Ministère Culture Communication

Délégation générale à la langue française et aux langues de France

# Enjeux culturels et linguistiques autour des données liées :

# Sémanticpédia et le programme « Sémantisation »

Thibault Grouas
7 juin 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** et **le web de données** apparaissent comme un enjeu majeur

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

- Les technologies sémantiques et le web de données représentent **une évolution considérable** des méthodes d'accès et de classement de l'information
- La présence de la culture et de la langue dans le web de données est 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

# 4. 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.

# 5. 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 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

# 6. Sémanticpédia ?

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

# 6. 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 (près de 25 millions de visiteurs par mois)

  - 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**.

# 7. Pourquoi Sémanticpédia ?

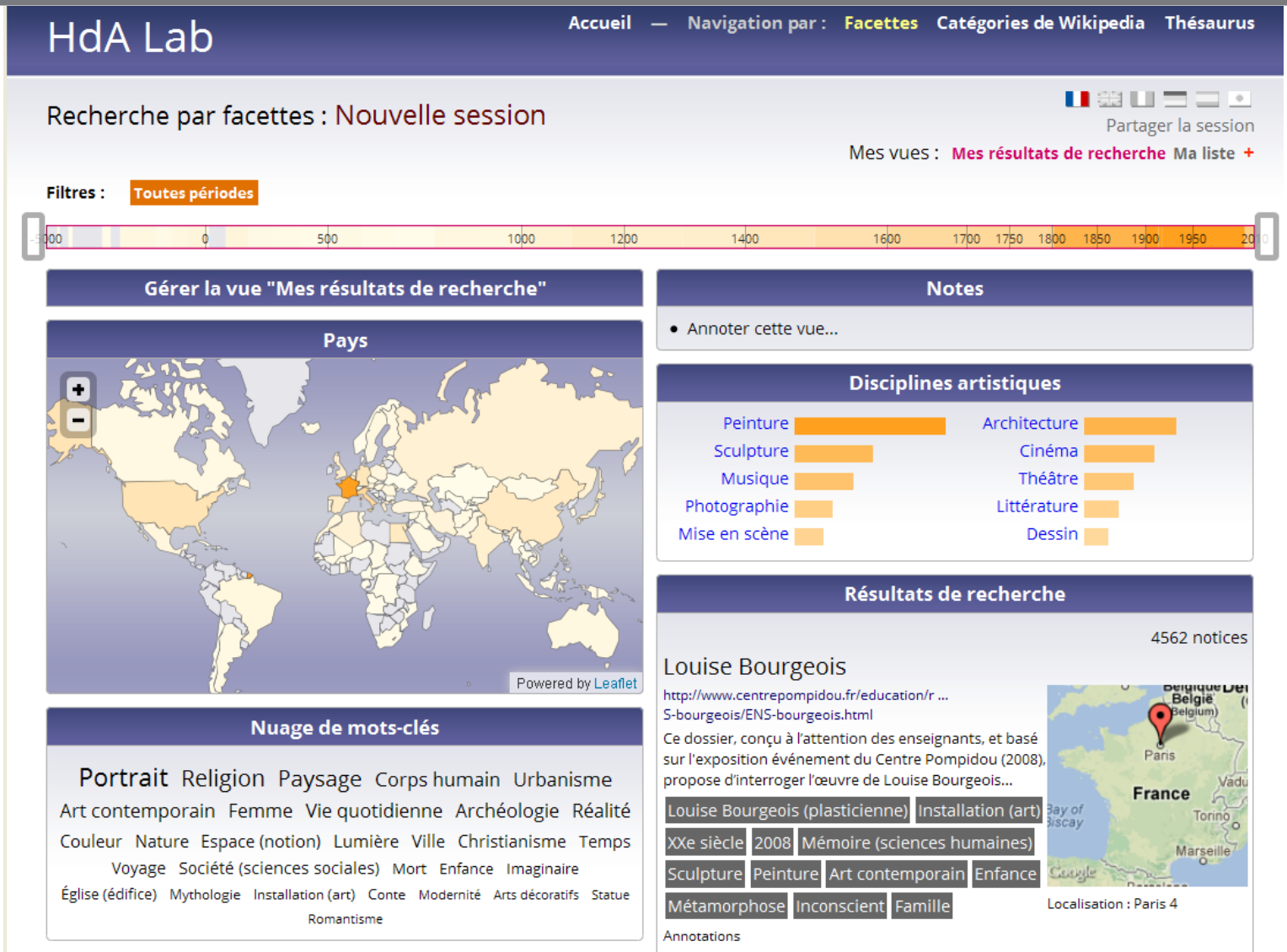
- Pour pousser plus loin l'expérimentation **Histoire des Arts « HDA Lab »**, menée par le DPN, riche en enseignements.
- Parce que Wikipédia, qui compte **plus de 1,3 millions d'articles** en français, est une ressource  culturelle unique en son genre.
  - Près de 40% des articles sur l'encyclopédie concerne directement **des contenus culturels**
  - Grande diversité des contenus : tous les formes d'expression culturelle 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 pour tous les contenus  qui ne sont disponibles qu'en français.

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

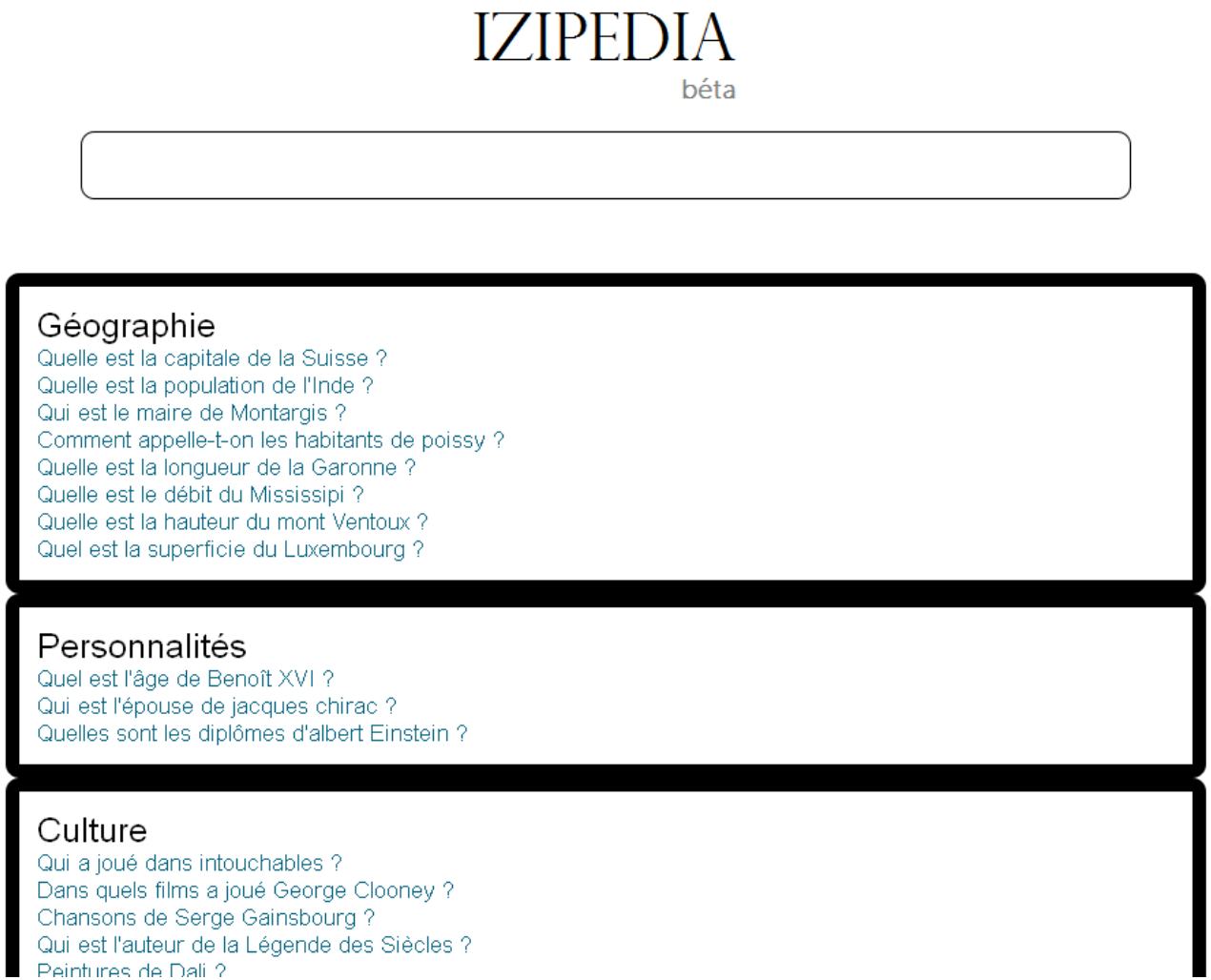
# 8. Le projet DBpédia en français

- **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

# 9. DBpédia en français : réutilisations



# 9. DBpédia en français : réutilisations



# 10. Le programme Sémantisation

- 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 et lancé en février 2013 a pour but, sur les 3 prochaines années, de mener des projets d'expérimentation au MCC autour des données liées.
- Le programme constitue un des engagements du **programme ministériel de modernisation et de simplification** (PMMS).
- Les projets qui en sont issus sont développés rapidement et à budget limité. Premier exemple : **Muséophile**.
    - Une application optimisée pour le mobile
    - présentant un accès multilingue aux musées du monde
    - selon différents choix personnels (artiste ou mouvement préféré)
    - Durée totale du projet : deux mois avec 8 langues disponibles
    - L'application sera mise en ligne début juillet.

# 11. Le projet JocondeLab

- Une  expérimentation a lieu cette année autour de **la base JOCONDE**, qui contient notamment **300 000 notices illustrées**.
- Cette expérimentation s'appuiera notamment sur les résultats de la preuve de concept « **HDA-Lab** ».
- Le projet associe plusieurs directions du MCC et est réalisé en partenariat avec l'**Institut de recherche et d'Innovation** (IRI) du Centre Pompidou.
- L'objectif est de démontrer l'intérêt du liage d'un corpus organisé tel que Joconde avec les données issues de Wikipédia via DBpédia.
- L'innovation se portera sur les modes d'accès à l'**image** et sur le **multilinguisme**
- Le site expérimental sera rendu public à la fin de l'année et devrait être disponible intégralement en au moins cinq langues dont une langue de France.
- Un bilan sera réalisé en fin de projet pour déterminer quelles innovations peuvent être reprises dans les **outils de travail** utilisés au quotidien au ministère de la Culture.

# 12. Autres projets de sémantisation envisagés

L'année prochaine, une autre expérimentation sera menée sur un **corpus d'archives sonores** en langues de France : Corpus de la Parole

- L'innovation portera sur **la navigation dans un corpus sonore**.
- D'autres expérimentations sont envisagées autour du Wiktionnaire (2 millions de termes en français) pour :
  - proposer des outils de traitement de la parole ou du langage,
  - de classement de termes,
  - de manipulation de synonymes,
  - de suivi des nouveaux termes issus du public...
- Ces expérimentations permettront au MCC d'**améliorer ses outils de travail** en fonction des résultats expérimentaux obtenus.

# Des questions ?

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