Ministère Culture Communication

Délégation générale à la langue française et aux langues de France

Enjeux culturels et linguistiques autour des données liées :

Le projet Sémanticpédia

Thibault Grouas
18 avril 2013

1. L'apport des technologies à la politique de la langue

- La Délégation générale à la langue française et aux langues de France (DGLFLF) élabore **la politique linguistique du gouvernement**
- Elle s'intéresse à toutes les langues parlées en France : le français, les langues régionales, les langues étrangères. On dénombre **75 langues parlées en France**
- Elle promeut activement **le multilinguisme**, notamment en Europe (Le multilinguisme, première langue européenne?)
- Elle dispose d'une **mission dédiée aux technologies numériques**
- Le numérique apparaît comme l'enjeu majeur pour les langues, comme un levier indispensable pour la conduite d'une politique.

2. Diversité des projets en terme de numérique

- Action sur les **technologies de la langue** (synthèse vocale, reconnaissance vocale, traduction automatique, correction d'othographe, sous-titrage...)
- Utiliser l'**internet collaboratif** pour favoriser l'enrichissement du français : Wiktionnaire, WikiLF
- Promotion du multilinguisme sur internet
- Encourager la **diversité linguistique** et promotion des cultures locales via le web 2.0 et les réseaux sociaux : Projet de **wikilivre des Outre-mer**
- Soutien à l'action de l'association Wikimédia pour la création de communautés de contributeurs en langues locales
- Diffusion la plus large possible des langues et cultures de France sur les réseaux : **le web sémantique** apparaît comme un enjeu majeur

3. Le web sémantique en soutien à l'influence culturelle française

- Les technologies sémantiques représentent **une évolution considérable** des méthodes d'accès et de classement de l'information
- La présence de la culture et de la langue sont un enjeu significatif pour la France
- Le web sémantique sera peut-être demain la clé de voûte de **l'accès à l'information** sur l'internet. Les outils de recherche pourraient bénéficier de nouvelles fonctionnalités
- Il est indispensable que **la culture et les langues de France** soient présentes sur la toile sémantique, risque majeur si ce n'est pas le cas
- Premier projet soutenu par la DGLFLF : projet SémanticPedia avec **Inria** et l'association **Wikimedia France**.

4. Le projet Sémanticpédia ?

Une **collaboration inédite** entre un acteur culturel public majeur, une expertise technologique, et une communauté de talents

4. Le projet Sémanticpédia ?

- Un **espace de collaboration** entre trois univers : la recherche, l'internet contributif et la culture

- Forte **complémentarité** des partenaires :

  - Wikimédia France : premier diffuseur de données culturelles en France sur internet (20 millions de visiteurs)

  - Ministère de la Culture : expertise unique sur les patrimoines et la création culturelle

  - INRIA : expertise technique sur le web et le web de données

- Objectifs : **croiser les expertises**, partager les retours d'expériences, mutualiser les efforts de recherche et développement.
- Créer des interconnections avec l'écosystème du Web sémantique ouvrant **des usages aujourd'hui insoupçonnés**.

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Ministère
Culture
Communication

5. Pourquoi Sémanticpédia ?

- Parce que Wikipédia, qui compte **plus de 1,3 millions d'articles** en français, est une ressource culturelle unique en son genre.
- Près de 40% des articles sur l'encyclopédie concerne directement **des contenus culturels**
- Grande diversité des contenus : tous les « corps de métier » de la culture sont représentés
- Une partie importante d'informations sont déjà « **structurées** » et potentiellement exploitables : infobox, catégories, portails...
- Le projet initial **DBpedia.org** ne couvrait pas le français : risque majeur pour tous les contenus culturels qui ne sont disponibles qu'en français.

  → Nécessité de conduire un projet de sémantisation centré sur Wikipédia **en langue française**

6. Le projet DBpédia en français

- **DBpedia.org** 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
- L'INRIA, à travers le projet **DBpédia en français** est le contact de référence pour DBpedia.org pour la France
- **DBpédia en français** permet notamment de couvrir beaucoup plus largement la culture francophone et notamment la création contemporaine, qui n'est pas encore décrite en anglais sur Wikipédia.
- **Plusieurs innovations** par rapport au projet initial : identifiants pérennes (URI), nouveaux extracteurs...
- Ce projet positionne la langue française au cœur du Web de données émergent

7. Web sémantique et multilinguisme

- Autre intérêt des technologies sémantiques : lorsque les corpus sont multilingues (Wikipédia), elles représentent **une chance pour le multilinguisme**
  - Possibilité de créer des interfaces de navigation nativement multilingues rapidement
  - Pour un éditeur, la charge liée à la traduction baisse considérablement
  - Possibilité d'utiliser des méthodes de classement (mots clés, rubriques...) multilingues

    → exemple avec Europeana : 11000 résultats avec « Baroque », 1200 avec « Barocco », 1600 avec « Barock ».

  → L'accès à la culture française est plus difficile pour ceux ne maîtrisant pas la langue française.
- Les technologies multilingues permettent donc d'**exporter la culture et la langue française**, de mieux la rendre visible aux locuteurs ne maîtrisant pas le français.

8. Web sémantique et langues de France

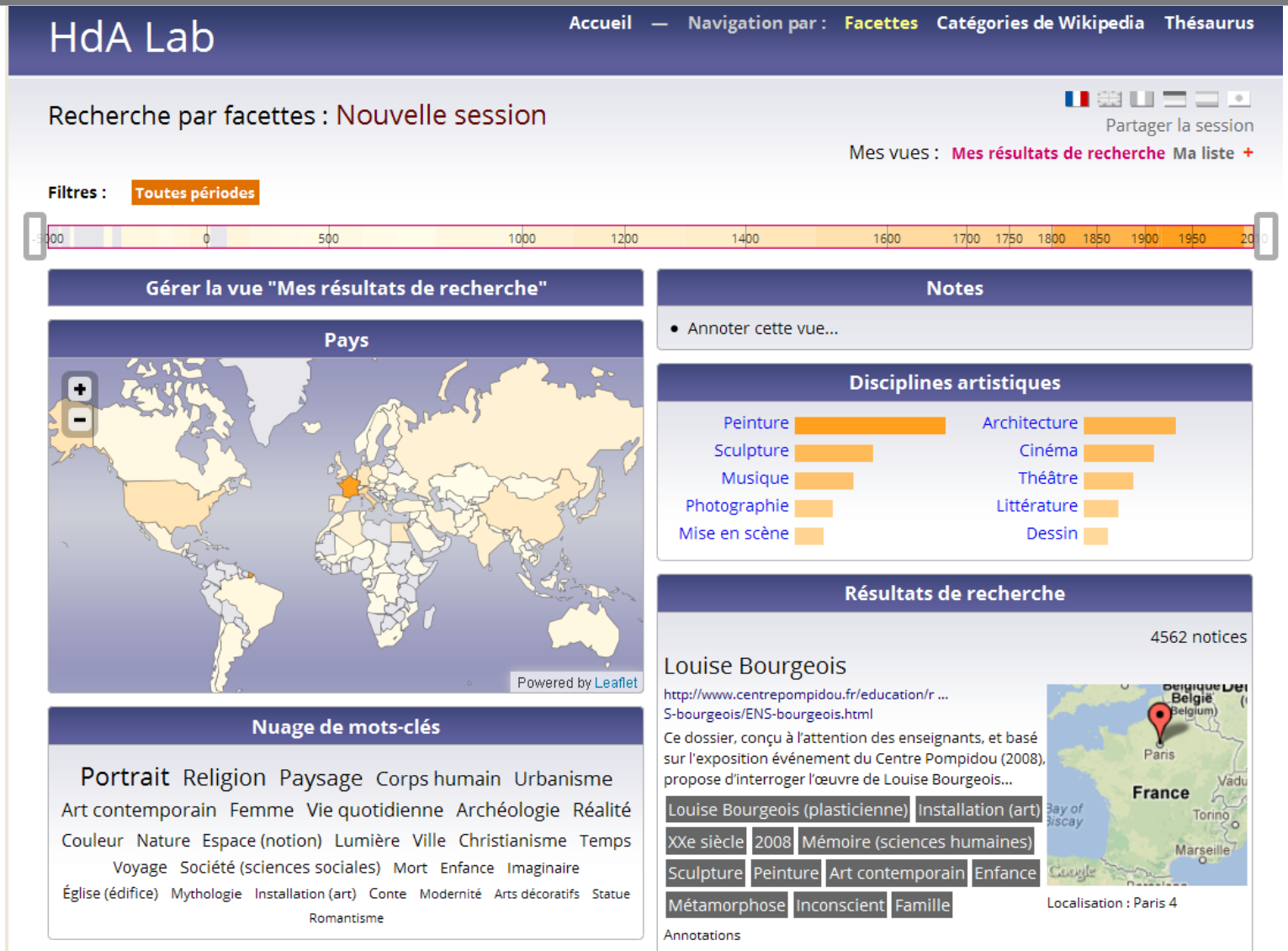
- La France parle **75 langues différentes**, certaines sont déjà bien présentes sur internet (basque, breton, catalan, occitan...) d'autres beaucoup moins visibles (langues de l'outre mer...)
- Intérêt croissant pour ces langues, dont certaines sont menacées.
- **Intérêt politique** : engagement du président de la République de ratifier la charte européenne des langues régionales
- **Wikipédia** et le **Wiktionnaire** constituent, pour beaucoup de ces langues, la seule présence sur la toile
- Les technologies sémantiques permettent de donner accès aux cultures en langues de France plus facilement (par exemple, article WP sur le Maraké, culture arawak en Guyane)
- Elles permettent aussi de **faciliter l'accès à la culture** pour les publics maîtrisant mal le français : interfaces de navigation et méthodes de classement en langues régionales

9. Web sémantique et dictionnaires

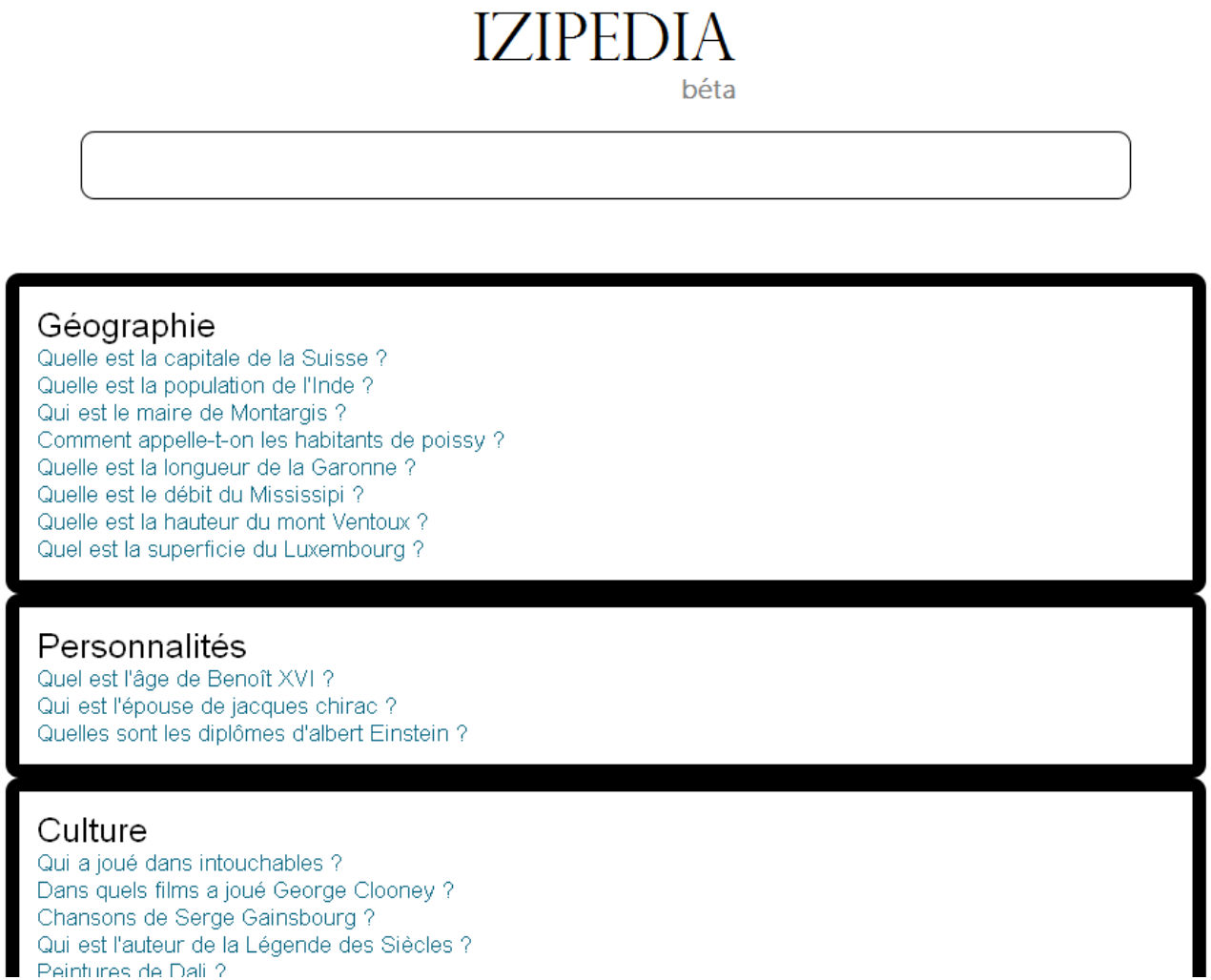
- Le **Wiktionnaire en français** est le 2e plus important après l'anglais, avec près de 2 millions de termes
- Le Wiktionnaire est un contenu déjà très structuré et complet (terme, définition, grammaire, équivalents étrangers, version vocale, phonétique...)
- Quelques intérêts d'une sémantisation du Wiktionnaire :
  - Développement facilité de systèmes d'organisation et de classement basés sur le français
  - Meilleur suivi de l'utilisation, de l'apparition et de la disparition de termes (Wiktionnaire comporte de nombreux termes nouveaux et désuets)
  - Développement d'outils de traitement automatiques interlingues : traduction automatique, reconnaissance vocale, synthèse vocale
  - Outillage propre à assurer une meilleure proximité de l'État avec les administrés qui maîtrisent mal le français (intérêt majeur pour la justice, la santé, les services d'urgence...cf. Actes des États généraux du multilinguisme)

→ Dans le cadre de **Sémanticpédia**, nous évaluons actuellement les possibilités offertes par le Wiktionnaire au format RDF sur interface SPARQL.

10. DBpédia en français : réutilisations



10. DBpédia en français : réutilisations



10. Les projets de sémantisation au ministère de la Culture

- Le nouveau **schéma directeur 2013-2015** des systèmes d'information du ministère de la Culture consacre une partie de son budget à l'**innovation**
- Le **programme « Sémantisation »** financé dans ce cadre a pour but, au cours des 3 prochaines années, de mener des projets d'expérimentation au MCC autour des données liées.
- Une première expérimentation aura lieu cette année autour de **la base JOCONDE**, qui contient notamment **300 000 notices illustrées**. L'innovation se portera sur les modes d'accès à l'**image**.
- Cette expérimentation s'appuiera notamment sur les résultats de la preuve de concept « **HDA-Lab** ».
- L'année prochaine, une autre expérimentation sera menée sur un **corpus d'archives sonores** en langues de France. L'innovation portera sur la navigation dans un corpus sonore.
- Ces expérimentations permettront au MCC d'**améliorer ses outils de travail internes** en fonction des résultats expérimentaux obtenus.

Des questions ?

Thibault Grouas
thibault.grouas@culture.gouv.fr
@tgrouas