la convention, a aussi débouché sur la création d'une série de services, d'applications et d'outils innovants tirant parti de ces données.

Cinq ans d'expérience sur DBpedia.fr

À l'heure d'écrire cet article, le graphe des données publiées sur DBpedia.fr a dépassé les 185 millions d'arcs (triplets) et atteint un record de 2,5 millions de requêtes, soit une moyenne de 45 000 requêtes par jour (68 700 en 2015). Ces logiciels et leurs configurations sont aujourd'hui stabilisés, opérationnels et en production. Ce niveau de service nécessite une machine virtuelle dédiée correspondant à un CPU Intel Xeon 2,6 Ghz (4 flots de traitement en parallèle) et 40 Go de RAM. Une documentation, des statistiques et des logs d'usage sont également maintenus. Des tutoriels décrivant la manière d'interroger les données et d'augmenter les extractions permettent à des acteurs externes de s'emparer à leur tour de ces résultats.

Fig. Nombre de requêtes SPARQL par jour sur DBpedia.FR (échelle logarithmique)

Les applications de DBpedia.fr

Outre la possibilité d'interroger ces données, la création d'un nouveau chapitre dans le giron de la communauté DBpédia a permis l'intégration et le croisement de nombreux autres jeux de données francophones. Au-delà de la recherche, ce sont toutes les fonctionnalités de gestion des ressources informationnelles qui sont susceptibles de bénéficier de ce réservoir de données ouvertes et de sa large couverture thématique : du filtrage à la notification, de la recommandation à la navigation. Du point de vue des détenteurs de données, DBpédia ouvre de nouveaux moyens d'accès à ces données et en accroît ainsi la valeur. Du point de vue des développeurs, DBpédia fournit des données croisées avec l'écosystème du *web of data*, aptes à alimenter leurs applications et à répondre aux besoins des utilisateurs.

Il est en pratique impossible d'imaginer toutes les applications potentielles qui en découlent. Et c'est précisément là que repose l'enjeu de l'ouverture des données en guise de soutien à l'innovation ouverte et au développement d'usages inédits. À ce titre, un concours dans le cadre de SémanticPédia a d'ailleurs été organisé en mars 2014 à l'initiative du ministère de la Culture et de la Communication et de Wikimédia France afin d'identifier et de récompenser les projets innovants participant à la valorisation de cet écosystème.

De multiples applications ont été réalisées par des acteurs publics et privés, sans oublier les particuliers : le moteur de réponse aux questions en langue naturelle Qakis ; le site Zone47, axé sur la recherche et la consultation d'œuvres d'art ; HdA Lab, le portail du ministère de la Culture consacré à l'histoire des arts, ou encore DiscoveryHub, moteur de recherche exploratoire développé au sein d'Inria pour faciliter la découverte de pans entiers de la connaissance.

Une plateforme d'innovation continue

Outre les données, la maintenance des logiciels et des plateformes assurant le fonctionnement de Dbpédia est le résultat d'une action collective menée par une communauté ouverte. Si la plateforme est libre et son code source ouvert, les collections de règles d'extraction tout comme les vocabulaires utilisés pour structurer les données extraites de Wikipédia sont, quant à eux, documentés et maintenus de façon collaborative *via* des wikis. Ainsi, la configuration et la maintenance d'une plateforme DBpédia résultent-elles d'une activité sociale, fruit d'un consensus auquel chacun peut prendre part (qu'il s'agisse d'ajouter ou de modifier des règles d'extraction et ainsi augmenter le nombre de données extraites ou d'améliorer leur qualité).

Parmi les contributions notables du chapitre francophone à la communauté internationale, la dernière en date n'est autre qu'un extracteur de l'historique d'édition des pages de Wikipédia représentant la bagatelle de 2 milliards de triplets RDF ! Sa publication devrait jeter un éclairage nouveau sur la dynamique de la communauté des « wikipédiens » en permettant d'identifier aisément les sujets les plus consultés, édités, voire à la mode. Sans oublier la possibilité désormais à portée de main de reconstituer le cheminement complexe présidant à l'écriture d'un article sur Wikipédia et, ce faisant, à la publication d'une ressource sur le web de données. La boucle est ainsi bouclée.■