

## Le modèle du Linked Open Data appliqué à des ressources orales

Michel Jacobson & Olivier Baude

Laboratoire Ligérien de Linguistique, UMR 7270, Université d'Orléans, France  
michel.jacobson@gmail.com, olivier.baude@univ-orleans.fr

**Résumé.** Afin de faciliter la réutilisation de corpus oraux par des personnes étrangères à leur collecte, les producteurs ainsi que les gestionnaires (documentalistes, archivistes, etc.) de ces corpus, doivent les organiser et les documenter. Jusqu'à récemment, la réutilisation de ces données était principalement envisagée sous les angles du droit des utilisateurs et de l'interopérabilité entre les machines et entre les logiciels. Depuis le début des années 2000, avec l'arrivée des technologies du « web sémantique » et plus récemment encore avec le mouvement du « Linked Open Data » (LOD), l'interopérabilité est aussi appréhendée au niveau sémantique. Les vocabulaires, ontologies, référentiels disponibles dans différents secteurs permettent aujourd'hui d'envisager d'autres pratiques de documentation. Enfin, les modèles de diffusion ou de mise à disposition des données du LOD ouvrent la porte à de nouvelles organisations pour la gestion de l'information. Nous discuterons de ces nouvelles orientations à travers un retour d'expérience de la plateforme de gestion de corpus oraux Cocoon (Collections de corpus oraux numériques). Seront discutés plus particulièrement les raisons des évolutions dans son modèle de données, ainsi que les avantages fonctionnels que cette plateforme entend tirer du LOD.

**Abstract.** In order to facilitate the reuse of oral corpora by people not involved in the data collection, corpora producers and administrators (librarians, archivists, etc.) must organize and document them. Until very recently, the reuse of these data was primary considered under the view of user rights and interoperability between machines and software. Since the early 2000s, with the emergence of “semantic web” technologies and, more recently, with the movement of “Linked Open Data” (LOD), the interoperability is also apprehended at the semantic level. Vocabularies, ontologies and repositories available in different domains allow us to consider other documentation practices. Finally, the new ways of disseminating and providing data in the LOD open the doors to new organizations for information management. We will discuss these new orientations through a feedback of the oral corpora management platform Cocoon (Collections de corpus oraux numériques). Will be discussed more particularly the reasons of the evolutions in its data model, as well as the functional advantages that this platform intend to take from the LOD.

## 1 Contexte

Les corpus oraux, tout au long de leur vie, peuvent passer par de nombreuses mains. Des collecteurs initiaux aux gestionnaires (documentalistes, archivistes, etc.) jusqu'aux ré-utilisateurs potentiels, ces corpus vont être structurés, organisés et outillés pour des usages variés. La documentation de ces corpus est une des tâches essentielle pour faciliter ces différents usages.

Depuis le début des années 2000, avec l'arrivée des technologies du « web sémantique » et plus récemment encore avec le mouvement du « Linked Open Data » (LOD), un nouveau modèle d'organisation de l'information vient influencer les pratiques documentaires. Ce sont ces nouvelles pratiques et orientations que nous discuterons ici en les illustrant par un cas concret de mise en œuvre dans le cadre d'une plateforme de gestion de corpus oraux. Seront discutés plus particulièrement les raisons des évolutions dans son modèle de données, ainsi que les avantages fonctionnels que cette plateforme entend tirer du LOD.

Avant les années 2000, le terme de « réutilisation » n'était pas employé aussi fréquemment qu'il ne l'est actuellement et les termes « ouverture », « partage » voire « interopérabilité » lui étaient préférés. Il convient de retracer brièvement l'évolution récente de l'usage de ces différentes notions.

La réutilisation des données peut être appréhendée selon différents points de vue. En terme de politique de la recherche, en 2003, la *Déclaration de Berlin sur le Libre Accès à la Connaissance en Sciences exactes, Sciences de la vie, Sciences humaines et sociales*<sup>1</sup>, qui avait pour but de promouvoir Internet « comme instrument fonctionnel au service d'une base de connaissance globale de la pensée humaine » a été signée par de nombreux Directeurs Généraux des Établissements Publics à caractère Scientifique et Technologique (EPST). Cette déclaration posait les bases d'une politique orientée vers le libre accès des données de la recherche et pointait la nécessité de disposer d'un cadre juridique pour l'encadrer.

L'interopérabilité a aussi été abordée sur le plan du droit à travers le problème des licences utilisateurs. Ainsi, après l'impulsion des licences libres pour les logiciels (GPL, Apache, BSD, CeCILL,...), de nouvelles licences, inspirées des premières, ont vu le jour pour permettre aux auteurs d'œuvres de définir comment les utilisateurs pouvaient utiliser et redistribuer ces œuvres. C'est ainsi que sont nées, par exemple, les licences Creative Commons. Par la suite, le mouvement d'ouverture des données publiques (*open-data*) engagé par de nombreux gouvernements a donné lieu à l'élaboration et l'usage de nouvelles licences. Pour ne citer que les plus répandues en France : la licence ouverte d'Étalab ou l'Open Database License (ODbL) de l'Open Knowledge Foundation. Ce sont les concepts d'ouverture et de partage qui sont ici mis en avant.

L'interopérabilité, a également été développée au plan technologique sur des aspects liés au caractère technique de l'outil informatique. C'est ainsi que les supports numériques ont connus une vague de normalisation leur permettant d'être lus plus facilement sur diverses plateformes. Puis les échanges de supports entre utilisateurs ont massivement diminués pour laisser place à des échanges « dématérialisés » sur les réseaux. A partir de là, l'interopérabilité est alors essentiellement pensée au niveau de protocoles d'échange, du codage et du formatage des données. Dans ce domaine, on a pu observer dans les dernières années la naissance de standards et de normes dont certains ont une importance majeure comme le standard Unicode pour le codage des caractères (normalisé au sein de l'ISO-10646) ou le standard eXtensible Markup Language (XML) du W3C. Sur ces briques de base, de nombreux autres formats (par exemple « Office Open XML » et « Open Document Format » pour la bureautique ; SVG<sup>2</sup> pour l'image

vectorielle), protocoles (par exemple SOAP<sup>3</sup>, OAI-PMH<sup>4</sup>) et langages (par exemple « Mathematical Markup Language » pour les expressions mathématiques, « Music Markup Language » pour la notation musicale) ont pu se construire. C'est le concept d'interopérabilité qui est dans ces cas le plus souvent invoqué.

Le domaine documentaire n'a pas échappé à cette tendance et les techniques d'interopérabilité évoquées plus haut ont aussi été largement utilisées pour définir des langages de description de ressources plus ou moins génériques. Notamment, le Dublin-Core (Norme ISO 15836) a apporté un modèle de description de base très largement utilisé et souvent repris dans des contextes d'utilisation métier en précisant l'acception ou en l'enrichissant. Dans le domaine culturel on observe que chaque communauté a développé ou fait évoluer ses propres modèles, en s'inspirant de ces briques d'interopérabilité de base. On retrouve ainsi de manière un peu caricaturale :

- pour le secteur des archives le modèle EAD (Encoded Archival Description) qui reprend en SGML puis à partir de 2002 en XML les principes de la norme ISAD-G ;
- pour le secteur des bibliothèques les modèles MARC (MACHINE-Readable Cataloging), avec une déclinaison en XML ;
- pour le secteur des musées le modèle CIDOC-CRM modèle conceptuel de référence pour l'information muséographique du Comité international pour la documentation de l'ICOM – Conseil international des musées.

Cette diversité des modèles s'explique en partie par les différences de nature des objets manipulés (objets de musées, fonds d'archives, exemplaires édités). Ils vont donc insister sur certaines particularités des objets, comme leur mode de production ou la structure hiérarchique des fonds pour les archives. Mais de nombreux autres aspects pourraient être facilement partagés entre ces trois domaines. Ce constat peut s'observer par exemple avec l'entrée des bibliothèques dans les travaux du CIDOC-CRM. Le nouveau modèle conceptuel des bibliothèques FRBR (Functional Requirements for Bibliographic Records) a été mis dans une forme *orientée objet* FRBR-OO afin d'en faire une extension du CIDOC-CRM. Enfin, on voit aussi naître des projets transverses autour de la notion d'objets culturels comme, en France, le projet HADOC (Harmonisation de la production des Données Culturelles) du ministère de la culture et de la communication.

Aujourd'hui, les formalismes du web sémantique (le langage RDF, les ontologies RDFS, OWL, le langage de requête SPARQL...) apportent de nouveaux outils pour le codage des modèles élaborés au sein des différentes communautés. Ces formalismes permettent d'explicitier la sémantique contenue dans ces modèles. L'aspect formel et standardisé permet aux machines, non pas de comprendre l'information mais de pouvoir la traiter de manière adaptée (inférences) et automatique.

Avec ces nouvelles avancées, le mouvement Open Data a pu trouver un cadre technique pour la mise en œuvre de ses principes d'ouverture et d'interopérabilité. Plus précisément le Linked Open Data pose les bases d'un modèle d'organisation intimement intégré dans l'écosystème du Web (identification des « choses » avec des URIs, disponibilité des données en format RDF, négociation de contenu, liage entre les entités) et mettant sur le devant de la scène la notion de « réutilisabilité ». Cette organisation définit une nouvelle strate au Web en l'orientant vers ce que l'on nomme également « Web de données » ou « Web 3.0 ».

C'est dans ce cadre que nous proposons de porter un regard sur l'expérience d'un projet d'exposition, de diffusion et d'archivage de données linguistiques orales, celui de la plateforme Cocoon.

## **2 Retour d'expérience sur le tournant du web de données par la plateforme Cocoon**

En 2006, à l'initiative du CNRS (Département SHS et Direction de l'information scientifique), naissent des centres de ressources numériques dont les périmètres sont définis en fonction de la nature des ressources qu'ils gèrent : les informations spatiales, visuelles, textuelles, orales, etc. Le CRDO (Centre de Ressources pour la Description de l'Oral) rassemblement de deux propositions s'est inscrit dans ce mouvement puis a donné naissance à deux entités : d'une part le SLDR intégré dans l'Equipex Ortolang en 2015 et d'autre part la plateforme Cocoon (Collections de corpus oraux numériques) gérée conjointement par deux laboratoires de linguistique, le LACITO et le LLL. Il convient de noter que les centres de ressources sur l'oral ont été associés au projet pilote sur l'archivage des données de la recherche au sein du CNRS.

Nous retraçons dans ce qui suit les différents choix opérés par la plateforme sur cette période qui couvre une dizaine d'année. Ce retour d'expérience permet d'illustrer les différents aspects touchés par la transition de cette organisation vers les techniques du web sémantique et vers le modèle du web de données pour la structuration de l'information et son exposition.

### **2.1 Le modèle de données OLAC**

Afin de définir un modèle pour ses données, la plateforme Cocoon a suivi dès sa création en 2006 les recommandations de la communauté OLAC (Open Language Archive Community). Ces recommandations tiennent essentiellement en deux points : l'utilisation du format OLAC pour le codage des métadonnées et l'utilisation du protocole OAI-PMH pour la diffusion de ces métadonnées. Le format de description d'OLAC est défini comme une spécialisation de celui du Dublin-Core. Il reprend les 15 éléments du Dublin-Core (DC) ainsi que la quarantaine d'éléments supplémentaires du Dublin-Core qualifié (DCQ). Il en donne une acception pour le domaine de l'archivage des langues et ajoute quelques vocabulaires contrôlés (pour les types de discours, les types linguistiques, les champs disciplinaires, les rôles des participants et l'identification des langues). La forme que prend cette spécialisation est un schéma XML classique facilitant ainsi la validation et la création d'outils d'édition.

Après presque 10 ans d'utilisation de ce modèle, nous présentons ici quelques limites auxquelles la plateforme Cocoon s'est heurtée et examinerons ensuite les pistes qu'elle suit actuellement pour tenter de les dépasser.

### **2.2 Les limites**

Une des premières difficultés dans l'utilisation du modèle OLAC, mais qui représente également une de ses plus grandes forces, est sa simplicité. Cette perception de simplicité est généralement héritée de la représentation que l'on se fait en première approche du DC : « indispensable, mais jamais suffisant ». A l'analyse, le schéma n'est pourtant pas si pauvre lorsqu'on fait usage de toute sa richesse en termes de rubriques d'informations et de précision d'encodage. Ainsi, s'il n'existe effectivement qu'une seule étiquette pour les informations de localisation géographique (spatial), il est possible d'en utiliser

plusieurs pour donner plusieurs localisations, plusieurs types de localisation, ou pour exprimer une localisation en plusieurs langues. Pour préciser le type de localisation il est nécessaire de préciser l'encodage (par exemple : un code ISO3166 pour identifier un pays, un point géographique en donnant ses valeurs de longitude, latitude, altitude, une zone en donnant les coordonnées du plus petit rectangle englobant ou encore un identifiant tiré du Thesaurus of Geographic Names). Enfin, il est aussi possible d'utiliser les mécanismes d'extension du DC pour préciser des acceptions plus étroites d'une rubrique ou pour réutiliser ses propres syntaxes et vocabulaires comme le fait par exemple OLAC pour le domaine linguistique. Pour autant, si la multiplication des étiquettes de même type permet d'être plus précis, le modèle ne permet pas d'indiquer les éventuelles relations qui peuvent exister entre elles. Par exemple si une étiquette précise le code pays et une autre la commune, rien ne permet d'inférer que la commune est dans ce pays. Autre exemple : si pour une ressource deux mots-clés sont donnés l'un en français et l'autre en anglais, rien ne permet de savoir s'il s'agit d'un même mot-clé ou de deux mots-clés distincts.

Une autre difficulté d'utilisation est une conséquence du choix de la plateforme Cocoon de faire des descriptions au niveau des documents. Ce choix a été guidé par la volonté de décrire le plus finement possible les ressources. Il est effectivement plus facile de décrire précisément le type ou le format d'un enregistrement ou d'une transcription que d'un regroupement des deux. Ce choix est également du au fait que le mode de production et le cycle de vie des documents n'est pas obligatoirement le même d'un type de document à un autre. Ainsi, les enregistrements sont (presque) toujours faits avant les transcriptions et ils ne sont pratiquement jamais modifiés après la phase de collecte. Les transcriptions, elles, sont parfois réalisées longtemps après l'enregistrement, éventuellement par d'autres chercheurs que ceux qui ont effectué l'enregistrement et parfois avec un objectif scientifique différent. Enfin, elles sont souvent modifiées pour apporter des corrections ou de nouvelles informations. Pour pallier cette difficulté tout en restant dans le modèle proposé la plateforme Cocoon a spécialisé certaines étiquettes de relation. Ainsi, pour indiquer les liens entre transcriptions et enregistrements ce sont les étiquettes `requires` et `isRequiredBy` qui ont été utilisées. Ont également été ajoutés des objets de type « collection » qui rassemblent des groupes plus ou moins vastes d'enregistrements, de transcriptions et éventuellement de sous-collections (par exemple toutes les ressources récoltées lors d'une enquête ou toutes les données d'un chercheur ou les données d'un laboratoire). Les liens entre les membres et leur collection sont de type `partitive` (`isPartOf`, `hasPart`). Une même ressource pouvant être représentée en différents formats, ce sont des relations `isFormatOf` qui ont été utilisées. Les transcriptions pouvant suivre différents modèles, ce sont des relations `conformsTo` qui ont été utilisées. On se doit donc de constater que l'interprétation correcte d'une ressource n'est pas évidente car elle requière de suivre l'ensemble de liens qui la lie à d'autres ressources.

Enfin un dernier problème réside dans la difficulté d'exposition des données. Pour donner aux utilisateurs une représentation intelligible des objets décrits dans la plateforme (collections, enregistrements, transcriptions, etc.), il faut composer des interfaces mélangeant des informations tirées de plusieurs ressources or il est difficile, pour une machine, de faire la part entre les ressources, leurs relations et leurs représentations.

### **2.3 Les évolutions en cours**

Afin de pallier ces limites, les gestionnaires de la plateforme Cocoon ont pris la décision de se lancer dans l'utilisation des technologies du web sémantiques et dans la mise en œuvre du modèle d'exposition des données du Linked Open Data. En effet, la principale piste de progrès dans la gestion des données de

la plateforme Cocoon est envisagée à ce jour par un changement du modèle de description. Ce changement tente de conjuguer harmonieusement la finesse du grain (au document) et la clarté du codage. Le modèle de donnée de départ est celui d'OLAC expliqué plus haut, alors que le modèle d'arrivé est basé sur le langage RDF. Nous discuterons dans ce qui suit les choix effectués dans la mise en œuvre de ce changement.

Ce travail de changement de modèle a suivi 3 étapes : (I) une étape d'identification des informations (métadonnées) que la plateforme doit gérer ; (II) une étape de définition formelle du modèle cible pour exprimer ces informations et (III) la migration des informations de l'ancien vers le nouveau modèle.

## **2.4 Les identifiants et les alignements**

Une étape préalable a consisté à aligner un certain nombre de rubriques d'informations sur des référentiels externes. Les objectifs initiaux étaient :

- de pouvoir identifier des informations plutôt que de simplement les nommer avec des littéraux. Identifier permet d'éviter de l'homonymie en distinguant les informations éventuellement de même nom. Par exemple, le nombre de chercheurs et de locuteurs référencés dans la plateforme étant croissant, la présence d'homonymies augmente au risque de perturber la recherche ou du moins l'interprétation de ses résultats ;
- de pouvoir enrichir les informations affichées avec les informations tirées des référentiels utilisés. En effet si la description de ces informations n'est pas au cœur du métier des gestionnaires de la plateforme, il est préférable d'utiliser les identifiants d'autres organismes dont c'est l'activité principale moyennant que ces identifiants soient pérennes. C'est ainsi que la description des lieux, des auteurs, des langues et des mots-clés a été déportée. Cela permet par exemple de donner pour une langue, ses différentes appellations, son système d'écriture, son rattachement dans un arbre phylogénétique, pour un auteur sa bibliographie ou sa biographie, pour un lieu ses niveaux englobant (région, pays...), sa monnaie, son histoire...
- de limiter la redondance en évitant la duplication dans les métadonnées des ressources d'une même information en différentes langues ou à différents niveaux de précision. La récupération de ces informations pouvant être faite en interrogeant les référentiels.

Une conséquence secondaire est que ce premier travail facilitera la transition vers le futur modèle de données cible en RDF en définissant ou réutilisant déjà un certain nombre d'URI. Une autre conséquence sera de pouvoir améliorer les fonctions de recherche par l'utilisation des concepts (mots-clés) dans toutes leurs richesses avec leurs formes génériques et spécifiques, leurs formes non préférentielles ou encore les formes équivalentes dans d'autres langues.

### **2.4.1 Premier exemple d'alignement : les fichiers d'autorité**

Dans un premier temps la plateforme Cocoon a aligné une partie des contributeurs (ceux dont le rôle était « déposant ») avec le référentiel VIAF (Virtual International Authority File). Ce référentiel est né de la volonté de lier entre eux les référentiels d'autorité des bibliothèques nationales et autres grands catalogues. Si un déposant a déjà publié, il y a de fortes chances que l'on puisse trouver une ou plusieurs notices le concernant dans des bibliothèques et donc qu'un identifiant VIAF lui ait été attribué. Cet

identifiant peut alors être utilisé pour identifier de manière unique l'auteur, ainsi que pour accéder à des informations le concernant. Dans la plateforme Cocoon cet alignement a aussi été utilisé pour pouvoir récupérer et afficher dynamiquement des informations complémentaires. Pour chaque déposant, une page est donc construite affichant le nom normalisé ou préférentiel de l'auteur ainsi que la liste de ses publications dans les principales sources françaises (Abes, BnF). L'entrepôt HAL (Hyper Article en Ligne) a aussi été ajouté car il permet de récupérer également de la littérature grise (pré-prints, articles non édités, etc.) et qu'il permet éventuellement un accès direct aux documents.

La récupération des informations de la Bnf se fait par l'interrogation en SPARQL de data.bnf.fr car l'alignement avec VIAF y est explicite. La récupération des informations de l'Abes se fait en récupérant dans VIAF l'identifiant Idref de l'auteur, s'il existe, puis en récupérant la notice correspondant à cet identifiant dans un format XML/RDF. Enfin la récupération des informations de HAL se fait en utilisant un webservice avec en paramètre l'identifiant VIAF de l'auteur. Toutefois l'introduction de cet identifiant dans HAL est très récent et demande une intervention des auteurs eux-mêmes pour le renseigner, de sorte que très peu d'auteurs l'ont fait pour l'instant.

Cet alignement s'est poursuivi pour d'autres contributeurs, notamment ceux dont le rôle est « chercheur » car ils ont également de fortes chances d'avoir déjà publié et qu'il soit donc possible de récupérer leurs identifiants VIAF. Pour les autres rôles, notamment les « locuteurs », la plateforme construit son propre référentiel qui peut porter en toute autonomie les propriétés de descriptions que les projets de recherche utilisent (par exemple des informations socio-linguistiques, des codes d'anonymisation, etc.).

**Carton, Fernand**

« Carton, Fernand » ( Personne )

Voir la notice d'autorité sur Virtual International Authority File (VIAF): <http://viaf.org/viaf/51686843>

Rechercher s'il existe dans l'entrepôt, des ressources cet acteur en tant que contributeur ou éditeur

**Références BnF (data.bnf.fr)**

- Dictionnaire du français régional du Nord-Pas-de-Calais. Paris : Éd. Bonneton , 1991. 125 p. <http://catalogue.bnf.fr/ark:/12148/cb354619790>
- Expressions et dictons du Nord-Pas-de-Calais. Paris : Bonneton , 2004. 192 p. <http://catalogue.bnf.fr/ark:/12148/cb39163250v>
- Expressions et dictons du Nord Pas-de-Calais. Paris : Bonneton , cop. 2007. 1 vol. (191 p.) <http://catalogue.bnf.fr/ark:/12148/cb41048350r>
- Le parler du Nord Pas-de-Calais. Paris : Bonneton , cop. 2003. 125 p. <http://catalogue.bnf.fr/ark:/12148/cb389777123>
- Le parler du Nord Pas-de-Calais. Paris : Bonneton , impr. 2006. 1 vol. (159 p.) <http://catalogue.bnf.fr/ark:/12148/cb40943010w>
- Atlas linguistique et ethnographique picard Volume 1. Paris : Ed. du Centre national de la recherche scientifique , 1989. 317 p. <http://catalogue.bnf.fr/ark:/12148/cb35346851z>
- Atlas linguistique et ethnographique picard Volume II. Paris : CNRS éd. , 1998. Non paginé <http://catalogue.bnf.fr/ark:/12148/cb36702970j>
- Introduction à la phonétique du français. Paris : Bordas , 1979. 250 p. <http://catalogue.bnf.fr/ark:/12148/cb346493014>
- Introduction à la phonétique du français. Paris : Bordas , 1987. 250 p. <http://catalogue.bnf.fr/ark:/12148/cb34955914w>
- Introduction à la phonétique du français. Paris : Dunod , 1997. 250 p. <http://catalogue.bnf.fr/ark:/12148/cb36197691q>
- Introduction à la phonétique du français. Paris : Dunod , 1994. 250 p. <http://catalogue.bnf.fr/ark:/12148/cb37460860b>
- Index lemmatisé et étymologique de l'"Atlas linguistique et ethnographique picard", volumes 1 et 2, de Fernand Carton et Maurice Lebègue. Amiens : Centre d'études picardes, Université Picardie-Jules Verne , 2010. 1 vol. (IV-218 p.) <http://catalogue.bnf.fr/ark:/12148/cb42410334s>
- La littérature picarde aux siècles "classiques", 17e et 18e siècles. Amiens : Langue et culture de Picardie, Office culturel régional de Picardie , impr. 2007. 1 vol. (526 p.) <http://catalogue.bnf.fr/ark:/12148/cb410598183>
- Atlas linguistique et ethnographique du Centre Volume 3. Paris : Éditions du Centre national de la recherche scientifique , 1976. 1 vol. (C.1098-1505) : cartes, ill. ; 50 cm <http://catalogue.bnf.fr/ark:/12148/cb40887531v>
- Les Parlers d'Aubers-en-Weppes... Arras : Société de dialectologie picarde , 1971. 27 cm, 175 p., ill., plan h.-t., couv. ill. 25 F <http://catalogue.bnf.fr/ark:/12148/cb353559904n>
- Recherches sur l'accentuation des parlers populaires dans la région de Lille. Lille : Service de reproduction des thèses de l'université , 1972. 23 cm, 363 p., ill., graph. h.t., multigr <http://catalogue.bnf.fr/ark:/12148/cb35937781p>
- Introduction à la phonétique du français. Paris ; Bruxelles ; Montréal : Bordas , 1974. 250 p. <http://catalogue.bnf.fr/ark:/12148/cb34561001m>

**Références ABES (idref)**

- François Cottignies dit Brûle-Maison : 1678-1740 : Chansons et pasquille / François de Cottignies / Ed. critique avec introd. étude grammaticale et glossaire / Arras : Archives du Pas de calais , 1965
- Un patoisant lillois au XVIIIe siècle : Jacques Decottignies / Fernand Carton / Villeneuve d'Ascq : Dactylogramme , 1964
- Les parlers d'Aubers-en-Weppes [Texte imprimé] / Fernand Carton & Pierre Descamps / Arras : Société de dialectologie picarde , 1971
- Recherches sur l'accentuation des parlers populaires dans la région de Lille [Texte imprimé] / Fernand Carton ; [sous la direction de Georges Straka] / Lille : Service de reproduction des thèses de l'université Lille III , 1972

Figure 1. « Mashup » affichant les références bibliographiques d'un auteur issues de différentes sources sur la base de son identifiant VIAF

## 2.4.2 Deuxième exemple d'alignement : les lieux d'enregistrement

Pour les données audio ou vidéo, la plateforme Cocoon a tenté de systématiser le renseignement du lieu géographique des enregistrements. Cette information peut être une donnée assez importante comme c'est le cas dans les enquêtes dialectologiques. Dans un premier temps, un alignement d'une partie de ces

informations sur le TGN a été effectué, puis plus récemment l’alignement s’est poursuivi sur les référentiels Geonames et Dbpedia. Les alignements sur Geonames ont été faits systématiquement, dans le cadre d’un projet particulier, pour les enregistrements des Atlas linguistiques de France (Picard, Bretagne, Gascogne, Alsace, etc.) au niveau de la commune. Puis, à partir de l’information d’identification INSEE de la commune récupérée dans Geonames, un alignement a été fait en rebond vers le Dbpedia français. Une fois ces alignements effectués, un prototype d’interface de navigation dans les enregistrements de ces Atlas a été réalisé en projetant toutes les coordonnées géographiques sur une carte de France et en permettant d’afficher pour chaque point les ressources disponibles dans Cocoon ainsi que l’image de la commune, le résumé de sa présentation et le lien vers Wikipedia pour en savoir plus. Compte tenu de l’utilisation assez fréquente de ces deux référentiels ou des informations qu’ils contiennent, le rapprochement d’informations réparties dans différents gisements d’information devient une tâche beaucoup plus facile à réaliser que par le passé et il devient possible d’imaginer toutes sortes d’applications et de réutilisation à buts culturels, scientifiques ou commerciaux.

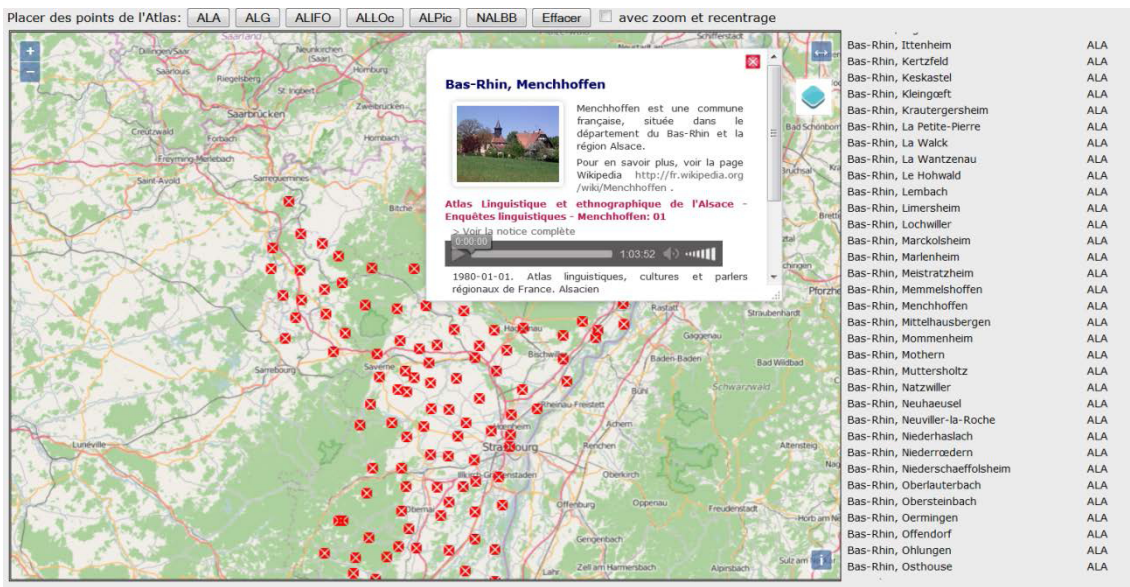


Figure 2. « Mashup » affichant différentes sources d’informations géographiques à propos d’un point d’une enquête dialectologique

## 2.5 Au-delà des alignements

Les alignements ne sont pas un but en soi, mais représentent pour la plateforme Cocoon une première étape de transition vers un nouveau mode de structuration et de mise à disposition de l’information. C’est également un service qui répond à des demandes d’utilisateurs comme par exemple le projet *Corpus de la Parole* de la Délégation générale au français et aux langues de France qui s’inscrit dans une politique affichée de développement de projets reposant sur un alignement des données.

La deuxième étape complémentaire est d’identifier les informations qui relèvent des compétences des gestionnaires de la plateforme ou pour lesquelles n’a pas été trouvé de référentiels adaptés.

Pour les ressources primaires de la plateforme Cocoon (enregistrements, transcriptions, collections), un système d’identification basé sur l’OAI avait déjà été mis en place. Ces identifiants sont repris dans des

URI de type POI<sup>5</sup> utilisant le système PURL<sup>6</sup>. Ce sont ces POI qui sont réutilisés dans le cadre du futur modèle pour identifier les ressources primaires. Les autres identifiants, dont l'utilisation n'est pas systématique (les ARK<sup>7</sup> affectés par le CINES lors de leur archivage, les Handles affectés par Isidore lors de leur moissonnage) seront uniquement véhiculés dans le modèle de données comme des propriétés d'identification.

Pour certaines informations, il n'a pas encore été trouvé ou choisi de référentiels. Ainsi, pour les organisations (laboratoires ou autres structures de recherche ou culturelles ayant contribué à la constitution des ressources), l'ISNI (International Standard Name Identifier) semble un bon candidat, mais une étude doit être menée pour évaluer la couverture que ce standard représente sur les données de la plateforme. Pour les typologies (les types de discours, les genres, les types de supports et de format, etc.), il existe de nombreux vocabulaires contrôlés qui peuvent être réutilisés. En particulier ceux de OLAC, qui sont déjà utilisés, seront sans doute repris. Toutefois leur forme de publication les rend difficilement réutilisables directement et un portage doit être fait pour leur affecter des URI ce qui entraîne des difficultés de maintien et de partage de ce travail. Une autre piste est celle de l'utilisation des vocabulaires de la BnF ou de la bibliothèque du congrès dont la forme se prête plus facilement à l'utilisation que l'on souhaite en faire. Les typologies de OLAC pourraient alors être traduites ou alignées sur ces vocabulaires.

Pour l'indexation matière, une expérimentation sur la collection « corpus de la parole est menée actuellement ». Dans le cadre de ce programme conjoint du CNRS et du Ministère de la Culture dédié à la conservation et la diffusion des corpus en français et langues de France à valeur patrimoniale, un projet orienté vers les développements du web sémantique a été initié en 2015. Sur proposition de la Délégation générale à la langue française et aux langues de France, le Ministère de la Culture a en effet sélectionné l'entrepôt de corpus linguistique « Corpus de la parole » pour une initiative innovante et exploratoire de « sémantisation » de données culturelles. Cette expérience fait suite au projet « Jocondelab » réalisé en 2014 à partir d'une base de données d'œuvres d'art numérisées. L'objectif de cette seconde expérience est de tester les apports du web sémantique sur des données linguistiques produites dans un cadre scientifique mais conservées et diffusées dans un cadre plus large. Ainsi pour mener à bien ce projet il a tout d'abord fallu enrichir les métadonnées avec une indexation utilisant le référentiel Rameau (Répertoire d'autorité matière encyclopédique et alphabétique unifié). Une autre piste pour le même travail pourrait être l'utilisation du référentiel Dbpedia qui, lui aussi à une vocation encyclopédique de même nature. Une étude de comparaison devrait être menée pour évaluer les impacts de l'utilisation de l'un ou l'autre référentiel.

Enfin, pour les locuteurs, il n'y a par nature aucun référentiel possible réutilisable puisque chaque projet produit ses propres éléments. Cocoon se dirige pour ce dernier vers la création d'un référentiel interne. Les choix à opérer pour ce cas de figure sont essentiellement ceux portant sur les vocabulaires employés pour décrire les individus. Sans doute le noyau du modèle tournera-t-il autour de l'ontologie FOAF<sup>8</sup> et des ajouts de propriétés issues d'autres vocabulaires seront fait au fur et à mesure des besoins. Pour ce référentiel, l'extensibilité du modèle est très importante dans la mesure où tout nouveau projet scientifique sera susceptible d'ajouter de nouvelles spécificités.

### **3 Choix du nouveau modèle de données**

Les principaux objectifs dans la construction du futur modèle, étaient que celui-ci exprime au mieux les relations entre les ressources et entre les ressources et leurs représentations, qu'il évite le plus possible la

redondance d'informations et qu'il soit extensible de manière à pouvoir répondre facilement à des demandes d'ajout de type d'informations.

Comme pour tout projet, les solutions se répartissent entre l'utilisation directe d'un modèle existant, l'adaptation à des besoins spécifiques d'un modèle proche ou la définition complète d'un nouveau modèle. Cocoon, dans un premier temps, a examiné un certain nombre de modèles candidats utilisés dans des projets de même type que celui que nous présentons ici : l'EAD, OAI-ORE et EDM.

Le modèle EAD est principalement utilisé dans le monde des archives. Il permet de faire des descriptions hiérarchiques multi-niveaux. Chaque « niveau de description » mutualisant des éléments de description dont héritent les niveaux inférieurs. Ce modèle, après une première définition dans le format SGML, a été traduit en XML. La version actuelle de référence est une DTD datant de 2002. Des travaux, toujours en cours à ce jour, visent à exprimer le modèle avec une technologie de schémas (Relaxng et XML-Schema) et à apporter quelques améliorations marginales. Ce modèle répond assez bien aux besoins de la plateforme d'éviter la redondance d'informations, ainsi qu'à celui d'exprimer les relations. Les inconvénients de ce modèle sont que les relations (entre niveaux) sont strictement hiérarchiques, que dans sa forme actuelle, il se prête assez difficilement à la description de fonds numériques et que son extensibilité est très réduite.

Le modèle OAI-ORE (Open Archives Initiative Object Reuse and Exchange) est organisé autour du concept central d'agrégation. Les spécifications de ce standard définissent comment des agrégations de ressources web doivent être décrites et échangées. Le modèle est défini en RDF et fait usage de vocabulaires tels que DC, DCT, FOAF et ceux propres aux web sémantique comme OWL et RDFS. Ce modèle répond bien au besoin de la plateforme en matière d'expression des relations en définissant formellement le principe d'agrégation et en séparant la description des agrégations de la description de leurs ressources. Ce modèle se veut agnostique dans la mesure où il ne dit rien sur la manière de décrire intellectuellement les agrégations ni les ressources Web, ce qui peut être fait avec les ontologies de son choix, héritant en cela du caractère extensible du formalisme RDF utilisé.

Le modèle EDM (Europeana Data Model) est un modèle défini dans le cadre du projet européen Europeana, un portail d'accès unifié à des ressources culturelles issues principalement de bibliothèques, de musées et de centres d'archives. Il est exprimé en RDF et reprend les concepts de l'OAI-ORE (agrégation, ressources web, carte des ressources) pour les adapter à ses besoins spécifiques. Il complète le modèle avec des classes (Event, Place, Agent, Concept) inspirées d'autres ontologies (FOAF, SKOS,...). Tout comme OAI-ORE, ce modèle répond bien aux besoins de la plateforme en matière d'expression de relations, de minimalisation de la redondance ainsi que d'extensibilité. Le problème est que cette extensibilité est dépendante du projet Europeana : seules rentrent dans le modèle les classes dont le projet a besoin.

Les responsables techniques et scientifiques de la plateforme Cocoon ont donc finalement fait le choix du modèle EDM qui est le modèle qui se rapproche le plus des besoins repérés. Toutefois, comme l'alimentation du portail Europeana n'est pas l'objectif qui a guidé ce choix, un certain nombre d'écarts au modèle sera effectué afin de l'adapter aux besoins spécifiques de la plateforme. Ces écarts seront effectués en respectant la compatibilité des modèles. Le principal ajout effectué sera celui des rôles des contributeurs, présents dans OLAC mais absent de EDM.

## 4 Conclusion

Faciliter la réutilisation des données est incontestablement l'un des enjeux majeurs des initiatives actuelles de gestion des données de la recherche. L'expérience concrète de la plateforme Cocoon dédiée aux données orales produites en sciences humaines et sociales est une contribution à ce défi. Les efforts faits sur la maîtrise du processus d'archivage et notamment sur les opérations de description fine et profonde des ressources ouvrent des perspectives particulièrement intéressantes dans le cadre du linked open data. Ces perspectives ne sont pas encore réellement exploitées dans le monde de la recherche et il n'est pas étonnant de constater que c'est l'opportunité d'un projet transversal entre le ministère de la recherche et celui de la culture qui a déclenché la réflexion et la mise en œuvre d'un changement du modèle d'exposition des données ce qui a un impact sur l'ensemble de la plateforme.

Il serait illusoire de penser que le linked open data se réduise à une opération de liage de données. Cet objectif repose avant tout sur des phases qui nécessitent d'identifier les choses contenues dans les réservoirs de données ce qui donne toute sa place au travail scientifique et technique d'usage de référentiels et de vocabulaires contrôlés, puis au rôle du modèle de données. Ceci est d'autant plus important qu'il s'agit d'un domaine très peu maîtrisé par les chercheurs producteurs de données. Or si l'on peut concevoir le linked open data comme un traitement de l'information offrant de nouvelles possibilités de construction du savoir, il convient de conduire ces travaux dans un dialogue toujours plus poussé entre les acteurs de la recherche, de la documentation, de la gestion des corpus et de l'usage de données.

## Références bibliographiques

- Baude, O., Marchello-Nizia, C., Mondada, L., Blanche-Benveniste, C., Calas, M.-F., Cappeau, P., Cordereix, P., Lamberterie, I. D., Goury, L., & Jacobson, M. (2006). *Corpus oraux?: guide des bonnes pratiques* (O. Baude, éd.). CNRS Éditions et Presses universitaires d'Orléans.
- Bermès, E., (2013), *Le Web sémantique en bibliothèque*, Editions du cercle de la librairie, 2013.
- Cordereix, Pascal, « Les fonds sonores du département de l'audiovisuel de la bibliothèque nationale de France », *Le temps des médias* n°5. Éditions du nouveau monde, p 253-264, 2005.
- Definition of the Europeana Data Model v5.2.6,  
[http://pro.europeana.eu/files/Europeana\\_Professional/Share\\_your\\_data/Technical\\_requirements/EDM\\_Documentation/EDM%20Definition%20v5.2.6\\_01032015.pdf](http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM%20Definition%20v5.2.6_01032015.pdf), (consulté le 05/12/2015)
- Feigenbaum, L., (2010), « The Semantic Web in Action », *Scientific American*,?  
<http://www.thefigtrees.net/lee/sw/sciam/semantic-web-in-action> (consulté le 15/08/2015)
- Huc C., Habert B., (2010) "Building together digital archives for research in social sciences and humanities", *Social Science Information*, vol. 49, n°3, p. 415-443.
- Jacobson M, Baude O., (2012) « Corpus de la parole : collecte, catalogage, conservation et diffusion des ressources orales sur le français et les langues de France », in *Ressources linguistiques libres, TAL*. Volume 52 – n° 3/2011, 47-69
- Principaux standards du web Sémantique : les URI, RDF et SPARQL,  
[http://www.bnf.fr/fr/professionnels/web\\_semantique\\_boite\\_outils/a.web\\_semantique\\_standards.html](http://www.bnf.fr/fr/professionnels/web_semantique_boite_outils/a.web_semantique_standards.html) (consulté le 10/12/2015)

Resource Description Framework, <http://www.w3.org/RDF/> (consulté le 10/12/2015)

---

<sup>1</sup> [http://openaccess.mpg.de/68042/BerlinDeclaration\\_wsis\\_fr.pdf](http://openaccess.mpg.de/68042/BerlinDeclaration_wsis_fr.pdf)

<sup>2</sup> Scalable Vector Graphics

<sup>3</sup> Simple Object Access protocol

<sup>4</sup> Open Archives Initiative - Protocol for Metadata Harvesting

<sup>5</sup> PURL-based Object Identifier

<sup>6</sup> Persistent uniform resource locator

<sup>7</sup> Archival Resource Key

<sup>8</sup> Friend of a friend