



## **Agrégation et diffusion des notices d'autorité en Linked Open Data<sup>1</sup>**

### **Auteurs <sup>2</sup>:**

#### **Xavier Agenjo,**

Directeur de projets de la Fundación Ignacio Larramendi et directeur de la Biblioteca Virtual de Polígrafos  
[xavier.agenjo@larramendi.es]

#### **Francisca Hernández.**

Bibliotecaire. Consultante de DIGIBÍS, Producciones Digitales  
[francisca.hernandez@digibis.com]

#### **Viedma, Andrés.**

Architecte de logiciels.  
DIGIBÍS, Producciones Digitales  
[andres.viedma@digibis.com]

*Traduit de l'espagnol par :  
Nadine Boddaert*

**Meeting:**

**80 — Cataloguing: breaking barriers — Cataloguing Section**

### **Résumé :**

Description, à travers une analyse de la Bibliothèque numérique des polygraphes (Biblioteca virtual de polígrafos), de l'agrégation et de la diffusion des données des notices d'autorité à travers le Linked Open Data (ou Web de données). L'objectif est de réunir information, données, textes numériques et sites WEB sur les polygraphes espagnols, hispanoaméricains, brésiliens et portugais de toutes les époques. De ce fait, ces agrégations réunissent de l'information sur la pensée, la philosophie, la politique, les sciences, etc. de l'Espagne, de l'Amérique hispanique, du Portugal et du Brésil, écrite dans n'importe quelle langue (latin, arabe, hébreu, espagnol, portugais...) et à n'importe quelle époque (depuis Sénèque au 1<sup>er</sup> siècle av. J.-C. jusqu'aujourd'hui).

La colonne vertébrale du système est constituée par les auteurs. Pour chaque auteur on crée une notice d'autorité conforme à MARC 21/RDA qui est enrichie de données biographiques, les attributs spécifiques sont catégorisés afin d'enrichir les relations et de favoriser la

<sup>1</sup> Toutes les URL et données chiffrées de ce texte correspondent à la date du 30 avril 2011.

<sup>2</sup> Nous remercions César Juanes, du Département I+D de DIGIBIS, pour la révision du texte de cette communication.

navigation sur le site (profession, activités, genre, appartenance, dates de naissance et de mort, lieu de naissance et de mort, langues et écritures). Dans ce sens, chaque notice d'autorité agrège de l'information provenant de multiples sources et vocabulaires. Pour chaque auteur on décrit également, selon MARC/RDA (12e mise à jour) les versions numériques de ses œuvres. De la même manière, chaque auteur est mis en relation avec d'autres auteurs, comme les traducteurs, éditeurs ou commentateurs, comme procédure pour suivre la transmission textuelle de leurs œuvres, en utilisant les notices auteur-titre structurées selon le modèle Œuvre-Expression-Manifestation-Item conformément à la proposition récente de MARBI.

La création de ces notices d'autorité s'est faite en collationnant différentes sources d'information de la plus haute fiabilité (fichiers d'autorité, encyclopédies, dictionnaires, biographies, etc.). La relation entre un auteur particulier et les données le concernant s'est effectuée de façon manuelle jusqu'en 2010. Depuis 2011, la Fundación Ignacio Larramendi développe différents outils pour obtenir tout type d'information à partir de sources Linked Open Data. La communication expose de façon détaillée la procédure pour obtenir les URIs des ressources LOD en générant des recherches automatiques dans des serveurs SRU/OpenSearch et SPARQL Endpoint, ou en effectuant cet enrichissement sémantique à travers les fichiers disponibles en LOD. L'export des données est effectué en MARC 21, de même qu'en Europeana Data Model 5.2.1, SKOS et VIAF, après les avoir enrichies sémantiquement à partir de vocabulaires tels que LCSH, VIAF, GeoLinkedData (de l'Instituto Geográfico Nacional) et GeoNames.

## **1. Introduction : la Bibliothèque numérique de polygraphes**

Les réflexions qui suivent sur l'agrégation et la diffusion des notices d'autorité dans un environnement Linked Open Data ne proviennent pas d'une élucubration théorique mais ont pour base une pratique bibliothéconomique actuellement en production. Ces travaux catalographiques sont effectués par la Bibliothèque numérique de polygraphes (Biblioteca virtual de polígrafos)<sup>3</sup> qui fait partie des bibliothèques numériques de la Fundación Ignacio Larramendi<sup>4</sup>. L'objectif fondamental de la bibliothèque numérique des polygraphes est de donner une importance et une signification particulières à l'œuvre des penseurs espagnols, portugais, brésiliens et hispano-américains. C'est l'une des caractéristiques principales du projet bibliographique de les mettre en relation avec d'autres œuvres aux caractéristiques similaires, pour les rendre plus largement accessibles internationalement.

---

<sup>3</sup> <http://goo.gl/3kn00>

<sup>4</sup> Pour un état de la question de ce projet il est possible de consulter la communication d'Agénjo Bullón, Xavier et Hernández Carrascal, Francisca. *La Biblioteca virtual Larramendi: fuente de información bibliográfica para el pensamiento iberoamericano en la Web 3.0*. In: *Jornadas virtuales iberoamericanas de bibliotecología*. [<http://goo.gl/VbwtN>] Ces pages ont été rédigées à l'automne 2010, depuis lors quelques changements importants ont eu lieu. La plupart de ces changements sont exposés dans *La Biblioteca Virtual Ignacio Larramendi desde la perspectiva LOD y EDM* de Xavier y Hernández, Francisca, qui a été présenté au Ier Seminario internacional de la Biblioteca de Galicia [<http://goo.gl/uFrXx>].

Dès le début, nous voulons souligner, même si cela paraît évident, l'importance de disposer d'un projet bibliographique préalable, aspect qui doit prévaloir sur l'aspect technologique, particulièrement pour un sujet aussi nouveau que le Linked Open Data. Précisément, ce sont ces objectifs bibliographiques qui marquent les spécifications fonctionnelles aussi bien de la pratique bibliographique que du développement de la Bibliothèque numérique des polygraphes. L'objectif du projet est de réunir et de mettre en relation l'information bibliographique et l'information tout court, sur la pensée ibéroaméricaine dans une perspective historique, en établissant également des relations avec d'autres « civilisations », pour utiliser un terme de Toynbee<sup>5</sup>.

Les sources d'information pour les notices d'autorité

Il est évident qu'en grande partie les sources d'information qui sont utiles pour les objectifs du projet bibliographique ne sont pas disponibles sur support numérique, ou, quand elles le sont, elles ne sont pas structurées de façon appropriée. Dans d'autres cas, lorsqu'elles sont disponibles sur le WEB, la conception de leur interface leur ôte toute capacité et rendement informatifs. Par exemple, le *Diccionario biográfico español de la Real academia de la historia* procure une information sur plusieurs milliers de personnages importants de l'histoire et de la culture hispanique, entendue au sens le plus large, mais sous une forme technologiquement très pauvre. Une autre source d'information fondamentale est *Hombres y documentos de la filosofía española* de Gonzalo Díaz y Díaz, publié en 7 volumes de 1980 à 2003, qui pêche par un excès d'érudition et une faiblesse dans la démarche heuristique, mais qui facilite l'accès à une quantité énorme d'information<sup>6</sup>. On pourrait dire la même chose du magnifique *Diccionario de filosofía* de Ferrater Mora<sup>7</sup>, beaucoup plus rigoureux du point de vue intellectuel, bien que plus parcimonieux en ce qui concerne l'information proprement dite.

Il est fort dommage que de magnifiques répertoires bibliographiques, publiés en leur temps sur papier et toujours protégés par les droits d'auteur, soient inaccessibles sur le WEB et en particulier avec les technologies telles que le Linked Open Data. On se trouve devant le paradoxe suivant : les grands projets de numérisation rétrospective s'arrêtent, selon les pays, à 50, 70 et même 80 ans après le décès de l'auteur ; il est ainsi possible d'accéder facilement sur le WEB à des contenus périmés, tandis que d'autres de beaucoup plus de valeur et à jour ne sont pas disponibles. Quand on parle des problèmes du libre accès (« *Open Access* »), on a coutume d'oublier un problème élémentaire, il ne s'agit pas de l'accessibilité ou non de

---

<sup>5</sup> Comme on s'en souvient, Toynbee dans *A Study of History* permettait d'identifier jusqu'à 21 civilisations différentes et même si cette information peut paraître dépassée, y compris l'application du concept de civilisation par Huntington dans son célèbre livre (et pour ce qui nous concerne dans le moins célèbre ouvrage qui vint plus tard), ce qui est certain, c'est qu'il semble y avoir des compartiments étanches entre les cultures que sans aucun doute l'échange d'information - spécialement envisagé comme le fait Tim Berners-Lee dans *Design Issues: Linked Data* [<http://www.w3.org/DesignIssues/LinkedData>] - devrait lever.

<sup>6</sup> Peut être consulté dans la *Biblioteca Saavedra Fajardo de Pensamiento político hispánico* [<http://goo.gl/NHiVW>]. La transcription ne semble pas intégrale.

<sup>7</sup> Ferrater, Mora J. *Diccionario de filosofía*. Madrid: Alianza, 1979. Il s'agit de l'édition définitive de Ferrater Mora ; postérieurement, une édition augmentée a été publiée par la Chaire Ferrater Mora, sous la direction du professeur Josep Maria Terricabras qui – fort heureusement – signale par un astérisque les ajouts.

l'information numérique, mais simplement que cette information n'existe pas sous forme numérique.

Cette situation, qui ne doit pas être perdue de vue dans un congrès de bibliothécaires chez qui la bibliographie doit être, plus que tout, une discipline fondamentale, nous fait oublier que, fréquemment, les meilleures éditions critiques qui collationnent les manuscrits et les variantes, qui comportent des études préliminaires, etc. ne sont pas accessibles sur la toile. Il existe, certes, de magnifiques projets, de plus en plus fréquents, pour lesquels un travail critique d'édition numérique est fait de façon significative, mais ils sont minoritaires. Du fait de sa qualité extraordinaire nous allons citer le projet *Mark Twain Papers & Project*<sup>8</sup> pour son exemplarité en tant qu'édition numérique d'un auteur de premier ordre.

Heureusement, pour la Bibliothèque numérique des polygraphes, on a essayé de recourir au moins à un spécialiste dans le but de dresser un état des lieux des études bibliographiques sur chaque auteur (polygraphe) en particulier. De plus, en général, on a établi un minimum de sources d'information qui, dans tous les cas, doivent être consultées : *Diccionario biográfico de la Real academia de la historia*<sup>9</sup>, le Fichier d'autorité de la Bibliothèque nationale<sup>10</sup>, *Virtual International Authority File*<sup>11</sup> (qui contient les notices d'autorité de la Bibliothèque nationale d'Espagne et de la Bibliothèque nationale du Portugal), *Library of Congress Authorities*<sup>12</sup> et les ouvrages imprimés déjà mentionnés *Hombres y documentos de la filosofía española* et le *Diccionario de filosofía*. Logiquement on utilise en plus d'autres sources d'information comme *Stanford Encyclopedia of Philosophy*<sup>13</sup>, *The Catholic Encyclopedia*<sup>14</sup>, *Enciclopedia católica*<sup>15</sup>, *Jewish virtual Library*<sup>16</sup>, *Islamic philosophy Online*<sup>17</sup>, *Proyecto Filosofía en español*<sup>18</sup>, *Biblioteca Miralle*<sup>19</sup>, *Biblioteca Saavedra y Fajardo*<sup>20</sup>, *Biblioteca virtual Miguel de Cervantes*<sup>21</sup>, *Proyecto Sarmiento*<sup>22</sup> parmi d'autres.

---

<sup>8</sup> <http://bancroft.berkeley.edu/MTP/>

<sup>9</sup> <http://www.rah.es/diccBiografico.htm>

<sup>10</sup> <http://catalogo.bne.es/uhtbin/authoritybrowse.cgi>

<sup>11</sup> <http://viaf.org/>

<sup>12</sup> <http://authorities.loc.gov/>

<sup>13</sup> <http://plato.stanford.edu/>

<sup>14</sup> <http://www.newadvent.org/cathen/>

<sup>15</sup> <http://ec.aciprensa.com/>

<sup>16</sup> <http://www.jewishvirtuallibrary.org/>

<sup>17</sup> <http://www.muslimphilosophy.com/>

<sup>18</sup> <http://www.filosofia.org>

<sup>19</sup> <http://www.bibliotecamiralles.org/escritores.html>

<sup>20</sup> <http://saavedrafajardo.um.es/Biblioteca/IndicesW.nsf/Inicio?OpenForm&m=2>

<sup>21</sup> [www.cervantesvirtual.com](http://www.cervantesvirtual.com)

<sup>22</sup> <http://www.proyectosarmiento.com.ar/>

## Des notices d'autorité aux notices bibliographiques

Parmi les sources générales on a inclus de façon préférentielle VIAF, qui permet de localiser les auteurs dans une série de sources d'information faisant autorité comme la Bibliothèque du Congrès, la British Library ou la Deutsche Bibliothek. L'un des grands avantages de VIAF c'est la possibilité de naviguer à partir du regroupement des notices d'autorité vers les différentes vedettes constitutives, provenant des diverses sources qui constituent VIAF. Cela permet de choisir prioritairement les vedettes en fonction de leur provenance, la Biblioteca Nacional de España pour les auteurs espagnols, la Biblioteca Nacional de Portugal pour les auteurs portugais, celle du Brésil<sup>23</sup> pour les Brésiliens et ainsi de suite. Et, enfin, comme pivot commun les notices provenant de la Bibliothèque du Congrès. L'essentiel est qu'à partir d'une recherche on puisse établir une nouvelle et très importante fonctionnalité à caractère universel, c'est-à-dire la formation d'un agrégat virtuel de toutes les notices bibliographiques liées aux notices d'autorité. Par conséquent on localise les noms d'un auteur et ses variantes, ce qui permet de réaliser un enrichissement très complet grâce aux notices bibliographiques associées, établies dans la majorité des cas par les agences bibliographiques nationales.

VIAF permet actuellement de se connecter à près d'une vingtaine de fichiers d'autorité, à travers lesquels on peut accéder à un nombre élevé de catalogues bibliographiques. Il est très important de signaler que seulement dans quelques cas les zones 100 des autorités sont liées aux zones 6XX des descriptions bibliographiques et que si cette fonctionnalité était toujours présente il serait possible d'accéder à toutes les œuvres sur un auteur<sup>24</sup>, en utilisant VIAF comme concentrateur ou *hub* des ressources bibliographiques non seulement d'un auteur, mais aussi sur les œuvres de ce même auteur.

## Les sources d'information pour les descriptions bibliographiques

De la même façon que nous avons défini une série de sources pour les auteurs, nous avons établi le même mécanisme pour les compilations bibliographiques et nous avons ainsi choisi le *Catálogo colectivo de patrimonio bibliográfico*<sup>25</sup>, *Heritage of the Printed Book Database* (HPB) du Consortium des bibliothèques européennes de recherche (CERL, Consortium of European Research Libraries)<sup>26</sup>, le *Novum Regestrum*<sup>27</sup> et *WorldCat*. La *Biblioteca virtual Menéndez Pelayo*<sup>28</sup> mérite une mention particulière étant donné que la Bibliothèque numérique des polygraphes poursuit le travail bibliographique et le recensement de l'œuvre des principaux penseurs espagnols et de leur influence, qui constitue le sujet de *La ciencia española*<sup>29</sup> par laquelle Menéndez Pelayo répondait à la fameuse question de Monsieur

---

<sup>23</sup> Ne faisant pas encore partie de VIAF

<sup>24</sup> L'accès à travers VIAF des Archives et bibliothèques du Canada est particulièrement intéressant [<http://www.collectionscanada.gc.ca/>], même si dans les objectifs de la Bibliothèque numérique de polygraphes il n'a pas été possible de trouver, au moins pour le moment, quelque auteur qui fasse partie du projet, mais il serait possible de le faire si la méthodologie s'étendait.

<sup>25</sup> <http://www.mcu.es/bibliotecas/MC/CCPB/index.html>

<sup>26</sup> Heritage of the Printed Book Database (HPB) del CERL <http://www.cerl.org/web/en/resources/hpb/main>

<sup>27</sup> <http://goo.gl/fqtHA>. Voir aussi *Novum regestrum: el Catálogo colectivo del patrimonio bibliográfico iberoamericano* de Xavier Agenjo Bullón, Francisca Hernández Carrascal, "Boletín de la ANABAD", ISSN 0210-4164, Tome 44, N° 4, 1994, p. 127-142 <http://dialnet.unirioja.es/servlet/articulo?codigo=50938>

<sup>28</sup> <http://www.larramendi.es/i18n/bvmpelayo/inicio.cmd>

<sup>29</sup> Menéndez Pelayo, Marcelino. *La ciencia española*. In : *Biblioteca virtual Menéndez Pelayo*. <http://goo.gl/iB5m>

Masson<sup>30</sup> : *Que doit-on à l'Espagne? Et depuis deux siècles, depuis quatre, depuis dix, qu'a-t-elle fait pour l'Europe?*.

En outre, il convient de mentionner deux projets de grande portée qui sont en cours de réalisation simultanément, bien qu'ayant des dimensions différentes, et qui touchent le développement de la Bibliothèque numérique des polygraphes, non seulement quant au contenu mais aussi en ce qui concerne les spécifications fonctionnelles. Ces projets sont Hispana<sup>31</sup> et Europeana<sup>32</sup>. En ce qui concerne Hispana, il n'y a pas beaucoup de publications qui traitent de sa création et de son évolution, ce qui est surprenant étant donné les caractéristiques et les dimensions du projet, mais, par contre, il en existe de nombreuses présentations<sup>33</sup>. A l'heure où nous rédigeons ces pages [30 avril 2011], Hispana rassemble 3 181 786 objets numériques provenant de 146 réservoirs espagnols. Sur Europeana il y a, évidemment, beaucoup de documents, mais en ce qui concerne le propos de cette communication il convient de signaler le nouveau modèle de données *Definition of the Europeana Data Model Elements*, version 5.2.1<sup>34</sup>, mis à jour le 7 mars 2011 et *Functional specification for the Europeana Danube release*, publié le 31 août 2010.<sup>35</sup>

Ce n'est pas par hasard si les développements qui sont effectués pour la Bibliothèque numérique des polygraphes et pour la Fundación Ignacio Larramendi sont si étroitement en lien avec les spécifications fonctionnelles d'Europeana. La raison est la suivante : de façon complètement majoritaire l'Espagne participe à Europeana à travers Hispana (avec 1 367 808 objets numériques, bien loin devant la Biblioteca virtual Cervantés<sup>36</sup> qui en apporte seulement 19 062, bien que ce second projet fasse l'objet de beaucoup plus de publicité). Il se trouve qu'aussi bien Hispana que la majorité des fournisseurs de données collectées par Hispana<sup>37</sup> utilisent le SIGB DIGIBIB, actuellement dans sa version 6.0, mais qui en sera à sa version 7.0 au moment où sera faite cette communication à Porto Rico.

---

<sup>30</sup> Nicolas Masson de Morvilliers (1740-1789)

<sup>31</sup> <http://hispana.mcu.es>

<sup>32</sup> <http://europeana.eu>

<sup>33</sup> *Hispana y las iniciativas del Ministerio de Cultura* María Antonio Carrato Mena *Jornada de difusión de Europeana Local*, 17 nov. 2010 [<http://hdl.handle.net/10421/4765>]. *La aplicación del modelo de datos de Europeana a la Biblioteca virtual de patrimonio bibliográfico : bvpb.mcu.es*, María Luisa Martínez-Conde *Jornada de difusión de Europeana Local*. 17 nov 2010 URI: <http://hdl.handle.net/10421/4783>. La plus récente est Carrato, María Antonia. *Hispana. I Seminario internacional de la Biblioteca de Galicia* [<http://goo.gl/6xIqq>]

<sup>34</sup> <http://goo.gl/ojLL>

<sup>35</sup> *Functional specification for the Europeana Danube release*,

<sup>36</sup> <http://www.cervantesvirtual.com/>

<sup>37</sup> Biblioteca Virtual de Prensa Histórica, Galiciana: Biblioteca Digital de Galicia, Biblioteca Virtual de Andalucía, Gredos (Universidad de Salamanca, Espagne), Biblioteca Digital de Madrid, Biblioteca Digital de Castilla-La Mancha, Centre de documentacion de la Fundación MAPFRE, Biblioteca Digital de Castilla y León, Biblioteca Virtual del Patrimonio Bibliográfico, Biblioteca Virtual de Derecho Aragonés, Biblioteca Digital Real Academia de la Historia, Catalogue Collectif du réseau des Bibliothèques des Archives d'Etat, Biblioteca Virtual del Principado de Asturias, Archivo de la Imagen de Castilla La Mancha, Biblioteca Valenciana Digital, Biblioteca Virtual de Aragón, Fundación Sancho el Sabio, Biblioteca Regional de Murcia, Biblioteca Digital de Aranjuez, Universidad de La Laguna, Biblioteca Virtual de la Diputación de Zaragoza, Biblioteca virtual de La Rioja, Fundación Ignacio Larramendi, Biblioteca Virtual de la Real Academia nacional de Farmacia.

Il n'est pas non plus possible de passer sous silence le fait que la société DIGIBÍS<sup>38</sup>, qui se charge des développements du logiciel de gestion bibliothéconomique DIGIBIB et archivistique DIGIARCH, comme de l'entrepôt OAI appelé OasIs, est une société propriété de la Fundación Ignacio Larramendi, qui utilise systématiquement la Bibliothèque numérique des polygraphes comme banc d'essai pour ses développements. C'est pourquoi (et c'est un souhait des clients) beaucoup de fonctionnalités décrites ici toucheront de façon importante une trentaine d'importantes bibliothèques espagnoles et plus d'une centaine, si on prend en compte celles qui reprennent globalement la Biblioteca virtual de patrimonio bibliográfico<sup>39</sup> et la Biblioteca virtual de prensa histórica.

Quand cette communication sera présentée à Porto Rico, le volet *Danube* du projet Europeana sera déjà entré en phase de production, prévue pour le 1<sup>er</sup> mai 2011. Naturellement nous ne pouvons être sûrs qu'il en sera ainsi en définitive, mais de toute façon, le fait que survienne un retard est complètement indépendant de ce qui a été planifié pour la Bibliothèque numérique des polygraphes, étant donné que ce qui est véritablement décisif a été le choix effectué par Europeana de développer de nouvelles fonctionnalités en ce qui concerne la structure de l'information sémantique. Du fait de ces caractéristiques, aussi bien Europeana qu'Hispana constituent deux sources d'information primordiales pour la Bibliothèque numérique des polygraphes et deux exemples méthodologiques qui, en ce qui concerne Hispana, en plus s'alimentent réciproquement.

Pour le moment, Europeana a livré une API reposant sur OpenSearch<sup>40</sup> qui permet, grâce à son intégration sur un site Web, la recherche et l'affichage des notices d'Europeana. Cette API, qui fonctionne très efficacement et a été intégrée à l'interface WEB de la Bibliothèque numérique des polygraphes, lance une consultation à partir d'un terme de recherche initial dans la base de données Europeana. C'est-à-dire que la même consultation effectuée dans une base de données de plusieurs milliers de notices se fait aussi dans la base de données Europeana qui compte près de vingt millions d'objets numériques (17 901 019).

## 2. Agrégats numériques

Selon la méthodologie établie pour la Bibliothèque numérique des polygraphes, le noyau sur lequel tourne tout le système d'information est constitué par ce qu'on appelle des "agrégats numériques" qui se rapprochent beaucoup des traditionnelles entrées encyclopédiques, y compris dans leur forme typographique, et prétendent même atteindre le niveau de complexité et de complétude de certaines entrées de Wikipedia (ou DBpedia, comme on le verra plus loin).

Le terme "agrégation" ou "agrégat" a différents usages selon l'environnement dans lequel il s'applique et peut générer chez le lecteur une certaine incertitude. Par exemple, dans les procédures de collecte des métadonnées, les collecteurs agrègent des métadonnées et peuvent se dénommer agrégateurs, comme c'est le cas dans Europeana. Cependant, le concept d'agrégat numérique qui est utilisé dans la Bibliothèque numérique des polygraphes correspond davantage à la définition de l'*Open Archives Initiative Object Reuse and*

---

<sup>38</sup> <http://www.digibis.com>

<sup>39</sup> On peut voir en particulier dans la *Biblioteca virtual de patrimonio bibliográfico* le sous domaine *Iberoamérica* dans les collections de la BVPB [<http://goo.gl/8Jyfa>]. Il peut aussi être utile de faire une recherche plein texte dans la Biblioteca virtual de prensa histórica [<http://prensahistorica.mcu.es>].

<sup>40</sup> <http://www.version1.europeana.eu/web/api>

*Exchange* (OAI-ORE)<sup>41</sup>, selon laquelle un agrégat constitue une ressource en soi, constituée par un ensemble d'autres ressources reliées entre elles. Un agrégat numérique serait donc l'ensemble des données et des ressources informatives réunies autour d'un auteur particulier, dont le noyau d'information ou noeud principal est formé par une notice d'autorité Nom de personne.

La procédure de contrôle des autorités qu'effectue la Fundación Ignacio Larramendi est beaucoup plus large que celle qui est menée traditionnellement dans les grandes bibliothèques ou agences bibliographiques. Nous n'insisterons pas sur la fonction d'identification des entités *personnes* mais sur le travail de contextualisation, tel qu'il est défini par FRAD<sup>42</sup>. Non seulement on collecte des données qui ne figurent pas habituellement dans les fichiers d'autorité, mais aussi on identifie et met en relation les notices d'autorité avec d'autres sources, des ressources et des données disponibles sur la toile. Dans ce sens, l'indication des sources d'information d'où proviennent les données (670) est non seulement une justification pour le choix du nom et de ses variantes, mais encore une indication des sources d'information des ressources bibliographiques choisies en ce qui concerne un auteur particulier.

De façon similaire à celle utilisée pour VIAF, chaque notice d'autorité reste définie, selon le format MARC 21, par sa vedette principale (1XX), les variantes du nom (4XX), les relations avec les autres noms (5XX) et les équivalences dans d'autres langues des vedettes principales (7XX). Ce dernier point est d'une très grande importance pour la Bibliothèque numérique des polygraphes étant donné que tous les auteurs sont choisis non seulement pour leur importance dans la culture ibéroaméricaine, mais aussi pour l'apport de la culture ibéroaméricaine à la culture universelle. Pour cette raison, il est absolument nécessaire de disposer des équivalences dans d'autres langues afin de suivre la voie de transmission des textes au niveau mondial, de même que la transmission et la répercussion des idées. La même situation existe en Espagne où coexistent le castillan ou espagnol avec d'autres langues officielles comme le catalan, le valencien, le galicien et le basque.

### Vedettes multilingues

Nous souhaitons attirer l'attention sur la nécessité pour le format MARC 21 de disposer d'un élément permettant de coder la langue des vedettes. Malheureusement, le *Discussion Paper 2001-DP05 Multilingual Authority Records in the MARC 21 Authority Format*<sup>43</sup> ne semble pas avoir abouti à quelque proposition, du moins pour le moment. Les zones 7XX permettent de faire figurer les vedettes équivalentes dans d'autres langues, mais elles ne comportent aucun élément permettant d'indiquer la langue de cette vedette, ce qui est particulièrement nécessaire dans des fichiers d'autorité multilingues. D'autres mécanismes présentent aussi des limites ; la position 008 /08 permet seulement de coder l'anglais et/ou le français comme langue de catalogage et le 040 \$b permet seulement d'établir la langue de catalogage qui dans de nombreux cas est différente de celle de la vedette.

Cette situation a trouvé une solution raisonnable dans MADS et en tout cas dans la proposition MADS.RDF<sup>44</sup>, mais il ne faut pas oublier que la majorité des notices d'autorité existent seulement en format MARC et que par conséquent il sera difficile de transférer

<sup>41</sup> <http://www.openarchives.org/ore/>

<sup>42</sup> [http://www.ifla.org/files/cataloguing/frad/frad\\_2009-es.pdf](http://www.ifla.org/files/cataloguing/frad/frad_2009-es.pdf)

<sup>43</sup> <http://www.loc.gov/marc/marbi/2001/2001-dp05.html>

<sup>44</sup> /RDF Primer. Final Public Review Document. [<http://www.loc.gov/standards/mads/rdf>]

rapidement ce type de données dans un environnement Linked Open Data. Par exemple, VIAF a hérité de ce problème et tandis qu'il est possible d'identifier l'établissement et le pays créateur d'une notice, il n'est pas possible de savoir, sauf par déduction, dans quelle langue est la vedette. Cela est particulièrement important dans les cas de vedettes établies dans une même aire linguistique qui peuvent être très similaires ou impossibles à distinguer (par ex. Geografía en espagnol et Geografia [sans tilde] en italien ou catalan).

#### Variantes pour les noms de personne

La recherche d'information dans d'autres fichiers d'autorité a non seulement pour but d'établir des relations d'équivalence linguistique entre les vedettes, mais aussi de faire figurer la plus grande quantité possible de variantes des noms ainsi que leurs relations avec d'autres noms. Cette procédure, adaptée au comportement du SIGB DIGIBIB, permet qu'en important des notices bibliographiques, celles-ci soient parfaitement reliées à la notice d'autorité correspondante et qu'il y ait, de plus, contrôle de la génération d'entrées d'autorité non valides à partir de la zone 4XX. Cette procédure est lente en soi au début, au moment de la création de la notice d'autorité, mais est beaucoup plus productive à moyen terme et permet une plus grande qualité. En outre, il convient de rappeler que le projet bibliographique original se met en œuvre de bien meilleure façon si on analyse les sources en partant du noyau central constitué par les notices d'autorité.

Cette procédure est particulièrement importante si nous considérons qu'il s'agit d'un ensemble d'auteurs qui vont de Sénèque (Cordoue 4 av. J.-C.- Rome 65 apr. J.-C.) à Martín de Riquer (Barcelona, 1914-) et que beaucoup d'entre eux ont utilisé différents noms en fonction de la langue dans laquelle ils ont écrit, sans compter que ces noms ont été traduits dans différentes langues avec leurs œuvres (par ex. Averroès est le nom latin de Ibn Rus). Disposer, par conséquent, d'une série de variantes de noms, correspondant en outre aux ressources qu'on désire importer, permet à chaque intégration d'une ressource de limiter les modifications des notices d'autorité dans la mesure où y figurent déjà un nombre important de ces variantes.

### 3. Notices d'autorité

La mise à jour n° 12 du format MARC 21, publiée en octobre 2010, consolide, selon la philosophie de RDA (*Resource Description and Access* = Ressources : Description et accès), une série de zones qui sont essentielles pour l'agrégation de l'information sur les personnes. Nous nous référons, comme c'est évident, aux zones 3XX. D'un point de vue historique<sup>45</sup>, les éléments qui dans un premier temps avaient pour objectif principal de différencier les homonymes, ont acquis peu à peu une valeur en eux-mêmes. Par exemple, les différentes sous-zones de la zone 100 qu'on demandait d'utiliser seulement pour différencier les auteurs portant le même nom, sont devenues obligatoires, en particulier le \$d, car il permettait de contextualiser un auteur immédiatement d'un point de vue chronologique, ce qui aidait dans de nombreux cas à le mettre en relation avec l'œuvre en évitant des erreurs possibles. Jusqu'à l'adoption de RDA par MARC 21<sup>46</sup> il n'y avait pas d'éléments spécifiques pour les attributs des personnes aussi significatifs que le lieu de naissance. Justement ces attributs, en référence non seulement au nom de la personne mais à sa personne, contribuent à donner aux notices

---

<sup>45</sup> Agenjo Bullón, Xavier y Palá Gasós, Pilar. *El fichero de autoridades del Catálogo colectivo del patrimonio bibliográfico*. In: *Boletín de la ANABAD*, ISSN 0210-4164, Tome 37, N° 4, 1987, p. 593-606.

[<http://dialnet.unirioja.es/servlet/articulo?codigo=801041>]

<sup>46</sup> <http://www.loc.gov/marc/RDAinMARC29.html>

MARC un aspect essentiellement encyclopédique, presque « dbpédique », si vous nous permettez le néologisme.

Pour les objectifs de la Bibliothèque numérique des polygraphes, la publication de RDA a supposé un renforcement de ses lignes de travail. En premier lieu, en ce qui concerne la récupération d'information sur les auteurs, selon les tâches effectuées par les utilisateurs définies dans FRAD comme contextualisation et recherche. En second lieu, dans la façon de structurer les ressources d'information selon le modèle Œuvre-Expression-Manifestation-Item<sup>47</sup>. Pour cette raison on a aussitôt appliqué les mises à jour successives du format MARC 21 à tel point que le logiciel DIGIBIB a ajusté le rythme de ses mises à jour à la publication des mises à jour de MARC 21.

Pour chaque auteur, dans la mesure du possible, on donne, en plus de la forme normalisée du nom, les variantes et équivalences dans d'autres langues, sa biographie (678), les dates de naissance et de mort (046), les lieux de naissance et de mort (370) et de toute autre activité (370), ses activités (372 et 374), la langue utilisée (377) ou le genre (375). Ce sont des attributs de l'entité *personne* qui ont pour but ultime d'étayer la recherche et la navigation sémantiques entre les données. Ces attributs permettent au site WEB de la Bibliothèque numérique des polygraphes d'offrir la possibilité de rechercher les auteurs par les dates de leurs activités, par les lieux associés à leur biographie, par les professions et activités, et non pas seulement par leurs œuvres, comme c'est le cas fréquemment dans les OPAC WEB.

Notices d'autorité dans d'autres schémas de métadonnées

Pour la Bibliothèque numérique des polygraphes l'information concernant les collectivités, personnes et familles provenant des archives est d'un intérêt comparable. Le besoin de convergence entre les données des archives, des bibliothèques et des musées est maintenant reconnue et avec des exemples d'une qualité remarquable. Il est maintenant fréquent qu'on se pose la question d'unifier ce type de données par une interface unique de consultation, ce qui contribue à mettre en valeur la capacité informative de ces établissements. Très récemment dans DIGIBÍ a été mené un travail présentant ces caractéristiques pour le projet de la Junta de Castilla-La Mancha consacré à l'Archivo de la imagen (Archives de l'image, photographies)<sup>48</sup> et à l'Archivo de la palabra (Archives de la parole, archives sonores)<sup>49</sup>. Pour cela on a réalisé un alignement (« *mapping* ») systématique<sup>50</sup> de MARC 21 vers EAD et EAC-CPF, de même que vers ISAD(G) et ISAAR (CPF).

De même, et comme cela était souhaitable, dans la Bibliothèque numérique des polygraphes on a pointé la nécessité de relier les matériaux bibliographiques et archivistiques en relation : c'est le cas par exemple pour Agustín de Montiano y Luyando (1697-1794), historien, critique et dramaturge espagnol - qui est passé dans la mémoire collective pour avoir été le

---

<sup>47</sup> MARC Discussion Paper No. 2011-DP03: Identifying Work, Expression, and Manifestation records in the MARC 21 Bibliographic, Authority, and Holdings Formats [<http://www.loc.gov/marc/marbi/2011/2011-dp03.html>]

<sup>48</sup> [http://clip.jccm.es/archivo\\_de\\_la\\_imagen/es/micrositios/inicio.cmd](http://clip.jccm.es/archivo_de_la_imagen/es/micrositios/inicio.cmd)

<sup>49</sup> [http://clip.jccm.es/archivo\\_de\\_la\\_palabra/es/micrositios/inicio.cmd](http://clip.jccm.es/archivo_de_la_palabra/es/micrositios/inicio.cmd)

<sup>50</sup> [http://www.sedic.es/p\\_boletin\\_novedades\\_mensual.asp#Digibis](http://www.sedic.es/p_boletin_novedades_mensual.asp#Digibis)

premier directeur de la Real academia de la historia, fondée en 1735 – étant donné qu'on conserve ses archives qui font partie des Bibliotecas virtuales Larramendi<sup>51</sup>.

#### 4. Lien entre les données

La Bibliothèque numérique des polygraphes participe à Europeana, à travers Hispana comme fournisseur de données et, comme on l'a dit, elle utilise sa base de données comme source d'information. Pour cela, elle s'est assigné comme objectif stratégique l'adoption du modèle proposé par Europeana Data Model (EDM)<sup>52</sup> et les technologies Linked Open Data. Aussi bien EDM que Linked Open Data sont en train d'ouvrir à toute vitesse le domaine d'activité et d'offrir d'énormes quantités d'information comme c'est le cas de VIAF<sup>53</sup> ou de DBpedia<sup>54</sup>. Depuis la proposition de Tim Berners-Lee sur le Linked Data (ou Web de données) en 2006 jusqu'à maintenant, on a mis à disposition des données sémantiquement structurées et réutilisables à un rythme jamais vu précédemment sur le WEB, exception faite probablement des notices bibliographiques en MARC 21.

Pour tout ce qui est relatif à l'apport bibliothéconomique au LOD, il convient de rappeler qu'au sein du W3C a été créé en 2010 le Library Linked Data Incubator Group<sup>55</sup> (LLD). Ce groupe a pour but de préparer divers rapports sur l'état de la question des ontologies pour les archives, les bibliothèques et les musées, sur les vocabulaires et ensembles de données intéressants pour ces domaines et l'usage que la communauté des bibliothèques en fait. Toute personne intéressée pourra se tenir informée en suivant l'activité de ce groupe, comme du Linking Open Data<sup>56</sup>.

Pour la Bibliothèque numérique des polygraphes, le Linked Open Data (LOD) et les technologies annexes présentent l'opportunité d'effectuer pleinement l'établissement des « agrégats numériques » comme nœuds de liens de ressources d'information. De fait, la Bibliothèque numérique des polygraphes est un cas d'usage analysé par LLD<sup>57</sup>. Il s'agit maintenant non seulement de capturer de l'information mais aussi de la lier à des nœuds appropriés du nuage LOD, selon des règles du jeu où il est souhaitable que chacun joue son rôle. Chaque ressource disponible dans un ensemble de données LOD est identifié par un URI avec des caractéristiques spécifiques<sup>58</sup>. Cet URI draine à son tour tout un réseau de relations, attributs, avec d'autres ressources qui définissent, selon le modèle qui est utilisé (ontologie ou schéma RDF), les entités et propriétés d'un domaine fonctionnel déterminé. Mais même si cela est important, ce qui ajoute véritablement de la valeur, c'est que cela permet d'identifier qu'une entité dans son environnement est bien identique ou très similaire à une autre entité définie dans un autre environnement. Par exemple, Maimonide est un

---

<sup>51</sup> En fait, il fut un prédécesseur d'Ignacio de Larramendi y Montiano (1921-2001) initiateur, parmi beaucoup d'autres initiatives, des Bibliotecas virtuales Larramendi o de DIGIBÍS. Voir aussi, Hernando de Larramendi, Ignacio. *Mecenazgo cultural de Ignacio Hernando de Larramendi y Montiano : crónica y testimonios*. [<http://www.larramendi.es/i18n/consulta/registro.cmd?id=954>]

<sup>52</sup> Doerr, Martin, Gradmann, Stefan, Hennicke, Steffen, ISAAC, Antoine, Meghini, Carlo, SOMPLE, Herbert van de. *The Europeana Data Model (EDM)*. In: 76<sup>th</sup> IFLA General Conference and Assembly, 10-15 August, 2010, Gothenburg, Sweden. <http://goo.gl/7eLs>

<sup>53</sup> <http://ckan.net/package/viaf>

<sup>54</sup> <http://dbpedia.org/About>

<sup>55</sup> <http://www.w3.org/2005/Incubator/lld/>

<sup>56</sup> <http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

<sup>57</sup> [http://www.w3.org/2005/Incubator/lld/wiki/Use\\_Case\\_Polymath\\_Virtual\\_Library](http://www.w3.org/2005/Incubator/lld/wiki/Use_Case_Polymath_Virtual_Library)

<sup>58</sup> <http://www.w3.org/TR/2007/WD-cooluris-20071217>

*philosophe* pour DBpedia et *le nom d'une personne* dans VIAF. La procédure consiste à lier les données par l'intermédiaire de ces URI et à disposer d'une série d'outils<sup>59</sup> qui permettent de localiser ces données, de les réutiliser, de déduire des données à partir d'autres données en s'appuyant sur les relations et sur des règles d'inférence, de mettre à jour les données d'origine et de les proposer de nouveau modifiées et enrichies. Et cela de façon itérative.

## VIAF et DBpedia

En 2010 la Bibliothèque numérique des polygraphes a démarré le processus de lien entre les notices d'autorité et les ressources similaires existantes dans le LOD. Évidemment, les sources d'information signalées dans la méthodologie sont ici aussi les sources pour obtenir les liens. Par exemple, on a fait figurer dans la zone 024 des notices MARC 21 l'URI des entités de VIAF. La ressource en RDF pour Maimonide (`rdf:about="viaf/100185495"`) regroupe les vedettes établies pour cet auteur provenant de 7 fichiers d'autorité différents, avec leurs 255 variantes du nom, mais est aussi liée au contenu de la DBpedia (<http://dbpedia.org/resource/Maimonides>). Si le bibliothécaire a de l'expérience et la connaissance du domaine, c'est-à-dire est un catalogueur chevronné, il déduira du tissu de relations de VIAF jusqu'à 12 vedettes différentes, d'autant de pays et de langues, correspondant à l'entité Maimonide<sup>60</sup>.

Évidemment, le processus d'obtention de ces données de façon manuelle est très précis mais lent, et très vite on a envisagé la possibilité de l'automatiser, en évaluant les avantages et inconvénients à le faire de cette manière. Indubitablement, le nombre d'auteurs qui entrent dans le champ du projet bibliographique de la BVP, autour de 1 000, permettrait de le faire manuellement par contre la productivité en temps et en données serait bien supérieur en l'effectuant de façon automatisée. En outre, une fois envisagés et développés, les instruments nécessaires pourraient être intégrés comme fonctionnalités spécifiques du SIGB DIGIBIB. Nous espérons qu'au moment de la lecture de cette communication, 13-18 août 2011, ces instruments feront déjà partie d'un module additionnel de DIGIBIB.

## SKOS : une autre voie vers le Linked Open Data

Les mêmes liens ont été établis en utilisant SKOS. Le résultat visé est d'indiquer qu'un concept est identique ou très similaire à d'autres concepts dans d'autres vocabulaires. A travers les vedettes matière des notices bibliographiques, et le lien des concepts avec d'autres vocabulaires aussi disponibles dans SKOS, il est possible d'obtenir un nombre très important d'œuvres d'un auteur. En outre, il est possible d'automatiser le processus de recherche de ressources d'information, en parcourant le tissu des relations entre concepts des différents vocabulaires, et de récupérer en même temps les ressources d'information liées à ces concepts.

En ce moment, la Bibliothèque numérique des polygraphes est en train de travailler, à titre d'essai, sur la *Lista de encabezamientos de materia para bibliotecas públicas* (Liste des vedettes matière pour les bibliothèques publiques) qui a été établie, en son temps, par la Dirección general del libro y bibliotecas et qui a été enrichie postérieurement, en la croisant avec le fichier d'autorité de la Biblioteca Nacional<sup>61</sup> et son équivalent avec les vedettes

<sup>59</sup> <http://www.w3.org/wiki/SemanticWebTools>

<sup>60</sup> (<http://dbpedia.org/resource/Maimonides>). Si

<sup>61</sup> *Autoridades de la Biblioteca Nacional* [Recurso electrónico]/ Biblioteca Nacional ; software, Chadwyck Healey -- Número 1 (nov. 1996)- -- Madrid : Biblioteca Nacional : Chadwyck-healey España, 1996- CD-ROM.

matière de la Bibliothèque du Congrès<sup>62</sup>. Étant donné que ce processus est analogue à celui effectué pour lier les concepts SKOS de RAMEAU<sup>63</sup> avec ceux de LCSH et les Schlagwortnormdatei (SWD)<sup>64</sup> il sera possible d'établir ces relations conceptuelles qui, sans aucun doute, ne posséderont pas un degré de granularité très élevé, mais qui, traitées et croisées avec un nombre déterminé d'auteurs, pourront vraiment permettre une mise en relation claire d'ensembles très larges d'information.

## Temps et espace

Deux autres éléments jouent un rôle essentiel dans l'agrégation et la mise en relation des données concernant un auteur : ce sont les données chronologiques et géographiques. En effet, un auteur vit et meurt à des dates déterminées, il travaille, a des relations sociales avec ses contemporains durant des périodes déterminées, il écrit des livres qui sont publiés, traduits et édités à des dates très concrètes<sup>65</sup>. La possibilité de mettre en relation et de lier les auteurs par cette nouvelle catégorie chronologique fait que cet ensemble d'œuvres et d'auteurs liés par des noms et des matières, seront de plus considérablement identifiés aussi chronologiquement, même s'il est certain que cela rendra plus sûrement visible la dispersion ou l'incohérence. Le codage du temps, présent dans de nombreux éléments des notices d'autorité et des notices bibliographiques, nécessite souvent des qualificatifs ou des attributs pour pouvoir exprimer la probabilité, l'inexactitude, la fiabilité, etc. dans les descriptions des personnes et des ressources, sans parler des événements. Ce type de données est fondamental pour générer des lignes du temps ou pour la réalisation d'opérations et de déductions à partir des dates, usage qui peut être vu dans WorldCat Identities<sup>66</sup>.

De la même façon, on est en train de travailler sur l'agrégation et la mise en relation des données géographiques. De fait, il existe dans le Linked Open Data une importante initiative pour ce type de données, GeoNames<sup>67</sup>, qui, de plus, est le seul cas connu jusqu'à aujourd'hui qui compte une réplique en Espagne, à travers le processus de conversion des répertoires géographiques de l'Instituto geográfico nacional en RDF et en LOD<sup>68</sup>. Une nouvelle fois on a eu recours au processus d'agrégation et, ainsi, nos auteurs sont *nés* et sont *morts* dans un *lieu* déterminé, ils ont eu des relations avec leurs contemporains dans des *endroits* déterminés, peut-être dans des écoles, des universités, des ordres religieux, des armées, des sociétés scientifiques ou culturelles, des groupes politiques, situés ou en lien avec un *lieu* déterminé. Par conséquent, et à travers eux, on pourra établir de nouveau les agrégations et lien appropriés.

Dans le cas de GeoNames et de GeoLinkedData le processus a été effectué de façon automatique, en transformant les données sélectionnées dans une structure d'information

---

<sup>62</sup> <http://id.loc.gov>

<sup>63</sup> <http://www.cs.vu.nl/STITCH/rameau/>

<sup>64</sup> <http://www.d-nb.de/standardisierung/normdateien/swd.htm>

<sup>65</sup> Il existe maintenant un nouveau format qui va essayer d'homologuer les façons très différentes par lesquelles les diverses civilisations ont établi les chronologies, y compris les chiffres par lesquels elles les ont représentées, autrement que par l'abécédaire, l'alphabet, l'écriture arabe, etc. Extended Time/Date Format ETDF [<http://www.loc.gov/standards/datetime/>]

<sup>66</sup> <http://www.worldcat.org/identities/>

<sup>67</sup> <http://www.geonames.org/ontology/documentation.html>

<sup>68</sup> Cette initiative a démarré avec la publication de diverses sources d'information issues del Instituto geográfico nacional, en les rendant disponibles comme base de connaissance RDF conforme aux principes du Linked Data. <http://geo.linkeddata.es/web/guest;jsessionid=6E6A9C1E1DAF40F81005E8F4DA3A27E1>

importable dans DIGiBIB. Un alignement (« *mapping* ») de ces données en MARC a été effectué, avec l'objectif supplémentaire de remplir automatiquement les zones 752 et 151. De cette façon, on a obtenu non seulement les noms disponibles dans ces vocabulaires, mais aussi les coordonnées géographiques. Il est bien évident qu'il doit y avoir postérieurement un travail manuel pour résoudre les problèmes qui peuvent se présenter au moment de différencier entre juridictions et localisations, ainsi que pour la localisation géographique des juridictions anciennes. De la même façon, comme on le verra plus loin, une chose est de gérer des *personnes* et une autre les *noms de personne*. Il est évident que gérer des juridictions, des noms géographiques et des localisations géographiques oblige à disposer de structures de données différenciées.

## 5. Utilisation de données dans le LOD

Sur le site WEB de VIAF on peut faire manuellement des recherches concrètes et capturer les notices obtenues dans leur version MARC XML ou RDF. À travers l'API de VIAF<sup>69</sup>, on peut faire des recherches, à travers SRU et OpenSearch, et l'on obtient les résultats en MARC XML, RDF et dans une grande variété de formats et de schémas. De même VIAF<sup>70</sup> est disponible comme Linked Open Data. Selon les recommandations pour la publication des données<sup>71</sup> dans le LOD, les ensembles de données (« *datasets* ») disposent fréquemment d'un serveur SPARQL Endpoint<sup>72</sup> qui permet de réaliser des recherches en SPARQL<sup>73</sup> ou bien qui peuvent utiliser des clients SPARQL<sup>74</sup> qui permettent d'interroger des fichiers RDF. Ceci signifie que VIAF ou DBpedia<sup>75</sup> peuvent être interrogés via SPARQL et que, par conséquent, les questions qui peuvent être faites permettent de faire un balayage logique de ces données et de récupérer plus d'information en lien qu'avec les méthodes exposées plus haut. Au moyen de SPARQL il est possible de définir une recherche beaucoup plus complexe ; prenons un exemple possible dans DBpedia, mais pas dans VIAF, « philosophes influencés par Maimonide ». La Bibliothèque numérique des polygraphes est en train d'avancer, nous le répétons, selon un plan bibliographique déterminé pour profiter de l'information qui est versée sous forme d'ensembles de données (« *datasets* ») dans le LOD<sup>76</sup>. Et cela, elle le fait aussi bien du point de vue du consommateur de données que de celui du producteur de données.

Ainsi, le premier pas a été l'application de l'Europeana Data Model comme réponse à la transformation qu'Europeana a planifiée pour la phase *Danube*. Il est possible qu'Europeana

---

<sup>69</sup> <http://www.oclc.org/developer/documentation/virtual-international-authority-file-viaf/using-api>

<sup>70</sup> <http://ckan.net/package/viaf>

<sup>71</sup> *How to Publish Linked Data on the Web*. <http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/BestPracticeRecipesforPublishingRDFVocabularies>. W3C Working Group Note 28 August 2008. <http://www.w3.org/TR/swbp-vocab-pub/>

<sup>72</sup> <http://www.w3.org/wiki/SparglEndpoints>

<sup>73</sup> SPARQL Query Language for RDF. [<http://www.w3.org/TR/rdf-sparql-query/>]. Actuellement le W3C travaille en SPARQL 1.1 Federated Query. *This specification defines the syntax and semantics of a SPARQL 1.1 Federated Query extension for executing queries distributed over different SPARQL endpoints*. [<http://www.w3.org/2009/sparql/docs/fed/service>]

<sup>74</sup> SPARQL Query Language for RDF. [<http://www.w3.org/TR/rdf-sparql-query/>]. En la actualidad el W3C está trabajando en SPARQL 1.1 Federated Query. *This specification defines the syntax and semantics of a SPARQL 1.1 Federated Query extension for executing queries distributed over different SPARQL endpoints*. [<http://www.w3.org/2009/sparql/docs/fed/service>]

<sup>75</sup> <http://ckan.net/package/dbpedia>

<sup>76</sup> <http://ckan.net/package>. Voir aussi le recensement effectué par la LLD <http://ckan.net/group/lld> ou DOIs as Linked Data <http://inkdroid.org/journal/2011/04/25/doi-as-linked-data/> ou Linked Periodicals Data <http://periodicals.dataincubator.org/html>

puisse réaliser des traitements plus ou moins automatisés pour transformer l'information reçue des fournisseurs dans le nouveau modèle de données EDM, mais il faut considérer que ceux-ci sont déjà plusieurs milliers et dans plus de 25 langues, ce qui laisse supposer que les seuls traitements automatiques pourront toucher un pourcentage indéterminé de données, avec des résultats très certainement impossibles à vérifier, et avec un grand risque d'établir des équivalences erronées. Indépendamment du fait qu'une procédure de ce type puisse être pratique et résoudre l'établissement de liens entre une partie des données, elle ne peut s'étendre à la totalité, ni ne peut être améliorée sans l'intervention des fournisseurs de données eux-mêmes. Prenons un cas typique, les noms géographiques sont très souvent des noms de juridictions qui ne correspondent pas aux coordonnées géographiques actuelles, ainsi, faire l'équivalence entre la Castille du XVe siècle et la Castille actuelle constitue seulement une approximation. Par conséquent, la Bibliothèque numérique des polygraphes a lancé un processus de publication de ses données liées de façon interne en LOD.

Convertir les notices bibliographiques MARC 21 en Europeana Semantic Elements 3.3.1 et ensuite en Europeana Data Model présente une difficulté relative, mais si on réalise uniquement ce processus, on offrira des données à Europeana sans faire le lien avec les sources externes ou les vocabulaires communs avec lesquels Europeana pourrait réaliser une mise en relation exacte des données. La méthodologie qui consiste à incorporer les URI de VIAF dans les notices d'autorité s'est révélée fort utile étant donné qu'à la génération d'EDM on peut proposer une référence à un vocabulaire externe pour la classe *Agent*. Le second processus de mise en relation s'est effectué à travers les LCSH disponibles dans *id.loc.gov*. De cette façon, la Bibliothèque numérique des polygraphes permet de publier les vedettes matière en SKOS, mais liées à LCSH. Ainsi, la propriété *dc:subject* des ressources de la Bibliothèque numérique des polygraphes sera en relation avec *skos:concept*, mais liée à LCSH. Cela permet indirectement la mise en relation avec RAMEAU<sup>77</sup> et SWD<sup>78</sup> au minimum. Autrement dit, les vedettes matière sont accessibles dans une structure sémantique et associées à d'autres vedettes matières équivalentes en anglais, français et allemand. L'avantage de la mise en relation des différents ensembles de données (« *datasets* ») avec un ou plusieurs vocabulaires particuliers est évident.

En plus de cela, ce qui ne manque pas d'être paradoxal, c'est que cette procédure elle-même peut être valable pour l'interconnexion des données au niveau national. L'une des conséquences d'avoir considéré LCSH comme source principale d'information pour l'établissement des vedettes matière dans différentes bibliothèques espagnoles, c'est qu'à travers cette relation il est possible de réaliser le lien entre les listes de vedettes matière espagnoles qui, cela ne doit pas être oublié, peuvent être présentes dans différentes langues. Cela est un effet collatéral du LOD assez bénéfique et, même si c'est par un détour, cela permettra d'interconnecter et de lier les données des catalogues bibliographiques et les notices d'autorité espagnoles, qui sont restées assez isolées et imperméables, beaucoup plus qu'il n'était souhaitable et compréhensible.

Nous devons aussi signaler ici un autre effet bénéfique d'Europeana en ce qui concerne l'amélioration des systèmes d'information d'un pays. Le Ministère espagnol de la culture qui est responsable d'Hispana a aussi mis en route le processus d'adaptation à EDM et, comme on l'a dit, l'une des premières activités a été de convertir la Liste des vedettes matière pour les bibliothèques publiques (LEM) en SKOS, en liant les entrées possibles avec LCSH.

---

<sup>77</sup> <http://ckan.net/package/stitch-rameau>

<sup>78</sup> <http://ckan.net/package/dnb-gemeinsame-normdatei>

Logiquement, la Bibliothèque numérique des polygraphes complétera le processus de mise en relation avec les vedettes matière en galicien, catalan et basque. A la date du 30 avril 2011, la Biblioteca virtual del patrimonio bibliografico, implémentée sur le SIGB DIGIBIB 7.0 permet déjà l'export de notices conforme à EDM 5.2.1.

Ces processus convergeront aussi bien dans l'application complète d'EDM comme dans la publication des ensembles de données (« *datasets* ») de la Bibliothèque numérique des polygraphes en LOD. Dans ce sens, la publication des données en LOD devra s'ajuster à l'avancement du processus bibliographique de façon à publier en LOD les données qui sont établies avec un degré raisonnable de certitude et avec suffisamment de liens<sup>79</sup>. De même que le projet bibliographique de la Bibliothèque numérique des polygraphes comporte un choix critique de sources, le projet bibliothéconomique de publication en LOD implique un choix similaire d'ensembles de données (« *datasets* »). Il ne peut y avoir de meilleur environnement que ce congrès de l'IFLA pour rappeler que les arbres que sont les schémas de métadonnées, les ontologies et les alignements d'ontologies ne peuvent pas nous faire perdre la vision de la forêt bibliothéconomique, des objectifs bibliographiques du système d'information qui va être publié en LOD.

## 6. Processus automatiques

En tant que consommatrice d'information, la Bibliothèque numérique des polygraphes est en train de concevoir différentes applications pour la capture, l'intégration et la diffusion des données. Nous avons déjà mentionné la réutilisation des données de GeoNames ou de LEM. En outre, au moment où nous rédigeons cette communication, d'autres applications sont en cours de conception pour la capture d'ensembles de données (« *datasets* ») du LOD au moyen de procédures semi-automatiques. Ces processus peuvent impliquer la capture de triplets en relation avec un ou plusieurs auteurs ou l'obtention des URI de la ressource uniquement. Ces données sont emmagasinées dans un entrepôt intermédiaire dans lequel il est possible d'effectuer des opérations de recherche, de sélection et de modification, de façon – après comparaison – à permettre leur intégration à la base de données bibliographiques ou dans le fichier d'autorité, tout en gardant la possibilité de choisir soit les attributs correspondant à une entité soit uniquement son URI et de mettre à jour les notices de base.

La même procédure sera effectuée avec les descriptions des ressources bibliographiques et d'information. Ce processus débutera prochainement étant donné que se présente le problème supplémentaire de l'identification correcte des œuvres. Si l'identification des noms de personne dans les fichiers d'autorité a hérité de problèmes tels que l'absence de données suffisantes pour cette identification, dans le cas des œuvres, ce manque de données est encore plus grand. Par exemple, le manque de titres uniformes et d'entrées auteur-titre uniforme, l'absence de relation entre le titre d'une œuvre et les traductions de cette œuvre rend particulièrement compliqué la réalisation de processus automatiques ou semi-automatiques.

D'autre part, toutes les sources de données bibliographiques ou d'autorité ne sont pas disponibles en LOD, même si le paysage est en train de changer peu à peu ; c'est pourquoi la Bibliothèque numérique des polygraphes utilise des serveurs SRU sur lesquels elle peut lancer des profils déterminés de recherche. Il est possible de configurer ces profils de recherche de façon à varier les questions en fonction du ou des éléments recherchés. L'avantage de cette conception est que les résultats obtenus peuvent être emmagasinés dans un entrepôt OAI qui réalise les fonctions de collecte périodique à partir des résultats des

---

<sup>79</sup> Weibel, Stuart. *Principles of Linked Data Recast*. Weibel Lines [http://goo.gl/sYSDe]

serveurs SRU. Nous sommes conscients des grandes possibilités que ce système peut offrir pour créer un service de collecte au niveau national ou même international. L'immense progrès d'Europeana jusqu'à aujourd'hui reposait sur la collecte des métadonnées via le protocole OAI ; un système similaire, qui pourrait s'appeler *Americanae*, pourrait être lancé par cette procédure. Le principal avantage serait la collecte sélective des métadonnées et nous espérons qu'il sera possible d'obtenir de très bons résultats aussi bien pour les ressources espagnoles qu'internationales.

## 7. Encyclopédie des autorités

Disons qu'avec ce modèle on est passé des notices d'autorité de noms de personne aux notices de personnes, ce qui n'est pas une question de dénomination mais de portée. De fait, quelques-uns des problèmes de catalogage qu'il faudra résoudre prochainement consisteront à conjuguer les attributs correspondant aux *personnes* avec ceux correspondant aux *noms de personne*. Un coup d'œil à quelques ontologies et ensembles de données (« *datasets* ») comme DBpedia ou VIAF nous montre les différentes définitions de classes et d'attributs que l'un et l'autre mettent en œuvre pour établir les noms de personne.

Si le centre du diagramme du nuage du Linked Open Data est DBpedia, c'est-à-dire Wikipedia, il paraîtrait normal que la représentation de nos fiches encyclopédiques d'auteurs rappelle ou ait une structure similaire à celle de Wikipedia, mais en utilisant le format MARC comme base. Ne pas prendre en compte l'immense quantité (plusieurs millions) de notices MARC existantes dans le monde serait absurde. D'un côté, nous avons l'extraordinaire granularité et souplesse de MARC 21 qui peut être facilement converti en structures XML. Il est possible, donc, pour n'importe quel système d'obtenir les données nécessaires de n'importe quel ensembles de données (« *datasets* ») disponible en LOD, mais il ne faut pas oublier qu'il s'agit non seulement de mettre en relation et de lier les données, mais aussi de les conserver. Dans ce sens, les structures de DBpedia, VIAF, MARC 21 et EAC-CPF peuvent être adaptées pour permettre de combiner l'information, mais cela peut être insuffisant, car la mise en relation des données peut donner lieu à l'obtention des données contradictoires, erronées, inexactes ou redondantes. C'est-à-dire qu'une nouvelle structure de données est nécessaire pour les notices de *personnes* et les *noms de personne*, destinée à la gestion et à la maintenance des données.

La structure des entrées d'autorité pourrait être une combinaison de schémas ayant pour base le format MARC, qui partagerait en outre l'analyse de MADS/RDF, EAC-CPF<sup>80</sup> et pourrait inclure quelques-uns des attributs de DBpedia<sup>81</sup>. Il est fort possible que l'influence qu'exerce DBpedia rende nécessaire d'élargir le nombre des attributs qui définissent une *personne*. Par un coup d'œil rapide à l'ontologie de DBpedia nous observons quelques-unes des propriétés de la classe *personne*, d'un véritable intérêt pour la Bibliothèque numérique des polygraphes comme "influenced", "influencedBy" ou "philosophicalSchool".

Cet aspect peut être bien mieux apprécié dans le sous-projet de la Bibliothèque numérique des polygraphes intitulé Escuela de Salamanca (École de Salamanca). Sous cette dénomination collective a été regroupé un ensemble de professeurs d'université espagnols et portugais (majoritairement des théologiens, juristes et économistes) qui ont participé de façon directe

---

<sup>80</sup> Bien que cela paraisse évident tout au long de cette communication, nous ne pouvons pas omettre de mentionner la récente publication du CIDOC-ICOM Linked Open Data Recommendation for Museums. [[http://www.cidoc-crm.org/URIs\\_and\\_Linked\\_Open\\_Data.html](http://www.cidoc-crm.org/URIs_and_Linked_Open_Data.html)]

<sup>81</sup> <http://mappings.dbpedia.org/server/ontology/classes/Person>

au renouveau de la pensée au XVI<sup>e</sup> siècle et au début du XVII<sup>e</sup>, après la découverte du Nouveau Monde, après le travail intellectuel et pédagogique de Francisco de Victoria, pour citer un exemple (et un nom éminent) de l'Université de Salamanque. L'objectif ultime n'est autre que de suivre l'influence de ces précurseurs de quelques branches du droit, de la politique et spécialement de l'économie<sup>82</sup>. L'Époque moderne implique un changement important du concept de l'homme en société et c'est justement l'École de Salamanque qui aborde ces problèmes à partir de nouveaux questionnements ; par exemple, Francisco de Vitoria, Domingo de Soto, Martín de Azpilcueta, Tomás de Mercado o Francisco Suárez, essayèrent de réconcilier l'enseignement thomiste avec le nouvel ordre social et économique. Ainsi nous trouvons un groupe d'auteurs qui est à l'origine des sciences économiques<sup>83</sup> ; en effet, préoccupés par la légitimité morale de la conquête, ils ont développé des théories très novatrices<sup>84</sup>.

En récapitulant, pour les auteurs de la Bibliothèque numérique des polygraphes, écrivains de plusieurs milliers d'œuvres que nous avons pu localiser dans une multitude de bibliothèques, nous voyons que nous avons mis en lien les œuvres avec les auteurs en utilisant VIAF ; que postérieurement nous avons développé ce réseau de lien en établissant une navigation numérique entre les sujets de ces livres, en utilisant les vocabulaires ou listes de vedettes tels que LEM ou LCSH, RAMEAU ou SWD ; et que ces auteurs sont maintenant mis en relation et agrégés spatialement et chronologiquement, pour former un réseau facilement lisible. Beaucoup de liens ont pu être établis grâce aux données et, surtout, aux métadonnées concernant non seulement le *nom de personne* mais aussi la *personne*, et sans aucun doute les collectivités pourront être traitées de la même façon que les personnes. Nous devons maintenant ajouter les spécificités de l'*oeuvre* parce que, dans une grande majorité de cas, l'oeuvre, à travers ses différentes expressions, s'est manifestée grâce aux *imprimeurs*, *éditeurs* et aux *libraires* (et les *bibliothèques* où sont conservés les *item* ou *exemplaires*) et toutes ces entités auront leur représentation aussi bien dans l'espace que dans le temps. Par exemple, en suivant l'École de Salamanque, nous verrons jusqu'à quel point le Convento de San Esteban (40° 57' 38"N 5° 39' 47" O) a joué un rôle essentiel parmi les enseignants et étudiants, y compris en tant qu'éditeur et imprimeur, travail qu'il poursuit d'ailleurs jusqu'à aujourd'hui. Les chapelles invisibles, expression provenant de l'époque des écoles de documentation, deviennent clairement visibles.

## Redéfinition des fonctionnalités des SIGB

Donc les possibilités qu'offrent et offriront les liens entre les données va conduire à une redéfinition complète des fonctionnalités des SIGB (Systèmes intégrés de gestion de bibliothèque), en l'occurrence de DIGIBIB. En premier lieu, comme nous l'avons mentionné, pour permettre l'intégration d'éléments d'information ou d'attributs absents du format MARC pour la gestion et l'export en tant que LOD. En second lieu, pour permettre la

---

<sup>82</sup> Agenjo Bullón, Xavier; Hernández Carrascal, Francisca y Juez García, Patricia: La Escuela de Salamanca considerada desde el punto de vista de la Web semántica y la información en la red. Communication présentée le 14 avril 2011 aux *X Jornadas de la Asociación de hispanismo filosófico: crisis de la modernidad y filosofías ibéricas*, tenues à Saint-Jacques-de- Compostelle, les 13-15 avril 2011. [<http://goo.gl/uFrXx>]

<sup>83</sup> Schumpeter Schumpeter, Joseph A. *History of Economic Analysis*. New York: Oxford University Press, 1954.

<sup>84</sup> Grice-Hutchinson, Marjorie. *The School of Salamanca: Readings in Spanish Monetary Theory, 1544-1605*. Oxford: Clarendon Press, 1952

sélection d'ensembles de données complètes ou choisies issues de sources externes ; ce qui suppose d'avoir la capacité de mettre à jour totalement ou partiellement une notice. Par exemple, comme c'est le cas dans la Bibliothèque numérique des polygraphes, d'avoir la possibilité de ne sélectionner que l'URI d'une ressource particulière. En troisième lieu, il sera nécessaire de modifier les interfaces de consultation en permettant la récupération et la combinaison en ligne des autres sources, grâce aux APIs s'appuyant sur SRU/OpenSearch ou également en effectuant des combinaisons avec SPARQL. En dernier lieu dans cette liste non exhaustive, il devrait être possible de gérer un autre type d'informations peu utilisées, comme les descriptions des ensembles de données (« *datasets* ») du LOD eux-mêmes.

## 8. Disponibilité d'ensembles de données (« *datasets* »)

En plus de la disponibilité des ensembles de données (« *datasets* »), la sélection critique des ces ensembles est très importante, comme elle l'a toujours été, car en fin de compte ce ne sont pas autre chose que des ressources d'information bibliographique dans une nouvelle structure ; l'analyse critique de la qualité des données du point de vue intellectuel et théorique demeure. Hormis ce qui est requis pour enregistrer ensembles de données (« *datasets* ») selon CKAN<sup>85</sup> ou VoID<sup>86</sup>, il est évident qu'il faut disposer d'informations concernant non seulement l'ontologie ou la structure des données mais aussi la fréquence de mise à jour, le nombre d'entrées et de relations, de même que les sources utilisées pour effectuer une quelconque fusion de données. Nous pouvons dire que certainement l'expérience la plus immédiate des données liées à des sources externes, comme cela est par ailleurs bien connu par les bibliothécaires et les bibliographes, réside dans la variété des assertions concernant une même *chose* qui, comme nous l'avons dit, sont parfois inexactes, d'autres fois erronées, redondantes et, dans d'autres cas, manquent de garantie de qualité quant à la source.

Le WEB sémantique et le LOD (ou Web de données) proposent une plus grande exactitude pour les navigations mais on peut apprécier que, même si présentement il y a peu de sources disponibles en LOD, il est facile de trouver des *choses* définies de différentes façons, avec des attributs très différents et que les liens *per se* ne contribueront pas à améliorer l'exactitude de la navigation. Il faut se rappeler sur ce point ce qui a été dit sur la nécessité de transformer en ensembles de données (« *datasets* ») du LOD une énorme quantité de sources bibliographiques de première qualité, qui ne sont pas encore numérisées et ne sont pas structurées parce qu'elles ne sont pas encore dans le domaine public. Nous considérons que c'est une responsabilité que les bibliothèques et les bibliothécaires ne peuvent éluder et que des projets massifs de numérisation qui convergent dans Europeana et Hispana sont en train d'être menés à un bon rythme. *Das Ding im sich, das ding am sich y das ding bei sich*, selon l'idéalisme allemand du XIX<sup>e</sup> siècle.

## 9. Conclusions

Nous ne cachons pas que tout le travail décrit ici a gravité autour de deux projets qui, lorsque nous achevons la dernière révision de ce texte, le 30 avril 2011, ne peuvent encore être

---

<sup>85</sup> Comprehensive Knowledge Archive Network. [<http://ckan.net/package/new>]

<sup>86</sup> *Describing Linked Datasets with the VoID Vocabulary* [<http://www.w3.org/TR/void/>]

connus. D'un côté, nous avons la mise en route de la phase *Danube* d'Europeana et, de l'autre, le résultat de l'application expérimentale de RDA dans un large ensemble de bibliothèques nord-américaines. Si à cela nous ajoutons que les conclusions du groupe de travail W3C Library Linked Data Incubator Group ne sont pas encore disponibles, il est évident que nos résultats ne sont pas autre chose qu'un ensemble d'actions effectuées parallèlement au développement et à l'implémentation de ce nouveau paradigme. Peut-être est-ce ici que réside le côté le plus intéressant de notre contribution, étant donné que la Bibliothèque numérique des polygraphes a mené des actions similaires à celles de l'implémentation de RDA et participe activement à Europeana (et c'est un cas d'étude du LLD). Par conséquent notre proposition de travail réel, accessible et disponible via le WEB pour qui voudrait la consulter, peut être considérée comme un apport de plus pour arriver à ce que l'information (et pas seulement l'information bibliographique) atteigne la plus grande accessibilité possible et, surtout, remplisse les objectifs proposés par Tim Berners-Lee quand il a lancé le Linked Open Data. Voilà le sens ultime de cette communication.