Introduction générale

B. Gaiffe

2 décembre 2011

Le projet CLARIN

CLARIN (Common Language Research INfrastructure) :

- Infrastructure européenne
- partage de ressources et d'outils
- on ne s'intéressera ici qu'aux aspects techniques...

De ce point de vue, CLARIN a les mêmes problèmes (a beaucoup de problèmes en commun) avec DARIAH ou avec l'infrastructure IR-Corpus.

Ce qui est techniquement bon pour un projet européen est probablement bon pour un/des projets français !

Partager des ressources dans une infrastructure

Quelques idées de départ :

- pas **UN/LE** serveur central unique, mais plusieurs centres à travers l'europe,
- pas une institution globale à laquelle les chercheurs adhéreraient mais chacun reste identifié dans son institution d'appartenance,
- pas **UNE/LA** langue unique !

Le problème est difficile, mais autant le poser de façon réaliste. Si on reporte le problème au niveau français, mêmes contraintes (à la langue près).

Conséquences en termes techniques

- identifiants uniques pour les ressources (Persistent IDentifiers),
- authentification “unique” (Single Sign On). Imaginez l’accès à des ressources sur plusieurs centres européens avec nécessité de s’identifier sur chacun !
- interopérabilité :
  - en termes de codage des ressources (Unicode, XML, etc.)
  - en termes de description des ressources, i.e. des metadonnées

Interopérabilité des metadonnées

Premier problème : identifier une/des ressources dans un vaste ensemble.

- solution 1 : Google (ou Lucène, ou...). Metadonnées écrites sous forme de texte “libre”
- solution 2 : Isidore (traduction des metadonnées en RDF)
  - pose la question intéressante des référentiels...
  - Isidore ne peut marcher que si il moissone de “bonnes métadonnées”
- solution 3 : des métadonnées typées et structurées

Si on est pragmatique, on fait feu de tout bois ! Mais dans tous les cas de figure, la structuration des métadonnées ne gâte rien !

Formats de metadonnées “structurées”

- metadonnées “à plat” (mais typées)
  - Dublin Core
  - OLAC
- métadonnées structurées
  - IMDI (Isle Meta Data Initiative)
  - MARC (Unimarc, etc.)
- metadonnées semi-structurées
  - entête TEI (Text Encoding Initiative)
  - EAD (Encoded Archival Description)

Bien entendu, on a en fait un continuum.

Un nouveau format de metadonnées structurées !

Ca semble une gageure !
La seule solution est de définir un format très lâche, qui soit personalisable.
L'objectif est double :

- permettre de définir des formats personalisés par des communautés
- encourager la réutilisation

Le problème est toujours le même :

- format trop rigide $\Rightarrow$ peu de chance d'être suffisament précis pour certaines communautés. Or on ne peut pas se permettre de perdre de bonnes descriptions
- format très personalisable $\Rightarrow$ risque de poser des problèmes d'interopérabilité.

Solution proposée par CMDI

- imposer l'ancrage sémantique des feuilles des description
  - concepts employés (ex: birthdate) doivent être définis et identifiés de façon inambigüe et partageable ;
  - valeurs possibles, soit typées (W3C date) soit listées (et les valeurs doivent alors également définies de façon non ambigüe et partageable
- encourager la réutilisation de "composants" de métadonnées.

L'éco-système des metadonnées

- des formats existants ;
- des propositions en cours (metanet, flarenet...)
- des metadonnées existantes
- RDF

CMDI vit plutôt bien dans cet écosystème !

- toutes les metadonnées d'un document ne sont pas forcément à reporter en CMDI (cf Dublin Core)
- CMDI n'est qu'une solution technique, mais :
  - CMDI est obligatoire dans CLARIN
  - CMDI est implémenté (des outils existent et s'améliorent continuellement)
  - CMDI sera, j'espère, un norme ISO