MINISTÈRE
DE LA CULTURE
Liberté
Égalité
Fraternité

Direction générale des patrimoines
Service des musées de France

JOURNEE PROFESSIONNELLE « LES MUSEES A L'HEURE DU NUMERIQUE : TRAVAILLER EN RESEAU, REUTILISER ET CONTRIBUER » PARIS, 07/06/2013

Mise en ligne : juin 2013

Enjeux culturels et linguistiques autour des données liées : Sémanticpédia et le programme Sémantisation

Thibault Grouas, délégation générale à la langue française et aux langues de France

Les technologies numériques concourent grandement aux objectifs de la politique de la langue du ministère de la Culture : préserver la langue française pour permettre de répondre aux besoins d'expression et de communication des citoyens et des institutions d'une part, favoriser le passage entre le français et les autres langues afin de favoriser le multilinguisme, d'autre part.

La délégation générale à la langue française et aux langues de France, en raison de son positionnement interministériel et de sa vision d'ensemble des enjeux linguistiques propres au développement des technologies numériques, est en mesure de jouer un rôle de catalyseur. Ce positionnement doit lui permettre d'orienter son action autour de trois priorités : prendre en compte la dimension linguistique des technologies numériques, contribuer à mettre celles-ci au service de la politique de la langue, veiller à la présence de la langue française sur la Toile et aux moyens qui l'encouragent.

Pour mener à bien cette politique en faveur du numérique, le ministère accompagne différents types de projets en faveur du français et de la diversité linguistique :

1. Sur les technologies de la langue

Les technologies de la langue constituent pour les pouvoirs publics une question d'intérêt général : elles contribuent à l'amélioration de la vie quotidienne de nos concitoyens, au développement de notre économie et au renforcement de nos échanges.
Ces technologies permettent de multiples applications telles que :
- la traduction automatique et l'aide à la traduction ;
- l'aide à la rédaction (correcteurs orthographiques et grammaticaux) ;
- la reconnaissance vocale et les commandes vocales ;

- la synthèse vocale ;
- l'indexation automatique de documents ;
- l'automatisation des processus de constitution de méta-données (résumés, mots-clefs, catégories...)

2. Sur l'internet et les médias collaboratifs

La présence du français sur l'internet passe notamment par une participation active de nos concitoyens aux différents projets collaboratifs. Il est donc particulièrement important pour le ministère de s'assurer du dynamisme de l'effort collaboratif en langue française.
à ce titre, deux aspects apparaissent particulièrement stratégiques :
- l'enrichissement des encyclopédies ou dictionnaires collaboratifs en ligne ;
- la traduction en français des logiciels, sites internet, réseaux sociaux et outils proposés aux utilisateurs.

3. Sur le Web des données et le web sémantique

C'est un enjeu culturel majeur de développer en français des outils de diffusion culturelle qui confortent le rôle historique de la langue française comme langue internationale de diffusion des savoirs.
Grâce aux nouvelles technologies du web sémantique, l'internet 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...) ouvrant la voie à des usages radicalement nouveaux des données numériques.

Le ministère de la Culture, souhaitant que la langue française soit au cœur du web des données, a lancé en novembre 2012 un partenariat stratégique dénommé Sémanticpédia avec Inria et l'association Wikimédia France, afin de développer l'écosystème naissant autour des données culturelles liées.

Le premier projet développé dans ce cadre a été l'extraction semi-automatisée des données structurées de Wikipédia en français et des langues qui y sont liées. Ce projet, utilisable gratuitement et librement sur internet et dénommé DBPédia en français, s'appuie sur le projet international DBpedia.org. Il a été développé par l'Inria et permet d'ajouter de manière simple des données en provenance de Wikipédia sur des sites ou services proposés par des acteurs publics ou privés.

Le projet HDA Lab, qui reprend les données du site portail Histoire des Arts en les croisant avec celles issues de Wikipédia via DBpedia, est un des premiers projets de ce type mené au ministère de la Culture.

Le programme « Sémantisation », inscrit au schéma directeur des systèmes d'information 2013-2015 du ministère et au programme ministériel de modernisation du ministère (PMMS), vise à tirer les bénéfices de cette nouvelle méthode de travail et de publication sur internet, qui permet l'enrichissement des contenus du ministère avec d'autres contenus tiers, tels ceux présents sur la base DBpédia en français.

Il facilite par ailleurs le développement d'interfaces de navigation multilingues sans nécessiter de traduction, ce qui ouvre la voie à une diffusion plus large des contenus culturels français dans le monde.

Dans le cadre du programme « Sémantisation », plusieurs projets sont envisagés :
- Sémantisation de la base Joconde, gérée par le service des musées de France (SMF) et comptant plus de 300 000 notices illustrées (2013). Ce projet proposera des innovations de l'interface permettant la navigation et l'accès à l'image dans plusieurs langues, et pourra préfigurer l'interface des prochains outils utilisés au ministère pour enrichir et alimenter les bases de données culturelles : MISTRAL et GINCO notamment.

- Sémantisation du Wiktionnaire, en lien avec la communauté de linguistes et chercheurs en traitement automatisé de la langue (2014-2015)

- Sémantisation du Corpus de la parole, base de données opérée par la DGLFLF et contenant plus de 1000 extraits sonores documentés en diverses langues de France (2014-2015).

En termes de réutilisations innovantes, il est par ailleurs proposé de travailler sur le développement d'applications s'appuyant sur DBpédia en français et/ou sur les données sémantisées de Joconde.

C'est dans ce contexte qu'est développée l'application pour terminaux mobiles Muséophile, dont l'objectif principal est de démontrer la faculté de développer rapidement (2 mois) et à moindre coût une application multilingue (8 langues proposées) donnant accès à un corpus de données large et multilingue à partir d'une interface utilisateur simple et ergonomique.