

wikipedia
dbpedia



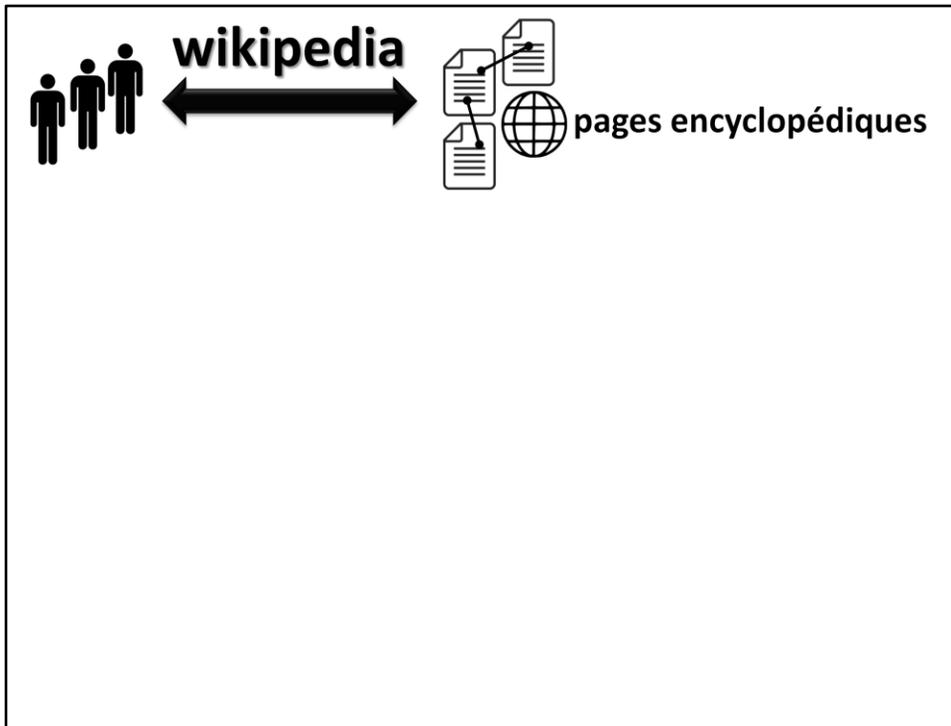
Fabien Gandon, <http://fabien.info> @fabien_gandon



Bonjour à tous.

Il me revient donc de présenter le premier projet de la convention Sémanticpédia qui vient d'être signée.

Il s'agit du projet DBpédia.fr c'est-à-dire le chapitre francophone du projet international DBpedia qui vise à publier sur le web un maximum de données disponibles dans l'encyclopédie Wikipédia et ceci dans des formats standards et ouverts utilisables par les logiciels.



Comme vous le savez Wikipédia est une encyclopédie libre et gratuite sur le web. Le nom de « wiki-pédia » vient du fait que cette encyclopédie repose sur le principe des wikis qui permet à une communauté d'utilisateurs de créer et de maintenir les pages d'un site web. Et cette communauté est impressionnante puisqu'avec plus de 400 millions d'utilisateurs mensuels Wikipédia est aujourd'hui le cinquième site web le plus visité au monde avec plus de 23 millions d'articles dans de multiples langues.

The diagram illustrates the concept of 'un sujet par page' (one subject per page) on Wikipedia. On the left, three human icons are connected to the word 'wikipedia' by a double-headed arrow. To the right of 'wikipedia' are icons for documents and a globe. On the right side of the diagram is a screenshot of a Wikipedia article for 'Tour Eiffel', showing the title, a map, and a photograph of the tower.

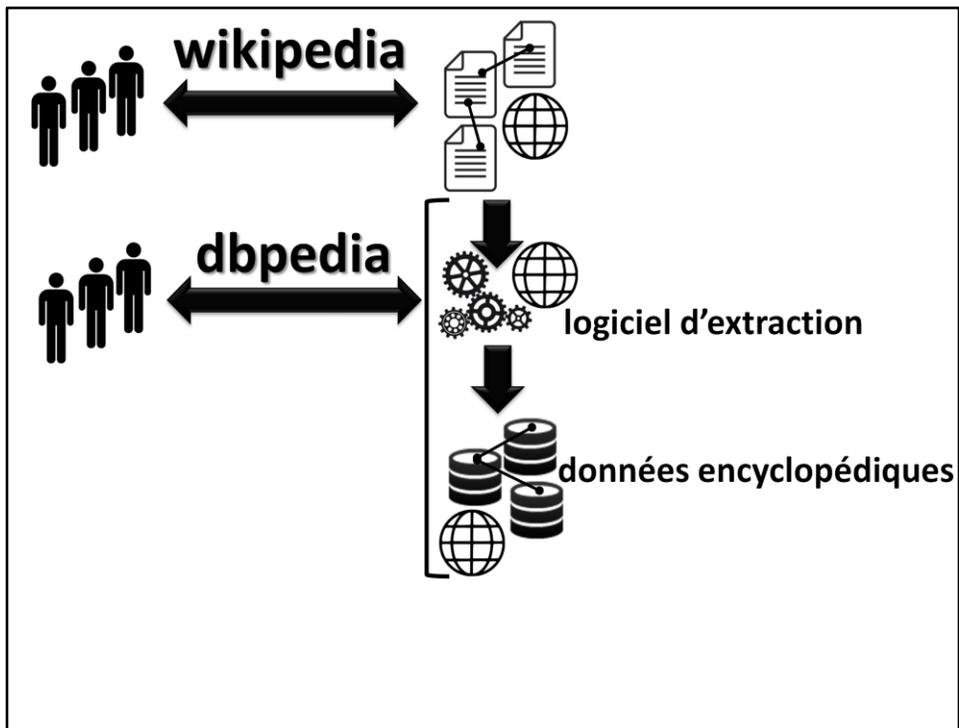
Typiquement une page encyclopédique de wikipedia est dédiée à un sujet par exemple la Tour Eiffel

wikipedia

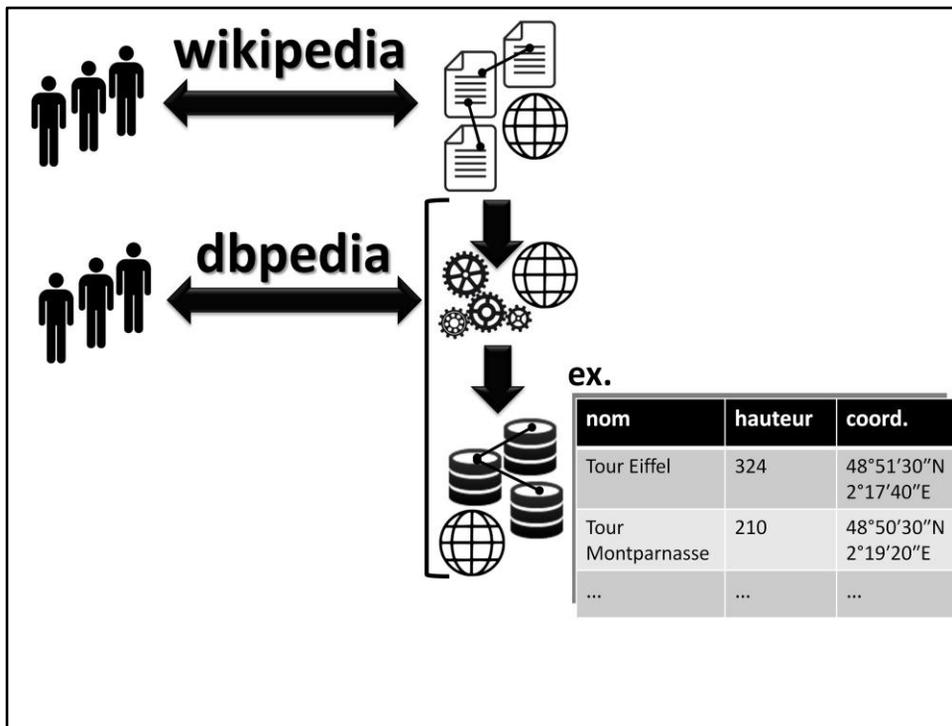
beaucoup de données

Coordonnées	
Coordonnées	Champ de Mars (Paris), 48° 51′ 30″ N 2° 31′ 40″ E﻿ / ﻿48.85833°N 2.52778°E﻿ / 48.85833; 2.52778
Élévation au sol	Carrel à 125 m
Hauteur	
Hauteur au 1 ^{er} étage	37,83 m
Hauteur au 2 ^e étage	276,13 m
États avec antenne	124 m (depuis 2005)
Plats	
Total	19 100 t
Charges réparties	7 200 t
Métal	Foré par les aciers de France (1900)
Nombre d'éléments	13 235 pièces métalliques
Coût	7 500 000 francs
Entrepreneur	Gustave Eiffel & C ^{ie}
Ingénieurs	Maurice Koechlin, Émile Nouguier
Architecte	Stephen Sauvestre
Statut	
Construction	1884 - 1887
Construction	1887 - 1889
Inauguration	2 mars 1889 et 3 jours
Inauguration	31 mars 1889
Propriétaire	Paris
Opérateur	
1881 - 2005	Société nouvelle d'exploitation de la tour Eiffel (SNTE)
2008 - 2010	Société d'exploitation de la tour Eiffel (SETE)
Personnel	≈ 500
Équipement	

Plus ce contenu est riche et plus il a de chances de receler des données, qu'il s'agisse par, exemple, de la hauteur ou des coordonnées GPS du monument.

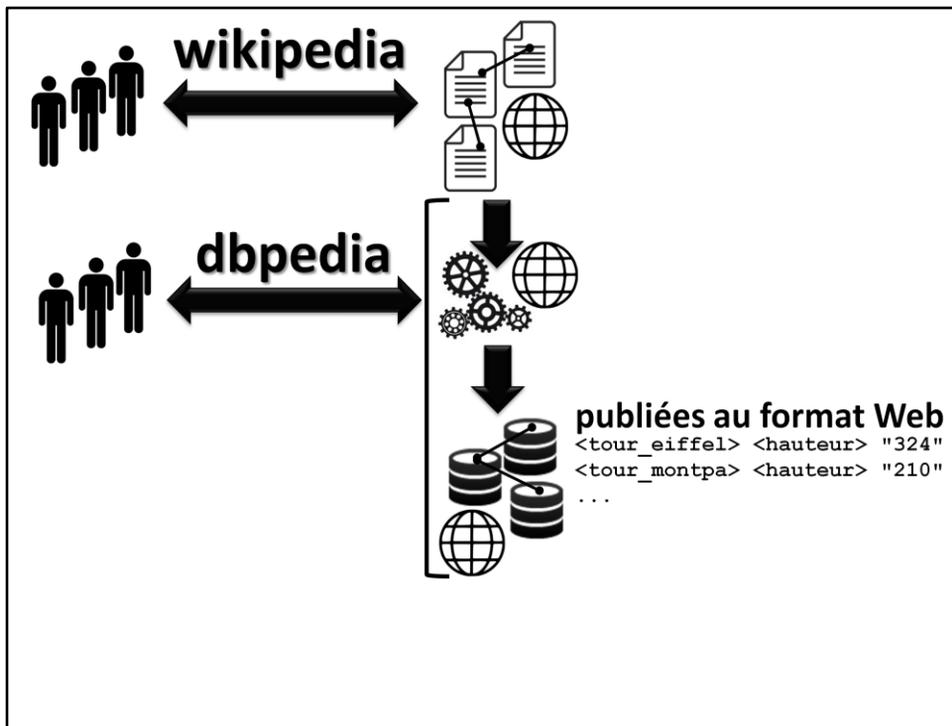


L'idée de Dbpédia fut donc d'extraire ces données en se basant dans un premier temps essentiellement sur le chapitre anglophone de Wikipédia.

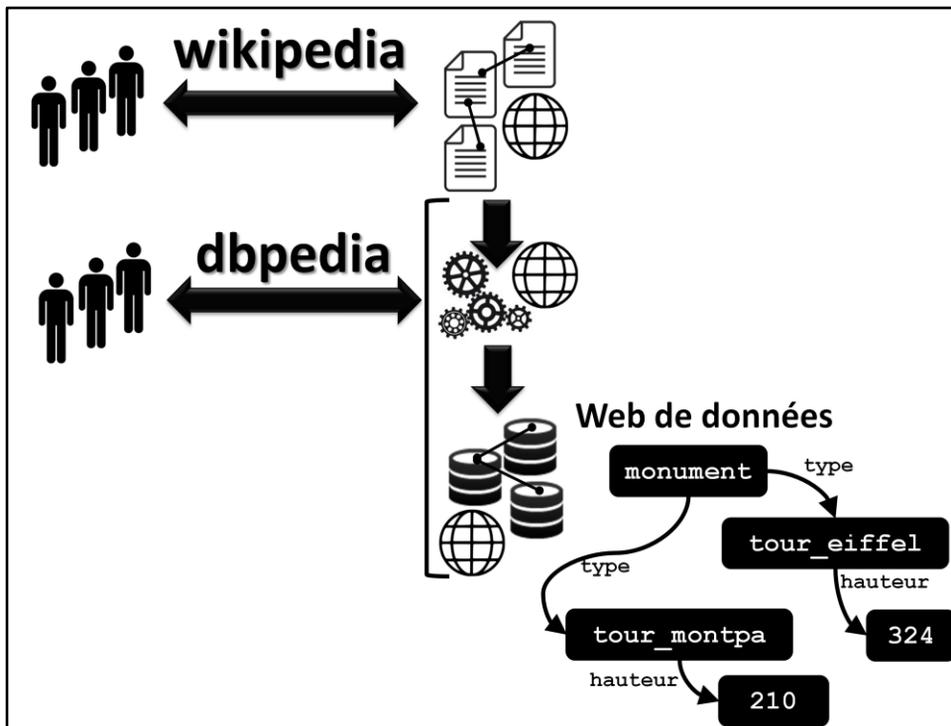


Ces données extraites sont stockées et maintenues de façon structurée permettant des traitements automatiques.

Ainsi, la liste des monuments avec leurs données structurées peut permettre de classer ces derniers par hauteur ou de calculer un itinéraire touristique dans une ville.

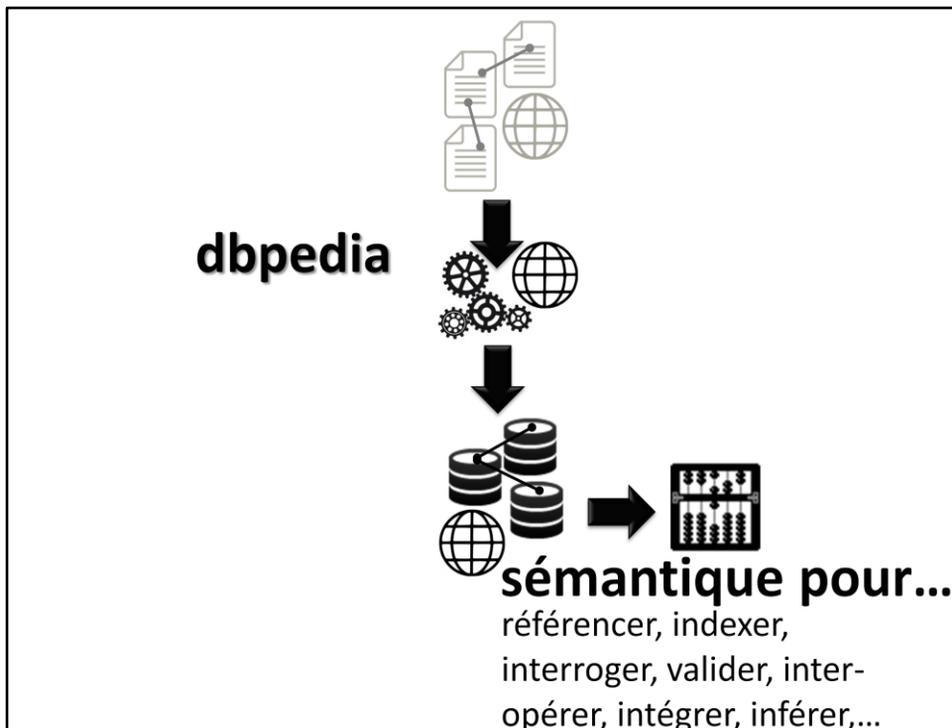


Ces données extraites sont ensuite publiées sur le web cette fois-ci non pas sous la forme de pages lues par des utilisateurs comme dans l'encyclopédie Wikipédia mais dans des formats de données standards et ouverts pour les mettre à disposition des programmes et machines connectées à internet.



Ces données viennent alors augmenter ce que l'on appelle le web de données et le web sémantique c'est-à-dire un web où on ne relie plus uniquement des pages mais bien des entités et leurs données directement exploitables par tous les outils qui ont accès au Web.

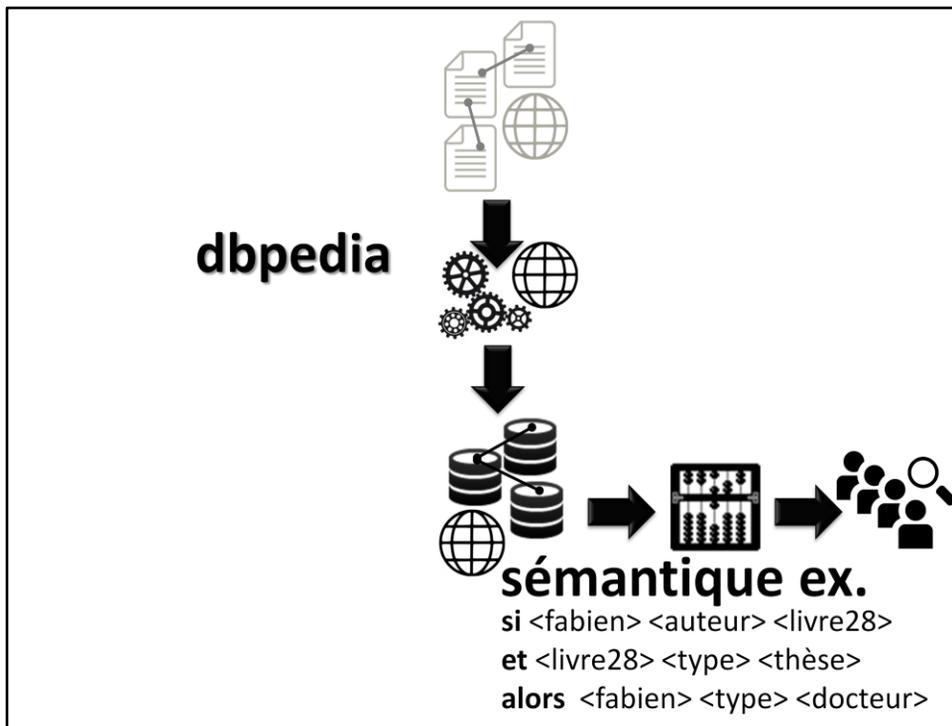
Liées et publiées sur le web, les possibilités de recherche, de parcours, d'accès et d'utilisation de ces données sont démultipliées.



On parle de « web de donnée » et de données liées car les entités, leurs données et les vocabulaires utilisés dans leurs descriptions utilisent des adresses web pour identifier non seulement les sujets, mais aussi les relations et les attributs qu'elles fournissent.

On parle de « web sémantique » par ce que ces vocabulaires, ou schémas, et les relations qu'entretiennent leurs termes sont eux aussi publiés et formalisés selon des langages standard du web.

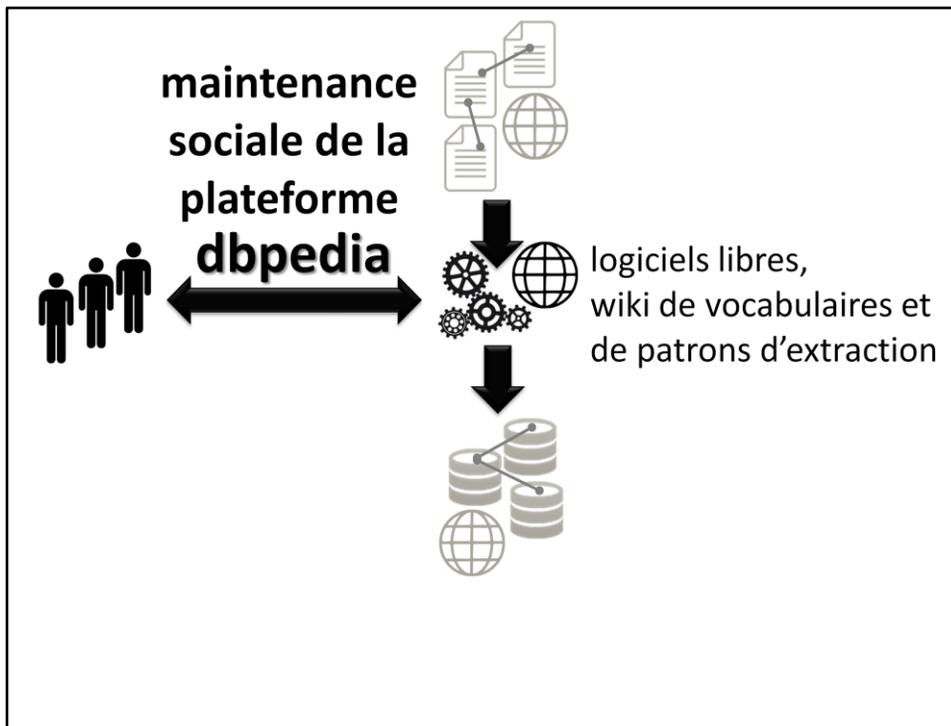
Cette sémantique partagée sur le web permet par exemple de référencer de façon unique un objet ou une caractéristique de cet objet, d'indexer des collections d'objets en partageant les références communes, d'interroger de façon très précise les données et leurs schémas, de valider des descriptions en vérifiant des règles de cohérence, d'assurer l'interopérabilité de systèmes distants en reposant sur un vocabulaire partagé ou des vocabulaires alignés, d'intégrer des jeux de données d'origines différentes, de raisonner sur les données et d'inférer de nouvelles données.



Par exemple, imaginons qu'une donnée indique que je suis auteur du livre 28 et qu'une autre donnée indique que cet ouvrage est une thèse.

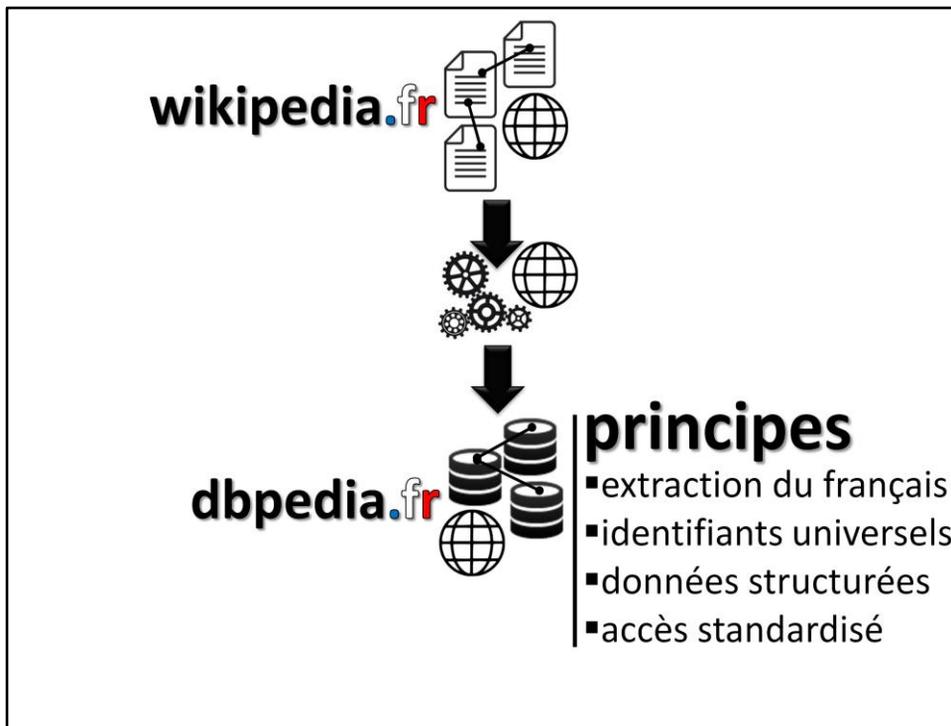
En raisonnant sur les données et leurs schémas, un logiciel pourra en déduire que je suis docteur et par conséquent inclure mon nom dans une requête demandant la liste des docteurs, même si cette donnée n'était pas initialement fournie.

Et il ne s'agit là que d'un exemple d'inférence parmi d'autres qui peuvent être échangées sur le web sémantique et dont les résultats viennent augmenter le web de données et se propager à son échelle.



Insistons enfin sur le fait que la maintenance des logiciels et des plateformes de Dbpédia est elle-même le résultat de l'action collective d'une communauté ouverte. Non seulement la plateforme est libre et les codes sources sont ouverts, mais de plus les collections de règles d'extractions tout comme les vocabulaires utilisés pour structurer les données extraites sont maintenus de façon collaborative là encore dans des wikis.

Ainsi la configuration et la maintenance d'une plateforme Dbpédia font elles aussi l'objet d'une activité sociale à laquelle chacun peut participer par exemple pour ajouter ou modifier des règles d'extractions et ainsi augmenter ou améliorer les données extraites.



Wikipédia est disponible dans de très nombreuses langues.

Le chapitre français lancé en 2001 contient aujourd'hui plus d'un million trois-cent mille articles.

Ce qui en fait le troisième chapitre de ce point de vue.

Ces articles en français regorgent eux aussi de données et notamment de données culturelles, qui n'étaient que très partiellement extraites par la version initiale de DBpédia focalisée sur le chapitre anglophone de wikipédia.

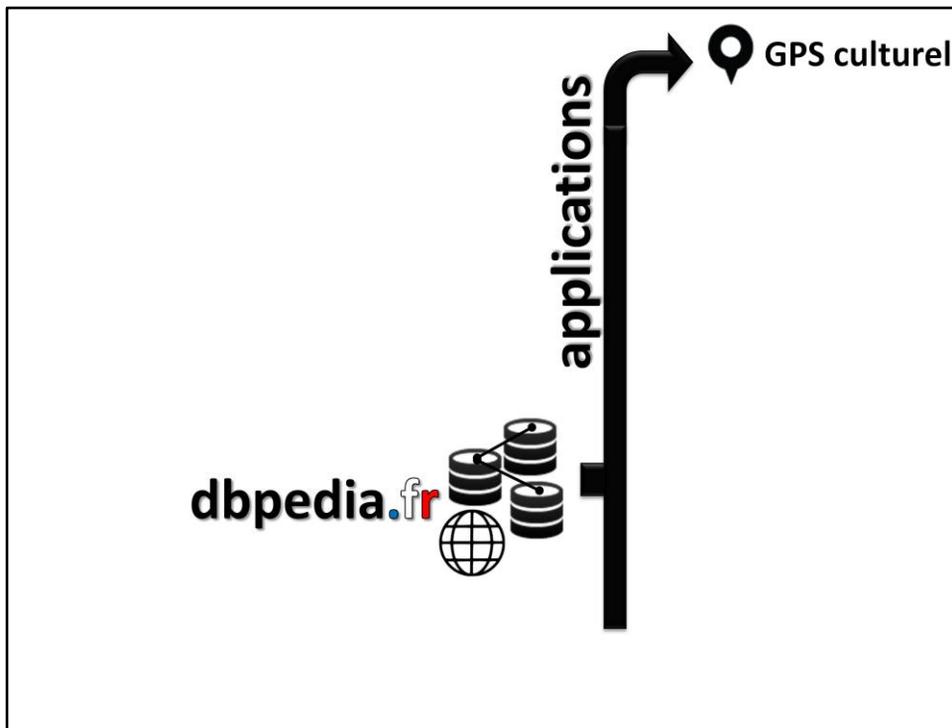
Ce premier projet de la plate-forme Semanticpédia a donc permis de concevoir, développer et publier le chapitre francophone de Dbpédia.

Loin cependant de se borner à rattraper un retard francophone, ce fut l'occasion d'innover en proposant un DBpedia tirant parti des dernières évolutions techniques et des retours d'usages analysés par notre équipe.

Plus précisément nous avons repris et augmenté les extractions de données pour le Français, la publication d'identifiants universels pour tous les sujets couverts, la structuration des données et leur mise à disposition selon les standards du web.

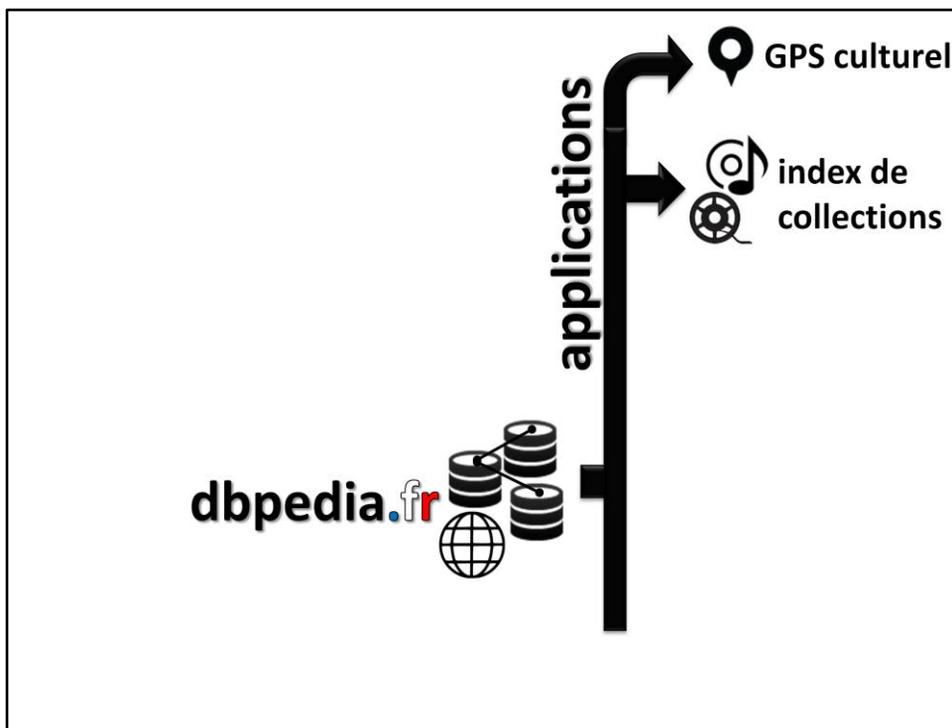
Plus de 130 millions de données sont maintenant extraites et publiées pour le chapitre francophone ce qui nous place aujourd'hui en deuxième position derrière les données extraites du chapitre anglophone.

Notons bien qu'un enjeu de la plate-forme Sémanticpédia est de structurer la contribution française et de la pérenniser dans ses services d'extractions des données, dans ses schémas d'identification des sujets et dans ses moyens d'accès ouverts.

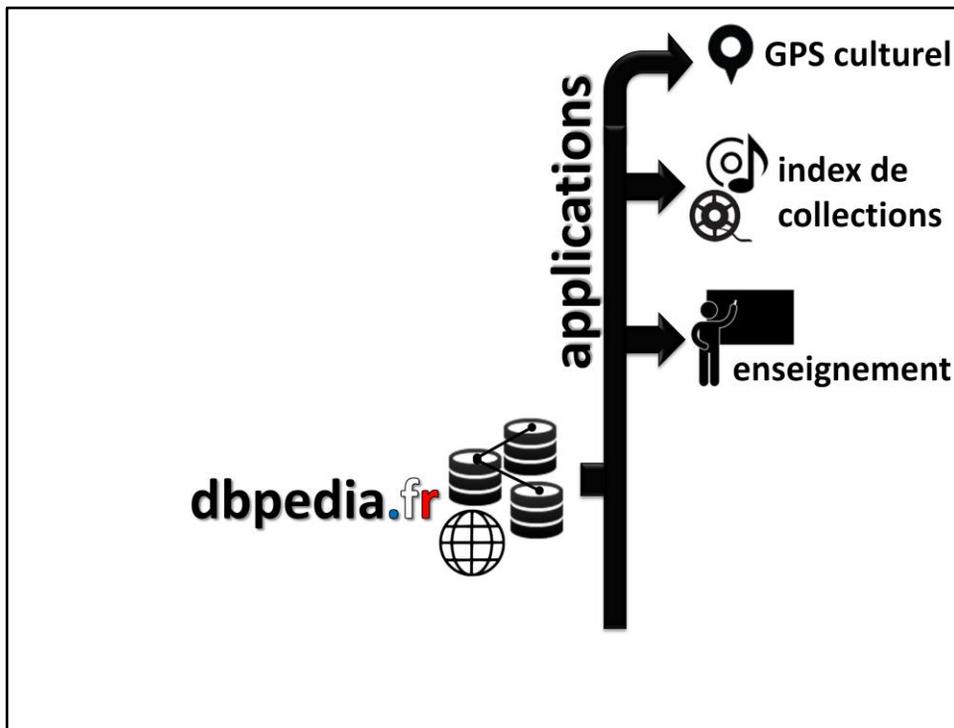


Comme vous le verrez dans les démonstrations qui suivront, ces données une fois publiées permettent de multiples applications. On peut imaginer

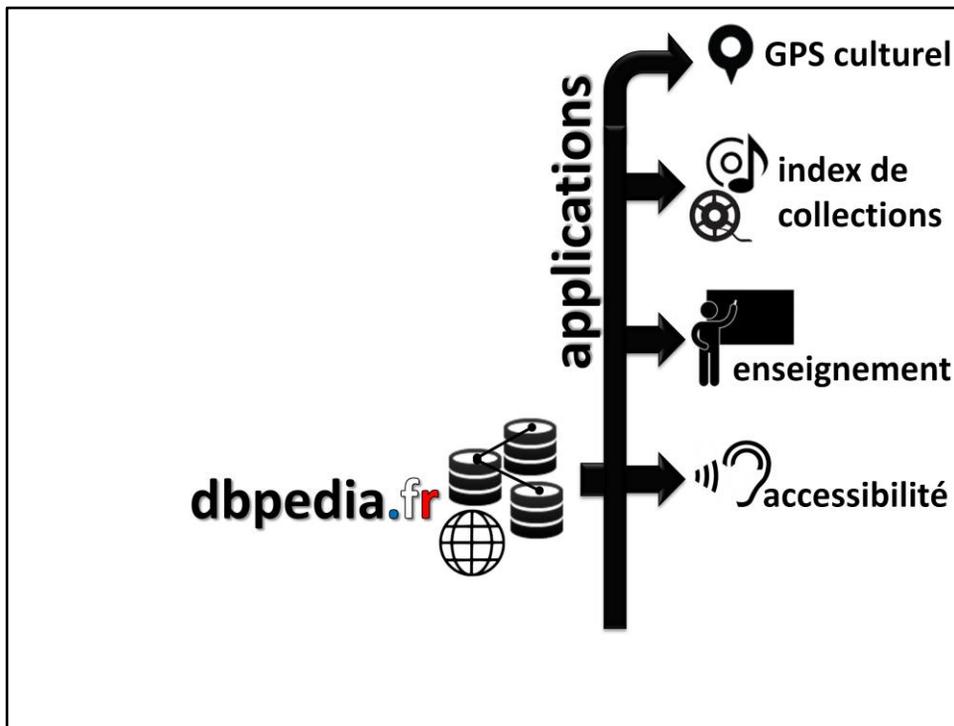
- Des GPS culturels qui nous guident et nous informent en même temps en se renseignant grâce au Web de données sur les lieux que nous traversons



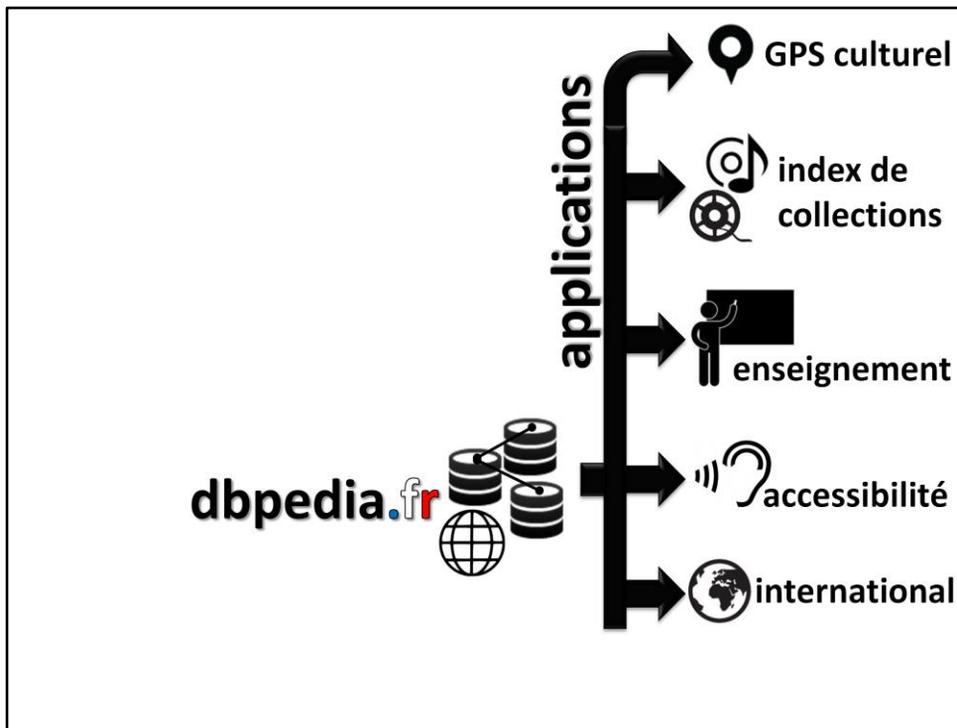
- De nouveaux portails pour organiser, accéder, naviguer et enrichir la description et les liens dans des collections culturelles pouvant avoir des sources très différentes



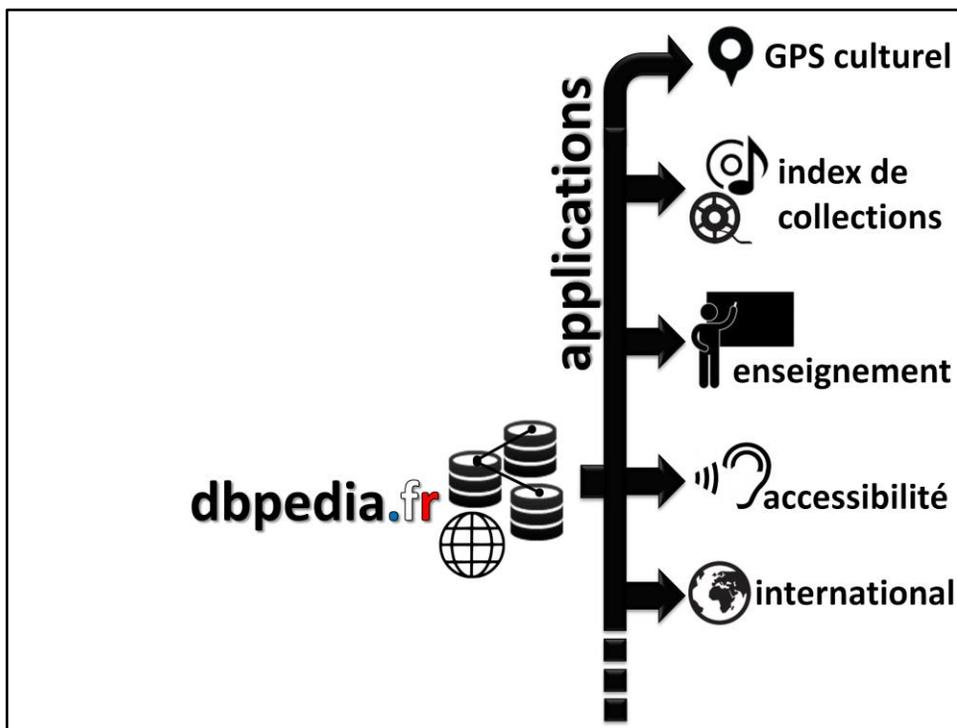
- De nouveaux logiciels interactifs pour l'enseignement par exemple pour fournir de multiples parcours de l'histoire de l'art



- Des interfaces et applications adaptant l'accès à ces données aux handicaps et aux contextes des utilisateurs

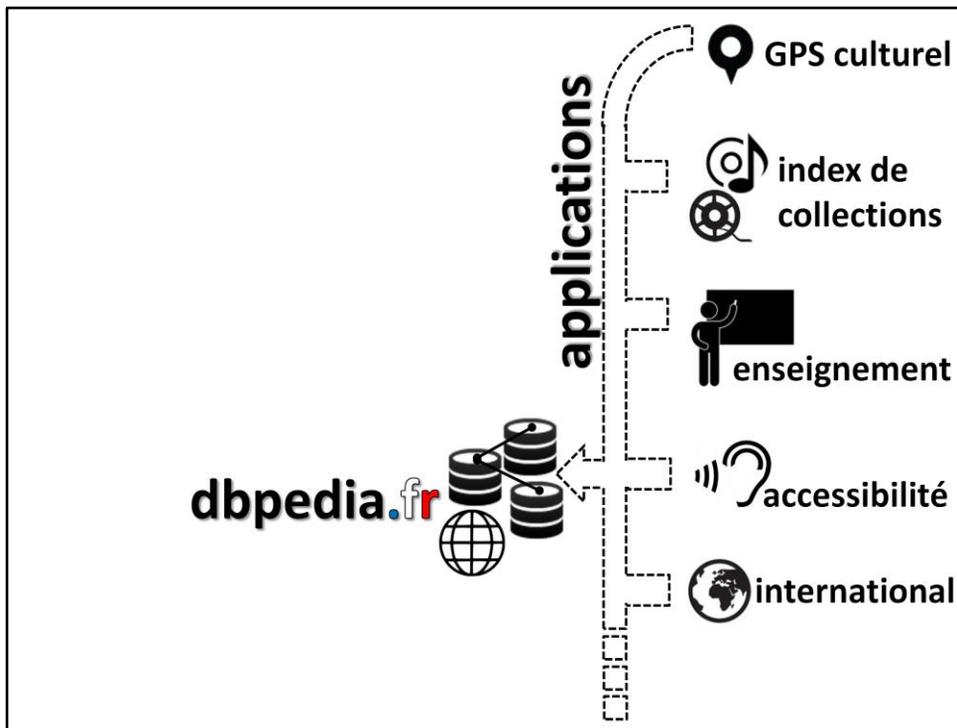


- L'internationalisation de ces données et leur intégration aux données d'autres aires linguistiques et d'autres cultures dans des applications pour des utilisateurs étrangers par exemple dans le tourisme



En fait il est en pratique impossible d'imaginer toutes les applications potentielles de ces données.

Et c'est d'ailleurs là que repose l'enjeu de l'ouverture de ces données qui doivent soutenir une innovation ouverte et des usages inédits, notamment dans le domaine de la culture.



Il reste aussi des défis à relever pour ce projet et pour les projets à venir dans le cadre de la convention Sémanticpédia.

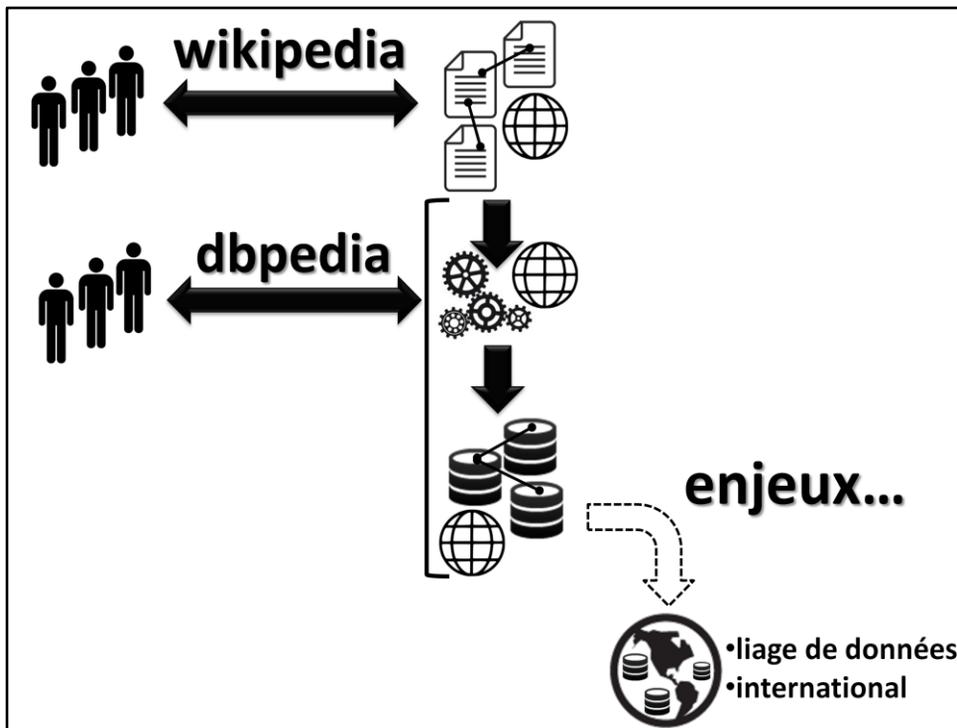
Par exemple, il y a un enjeu à trouver comment permettre à ces applications de participer au cycle de vie complet des données notamment en les enrichissant et en les maintenant à jour.

Comment un GPS culturel pourrait-il aussi permettre de compléter les données manquantes ?

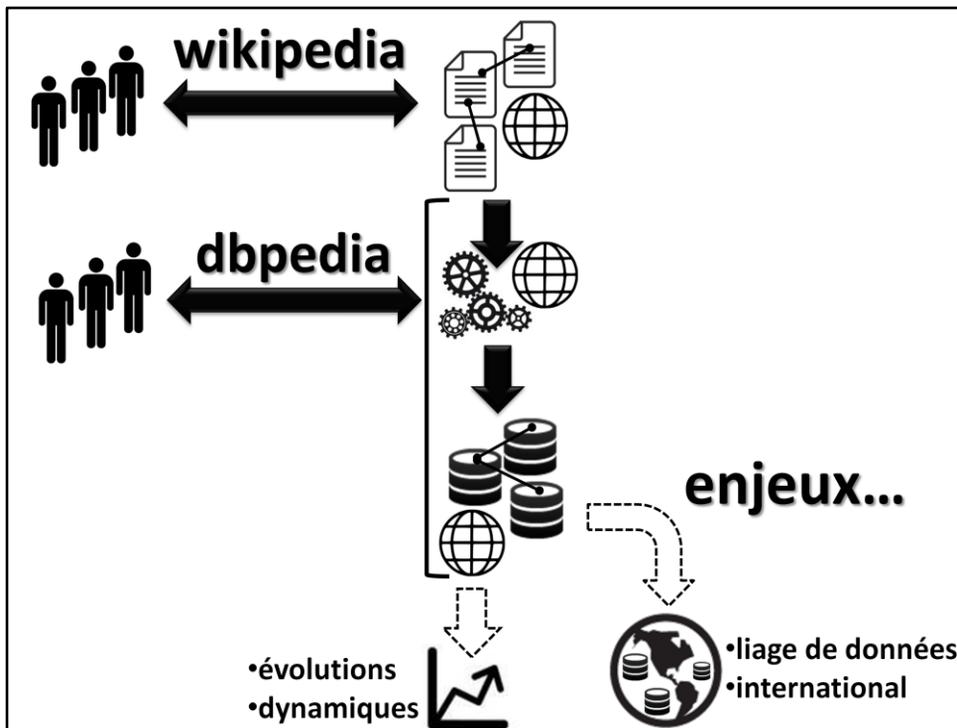
Comment un photographe pourrait-il enrichir la base en versant des photos de son choix ?

Comment un expert d'une institution culturelle pourra-t-il corriger une erreur qu'il a trouvée ?

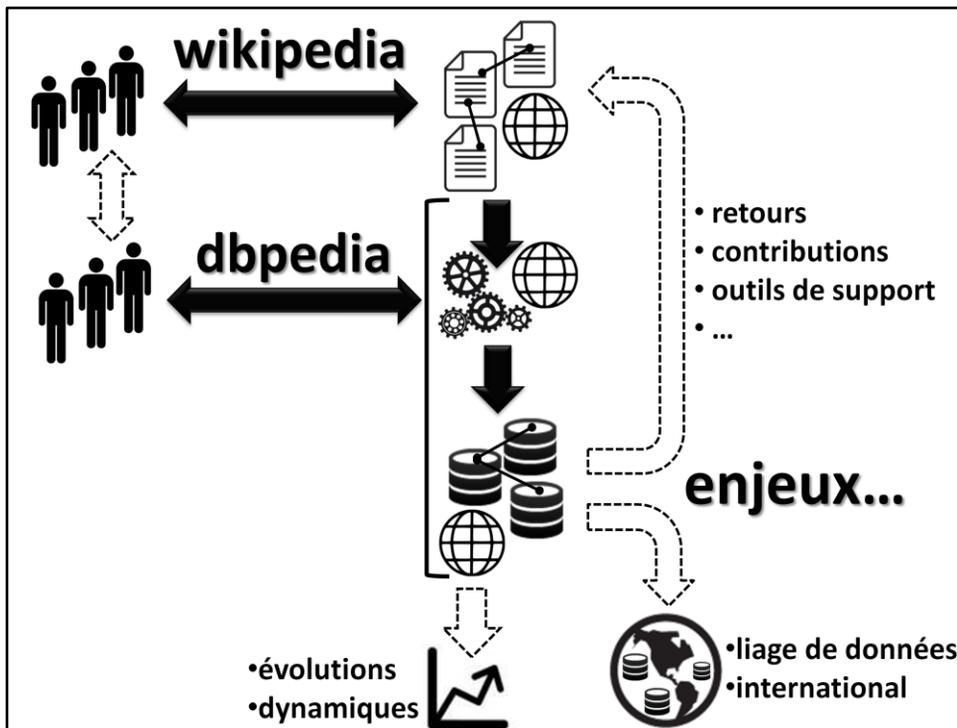
Et de façon générale comment passer de l'utilisateur à l'utilis-acteur des données de la culture.



Un autre enjeu serait aussi de lier ces données au reste des données du web, de croiser les bases en ligne pour qu'elles s'enrichissent mutuellement et d'internationaliser ces données pour qu'à travers ce processus les données francophones de la culture Française soient davantage disponibles dans d'autres langues et accessibles à d'autres cultures et d'autres publics.



Ces données changent aussi en permanence et il y a un enjeu non seulement à soutenir leur cycle de vie mais aussi à analyser ces changements comme des données portant elles-mêmes sur les évolutions, les tendances et toutes les dynamiques qui touchent aux thématiques couvertes dans l'encyclopédie.



Enfin il y a un véritable enjeu de pérennisation à mettre en place un cercle vertueux entre l'encyclopédie Wikipédia et les données extraites et publiées par Dbpedia. Il s'agit notamment de permettre de réintégrer des retours, des contributions et des résultats de calculs sur les données au sein des pages de l'encyclopédie dont elles sont issues.

Et ceci pour enrichir et améliorer son contenu, aider à détecter des manques, des erreurs aussi bien dans les pages que dans les données extraites ou dans les données calculées.

Il s'agit à terme de fournir des nouveaux outils de support à la communauté de Wikipédia qui est, à l'origine, la source qui permet toutes les applications que nous venons d'envisager et dont nous allons voir maintenant quelques exemples.

wikipedia
dbpedia



Fabien Gandon, <http://fabien.info> @fabien_gandon



Merci.