

Des données culturelles accessibles à tous : Sémanticpédia

Développer en français des outils de création, de transmission et de diffusion culturelle conforte le rôle historique de la langue française comme langue internationale de référence en matière de production des savoirs. Il s'agit également d'un enjeu national, lié à l'accès à l'éducation et à la connaissance par les jeunes générations et au partage du patrimoine culturel français à travers sa mise en ligne.

Repères

Délégation générale à la langue française et aux langues de France

De DBpédia à Sémanticpédia

DBpédia est un projet sous licence libre d'extraction des données de Wikipédia anglophone, lancé en 2007, pour en proposer une version Web sémantique structurée. Il est mené par l'université de Leipzig, l'université libre de Berlin (*Freie Universität*) et l'entreprise *OpenLink Software*.

Dans ce cadre, l'Inria a proposé au ministère de la Culture et de la Communication un projet inédit de constitution d'une version entièrement sémantisée de l'encyclopédie Wikipédia francophone. Plus précisément, la version française du *dbpedia.org* a été le premier projet conçu par le biais de cette plate-forme collaborative.

En 2011, une plate-forme de collaboration, dénommée Sémanticpédia, a été ouverte entre le ministère de la Culture et de la Communication (Secrétariat général, Département des programmes numériques et Délégation générale à la langue française et aux langues de France), l'Inria (l'Institut national de recherche en informatique et en automatique) et Wikimedia France pour réaliser des programmes de recherche et développement appliqués à des corpus ou des projets collaboratifs culturels, utilisant des données extraites des projets de Wikimedia.

Son déploiement rapide permettra d'éviter que les applications, mises en place par les entreprises ou les établissements publics francophones, ne s'appuient sur la version anglaise de Wikipédia. Celle-ci comporte, en effet, un nombre beaucoup plus restreint de pages culturelles relatives aux œuvres françaises ou aux personnalités s'exprimant en langue française.



Délégation
générale à la
langue française
et aux langues
de France

6 rue des Pyramides
75001 Paris
Téléphone : 01 40 15 73 00
Télécopie : 01 40 15 36 76
Courriel : dglf@culture.gouv.fr
Internet : www.dglf.culture.gouv.fr
ISSN imprimé : 2260-2887
ISSN en ligne : 2260-4146

Pourquoi une version francophone de DBpédia

Le projet DBpédia en français donnera accès sur la toile, en français, à la diversité du patrimoine culturel. Il permet, en effet, de rendre accessible aux formats du Web sémantique l'ensemble des données structurées figurant sur le portail Wikipédia francophone.

Ce projet ouvre la voie à de nombreuses applications concrètes : par exemple, la mise à disposition du public d'outils de recherche « nativement » multilingues, de géolocalisation automatique des monuments ou, encore, la possibilité d'effectuer des recherches transversales et thématiques sur des centaines de sites culturels. En outre, il positionne la langue française au centre du « Web de données » émergent.

Placé au cœur du Web sémantique du fait de son important volume de données, le projet permettra à tous les sites internet et applications mobiles à vocation culturelle de faire des requêtes sur DBpédia en français pour afficher des contenus de façon transparente et organisée sur leurs pages. Pourront figurer, par exemple, la liste des peintres impressionnistes français, les illustrations par ordre chronologique des 20 premiers monuments classés historiques en Aquitaine au XIX^e siècle...

La plate-forme Sémanticpédia envisage de développer ensuite *Le Wiktionnaire*.

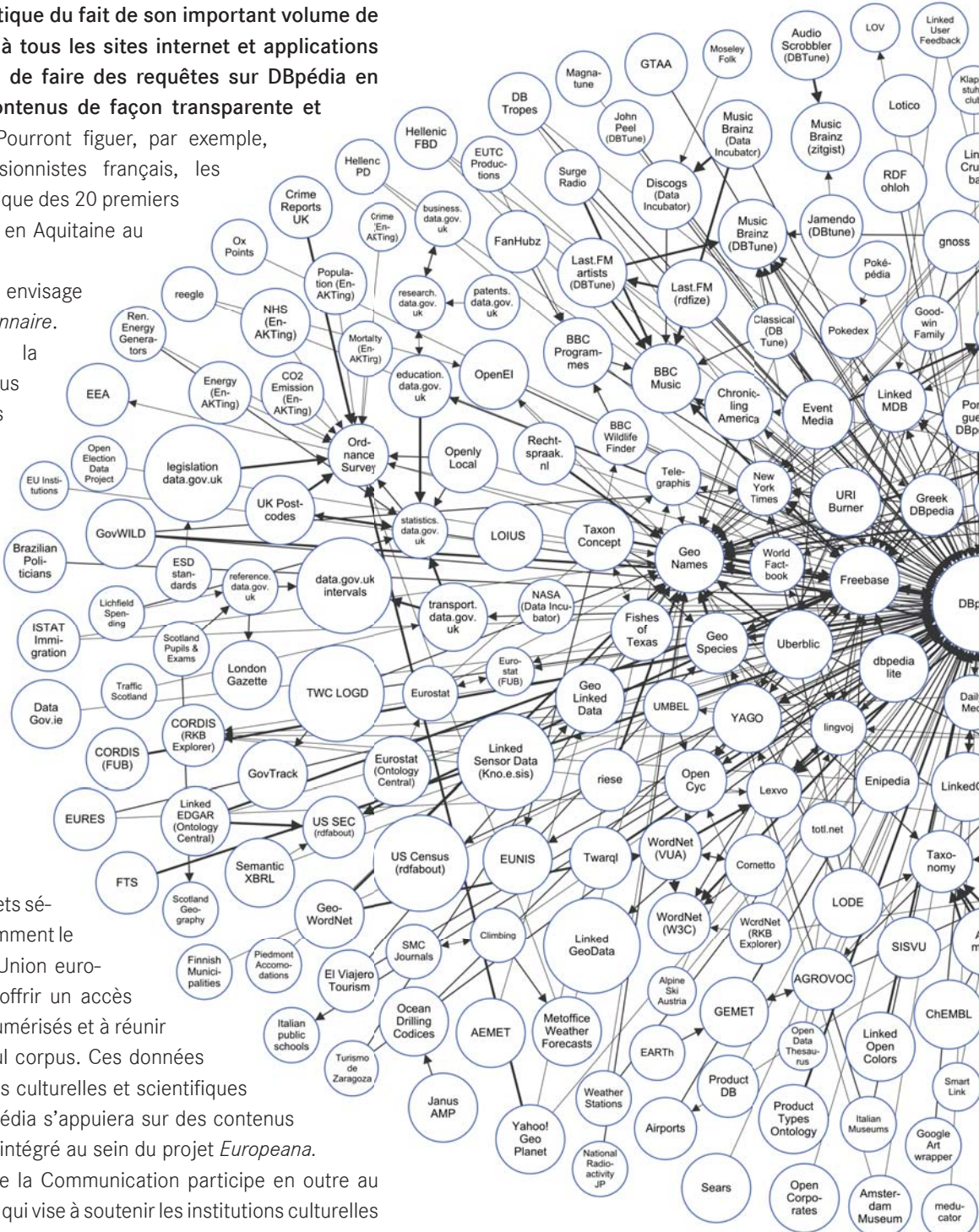
Ce projet collaboratif de Fondation Wikimedia compte plus de 2,2 millions d'entrées dans sa déclinaison linguistique en français, ce qui le place en deuxième position de ce type de dictionnaires, après la déclinaison en anglais. Cet outil offre en effet des possibilités immenses tant il est riche en termes nouveaux ou anciens, et peut donner lieu à de nombreuses applications très innovantes pour valoriser la langue, son utilisation, son enrichissement.

Actuellement, d'importants projets sémantiques se développent, notamment le projet *Europeana* financé par l'Union européenne. Cette initiative vise à offrir un accès ouvert aux contenus culturels numérisés et à réunir toutes les données dans un seul corpus. Ces données sont fournies par des institutions culturelles et scientifiques à travers l'Europe. Comme DBpédia s'appuiera sur des contenus culturels français, il pourra être intégré au sein du projet *Europeana*. Le ministère de la Culture et de la Communication participe en outre au projet européen *Linked Heritage*, qui vise à soutenir les institutions culturelles publiques dans l'optique de fournir leurs données à *Europeana*.

Le Web de données, un enjeu majeur

L'avenir de nos langues se joue donc aujourd'hui **essentiellement sur les réseaux** et l'observation des usages numériques montre combien il est important d'assurer la présence du français et des langues de France dans cet univers, qui est **une chance pour la diversité linguistique et le multilinguisme** : on compte à ce jour 285 déclinaisons linguistiques de Wikipédia, dont les 218 premières comprennent plus de 1 000 articles, alors que l'encyclopédie libre n'était disponible, à sa naissance en 2001, qu'en anglais.

Il importe donc d'être vigilant quant à la diffusion de la culture et des langues sur la toile.

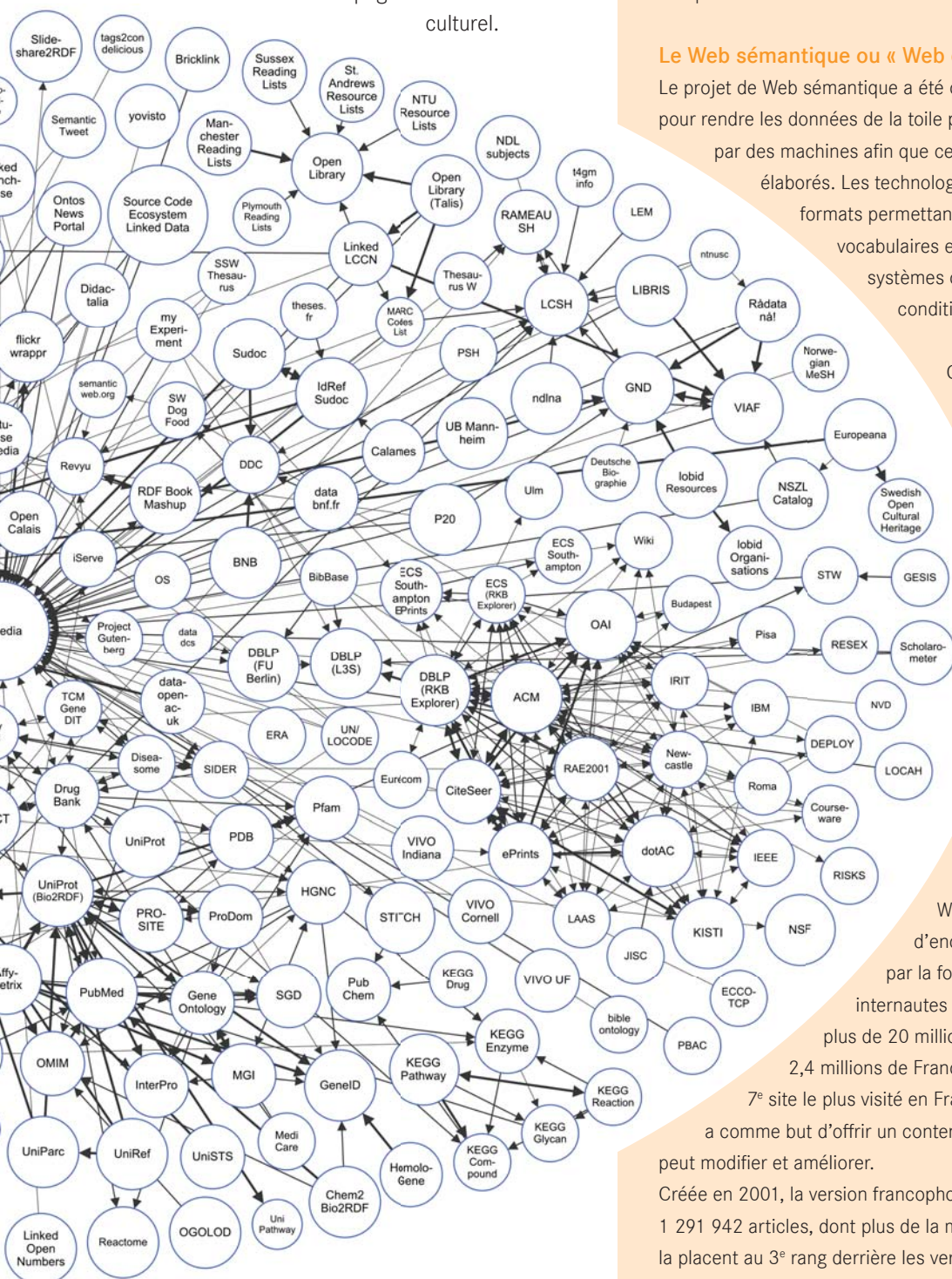


Représentation des relations entre DBpedia et divers

À cet égard, deux enjeux semblent plus particulièrement stratégiques :

> d'une part, le **Web des données** ou « Web sémantique », qui est en passe de devenir le moyen privilégié d'accéder à l'information et à la culture en ligne. Dès lors il importe **d'assurer la visibilité d'une langue et d'une culture** dans un contexte fortement multilingue :

> d'autre part, la mise à disposition de données « ouvertes » et librement réutilisables, afin d'encourager leur dissémination sur les réseaux et notamment sur les portails collaboratifs tels que Wikipédia, dont près de la moitié des pages est à caractère culturel.



Glossaire

Sémantiser

On appelle « sémantiser », dans ce contexte, l'utilisation des standards du « Web de données » sur un corpus existant – ici, les pages de l'encyclopédie Wikipédia – en particulier le contenu des « infoboxes » et des « templates » (métadonnées contenues dans le code wiki).

W3C

Consortium international fondé en 1994, le W3C est l'organisme de normalisation en charge de la gouvernance technique de la toile. C'est en son sein que furent développés les premiers standards qui en forment aujourd'hui encore l'assise, ainsi que le projet de Web Sémantique, plus tard rebaptisé « Web de données ».

Le Web sémantique ou « Web de données »

Le projet de Web sémantique a été conçu à partir du milieu des années 90 pour rendre les données de la toile plus facilement utilisables et interprétables par des machines afin que celles-ci puissent opérer des raisonnements élaborés. Les technologies du Web sémantique proposent des formats permettant de créer ces données, de constituer des vocabulaires et les règles qui les décrivent, et de bâtir des systèmes capables de les manipuler dans de bonnes conditions d'interopérabilité.

Grâce à ces technologies, la toile est en passe de devenir une base de connaissances mondiale. Ce que l'on appelle le « Web de données » consiste en effet à interconnecter d'immenses référentiels (terminologies, catalogues d'œuvres, etc.) ouvrant la voie à des usages radicalement nouveaux des données numériques. En effet, les données sont ici structurées et reliées entre elles, ce qui permet de constituer un réseau global d'informations, ouvrant de nouvelles possibilités pour le multilinguisme.

Wikipédia



Wikipédia est un projet collaboratif d'encyclopédie multilingue sur la toile, proposé par la fondation Wikimedia, rédigée par les internautes et publiée sous licence libre. Consultée par plus de 20 millions de Français différents chaque mois, soit 2,4 millions de Français différents chaque jour, Wikipédia est le 7^e site le plus visité en France. L'encyclopédie mondialement connue a comme but d'offrir un contenu libre, objectif et vérifiable, que chacun peut modifier et améliorer.

Créée en 2001, la version francophone compte actuellement (septembre 2012) 1 291 942 articles, dont plus de la moitié à un contenu culturel. Ces statistiques la placent au 3^e rang derrière les versions anglophone et germanophone.

Sources : Wikipédia, Bibliothèque nationale de France.

Sitographie

semanticpedia.fr

Le site de la plate-forme collaborative réunissant le ministère de la Culture et de la Communication, Wikimédia France et l'Inria.

wikimedia.fr

Le site de l'association Wikimédia France.

wikipedia.fr

Encyclopédie en ligne Wikipédia.

fr.wiktionary.org

Dictionnaire contributif en langue française.

inria.fr

Le site de l'institut national de recherche en informatique et en automatique.

dglff.culture.gouv.fr

Le site de la délégation générale à la langue française et aux langues de France.

fr.dbpedia.org

Le site du projet DBpédia francophone.

dbpedia.org

Le site du projet DBpédia anglais.

bnf.fr

Le site de la Bibliothèque nationale de France.

europeana.eu

Le site de la bibliothèque numérique *Europeana*.

www.linkedheritage.eu

Le site du projet européen *Linked Heritage*.

hdalab.iri-research.org/hdalab

Le laboratoire d'expérimentation du portail Histoire des Arts, *HdA Lab*.

