

Délégation générale à la langue française et aux langues de France

SÉMANTICPÉDIA : CONSTRUIRE LE WEB DE DONNÉES CULTUREL

I. Présentation du partenariat et des premières réalisations

Espace de collaboration entre trois univers (la recherche, l'internet contributif et la culture), ce programme vise à croiser les expertises, à partager les retours d'expériences et à mutualiser les efforts de recherche et développement. **Il met le ministère en pointe sur le développement (stratégique pour l'avenir) du « web de données », « web sémantique » ou « web 3.0 ».**

Le service **DBpédia en français**, maillon central de ce projet, rend librement accessible au format du web sémantique l'ensemble des pages de l'encyclopédie Wikipédia en français et dans bien d'autres langues. Cet outil, réalisé par Inria dans le cadre de notre convention, est aujourd'hui hébergé par Wikimedia France et utilisé par plusieurs dizaines de sites internet. Ce service est aujourd'hui complété par un autre outil, « DBpédia-live », développé par une jeune entreprise innovante. Il fournit des résultats extraits en temps réel de l'encyclopédie Wikipédia, ce qui permet d'imaginer des applications fortement liées à l'actualité.

Grâce à un programme de « sémantisation » inscrit au schéma directeur de ce ministère et que nous pilotons avec l'aide de la sous-direction des systèmes d'information (SDSI), nous avons pu mener plusieurs projets d'alignement de données culturelles de sources diverses.

– Un premier projet appelé « **Muséosphère** » a vu le jour cette année et permet, à partir des données de Wikipédia et de la base Muséofile de ce ministère, de trouver facilement un musée dans le monde selon plusieurs critères.

– Un second projet appelé « **JocondeLab** » permet d'accéder à près de 300 000 notices décrivant des œuvres des musées en de France en 14 langues. Ce projet, qui a été mené grâce au travail considérable de liage effectué par le service des musées de France et avec l'aide de notre partenaire de recherche, l'institut de recherche et d'innovation du Centre Pompidou, montre l'utilité du « web de données » pour valoriser la culture dans un contexte multilingue, et ce dans un contexte budgétaire restreint (environ 50 000 euros) et des délais très courts (huit mois).

Ce projet, qui associe DBpédia à la base Joconde a été **auditionné en mars 2014** par la Mission commune d'information sur l'accès aux documents administratifs et aux données publiques du Sénat. Le projet illustre le rapport intitulé « Refonder le droit à l'information publique à l'heure du numérique : un enjeu citoyen, une opportunité stratégique » dans lequel il est précisé que « JocondeLab constitue **un exemple de pilotage cohérent et efficace** ». Ce projet a par ailleurs été récompensé en mars dernier par le **prix « Data Access »** des « *Data Intelligence Awards* ».

II. Des retombées en dehors du ministère

Le ministère de la Culture n'est pas l'utilisateur exclusif de cet outil DBpédia : dans la sphère publique, **la Cité de la musique** et **la Cité de l'architecture et du patrimoine** mènent tous deux des projets réutilisant le jeu de données DBpédia. Du côté des régions, la fédération des langues Occitanes « Lo Congrès » souhaite s'appuyer sur le Web de données et ses outils pour développer la présence numérique de l'occitan. Dans la sphère civile, de nombreux projets ont vu et voient le jour. Un bon exemple est le robot SAMI, présenté à **Futur en Seine** par la société *Vocal Apps* et qui répond aux questions orales des visiteurs grâce à DBpédia.

Dans l'objectif d'encourager ces démarches s'est tenue le 29 avril 2014 la remise des prix du **concours Sémanticpédia** organisé par Wikimedia France. Trois lauréats ont été récompensés à partir de la vingtaine de dossiers, pour la plupart très innovants, transmis au jury.

III. Perspectives

Nous souhaitons continuer à explorer les potentialités offertes par ces nouveaux outils, d'une part, en les améliorant : nous soutenons ainsi les projets visant à **améliorer la qualité et la quantité d'informations** présentes dans l'outil DBpédia, comme, par exemple, à partir de l'analyse syntaxique des articles de l'encyclopédie Wikipédia (projet Tagmatica) ;

D'autre part, nous prévoyons de mener (après Jocondelab et Muséosphère) **une troisième expérimentation de liage de données** l'année prochaine à partir d'une banque de données sonores sur les langues de France que la DGLFLF constitue depuis une dizaine d'années et intitulée « Corpus de la parole », en faisant appel à l'expertise de la BNF et du CNRS en plus de celle déjà acquise ;

S'agissant d'ailleurs de la pluralité linguistique interne, nous envisageons – en nous appuyant sur la circulaire de la ministre relative aux langues régionales du 31 mars et à l'initiative du Cabinet – de proposer un projet de valorisation des langues régionales pour la signalétique en ligne et hors ligne des **monuments nationaux** à partir du web de données et plus précisément de la ressource DBpédia.

Enfin, nous participons activement à la préparation d'un cadre plus ambitieux pour le partenariat Sémanticpédia, avec l'appui d'un atelier piloté par le département des programmes numériques et par nos partenaires de recherche, Inria et Wikimedia France. L'objectif de ce projet est d'associer de manière plus étroite **les établissements publics culturels** et **les directions régionales des affaires culturelles** au partenariat, de faciliter l'alignement de différents jeux de données en plus de l'encyclopédie Wikipédia, et de rédiger une « **convention Sémanticpédia version 2** » permettant de mettre en œuvre et de piloter plus facilement des projets autour du Web de données.