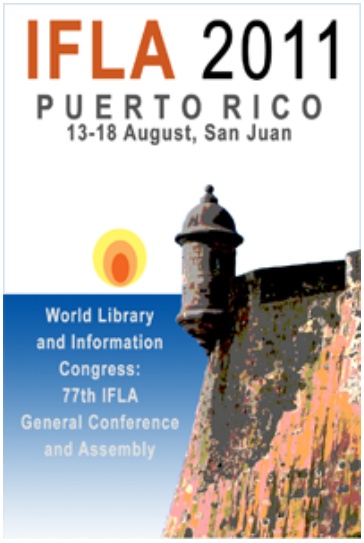

Data aggregation and dissemination of Authority Records through Linked Open Data[1]

Authors[2]:

Xavier Agenjo
Project Manager Fundación Ignacio Larramendi and director of Polymath Virtual Library
[xavier.agenjo@larramendi.es]

Francisca Hernández
Metadata Librarian, Consultant, DIGIBÍS Producciones Digitales
[francisca.hernandez@digibis.com]

Andrés Viedma
Software architect, DIGIBÍS Producciones Digitales
[andres.viedma@digibis.com]

Meeting: **Cataloguing Section**

Abstract:

Throughout the analysis of the Polymath Virtual Library the data aggregation and dissemination of Authority Records through Linked Open Data is described. The aim is to bring together information, data, digital texts and websites about Spanish, Hispano-American, Brazilian and Portuguese polymaths from all times. As such it aggregates information about the thinking, philosophy, politics, science, etc. from Spain, Hispano-American, Portugal and Brazil written in any language (Latin, Arabic, Hebrew, Spanish, Portuguese ...) and at any time (since Seneca in the first century BC to the present).
The backbone of the system are the authors. For each author a MARC21/RDA authority record is made and is enriched with biographical data. Specific attributes are categorized to enhance relationships and navigability of the site (profession, occupation, gender, membership, birth and death dates, places of birth and death and languages or script use). In that way each authority record aggregates information from multiple sources and vocabularies. Also for each author it is described, following MARC/RDA (12 update), the digital versions of their works. Similarly, each author is

[1] All figures and URL have been verified and updated as at April 30, 2011

[2] We thank César Juanes, DIGIBIS R&D Department, for reviewing this paper.

related to other authors (translators, publishers, commentators, etc.), as a way of following the textual transmission of their work using author-title authority records which will be categorized as WEMI following the next decision of MARBI.
The creation of these authority records is made through the consultation of different trusted sources (authority files, encyclopedias, biographical dictionaries, etc).The relationship of a particular author with the data found was done manually until 2010. In 2011 the Fundación Ignacio Larramendi is developing different tools to semi-automatically obtain data from Linked Open Data sources. The paper details the process of obtaining URIs of LOD resources generating automatic queries against SRU/OpenSearch servers and SPARQL Endpoint, or doing this semantic enrichment through files available on the LOD. Data export is made in MARC 21, Europeana Data Model 5.2.1, SKOS and VIAF, after semantic enrichment from vocabularies like LCSH, VIAF, GeoLinkedData (from the Instituto Geográfico Nacional) and GeoNames.

1. Introduction: The Polymath Virtual Library

The following reflections on aggregation and dissemination of authority records in Linked Open Data are not based on the theory, but on a library practice currently in production. These cataloging works are made in the Polymath Virtual Library[3] that is a part of the Virtual Libraries of the Fundación Ignacio Larramendi[4]. The fundamental objective of the Polymath Virtual Library is to give a special relevance and significance to the work of Spanish, Portuguese, Brazilian and Latin American thinkers (polymaths) and, this is one of the main features of the bibliographic project, to relate them to other works of similar characteristics, to make them widely available internationally.

This review highlights from the beginning, although it seems obvious, the importance of a prior bibliographic design, something that should take precedence over technological issues, especially on a matter as new as Linked Open Data. Indeed, these bibliographic objectives are the foundation for the functional requirements of both cataloging practice and the specific development of the Polymath Virtual Library. The project aims to collect and link bibliographic information, or merely information, about Iberoamerican thought from a historical perspective, establishing in turn relationships with other 'civilizations', to use the concept of Toynbee[5].

[3] http://goo.gl/3kn00

[4] To review the state of the art of this project a paper could be seen Agenjo Bullón, Xavier y Hernández Carrascal, Francisca. *La Biblioteca Virtual Larramendi: fuente de información bibliográfica para el pensamiento iberoamericano en la Web 3.0*. En: *Jornadas Virtuales Iberoamericanas de Bibliotecología*. [http://goo.gl/VbwtN]. This pages were written in autumn 2010, some important changes have been implemented since then, some of them can be seen in: Agenjo, Xavier y Hernández, Francisca. *La Biblioteca Virtual Ignacio Larramendi desde la perspectiva LOD y EDM* that was presented at the I Seminario Internacional de la Biblioteca de Galicia [http://goo.gl/uFrXx].

[5] As may be recalled Toynbee in his *A Study of History* could determine up to 21 different civilizations and although this information may seem outdated and even the application of the concept of civilization by Huntington in his famous book (and for our pourposes in that one that followed after) the fact is that there seems to be isolated compartments between cultures that undoubtedly the exchange of information, especially raised as in Tim Berners-Lee in his seminal *Dessign Issues: Linked Data* [http://www.w3.org/DesignIssues/LinkedData], should solve.

The information sources for authority records

It is clear that much of the information sources that are useful for bibliographic project objectives are not available in digital format, or being digitized they lack an appropriate structure. And in other cases, in which the resources are available on the Web, the interface design reduce their informative capacity and performance. For example, the *Diccionario Biográfico Español de la Real Academia de la Historia* provides capital information on thousand of persons relevant for Spanish history and Hispanic culture, understood in the broadest sense, but in a impoverished technological way. Another source of information is *Hombres y documentos de la filosofía española* of Gonzalo Díaz y Díaz, in 7 volumes published between 1980 and 2003, that sins of excessive erudition, with little heuristic approach, but provides access to an enormous amount of information[6]. The same could be said of the great *Diccionario de filosofía of Ferrater Mora*[7], much more rigorous from an intellectual standpoint, but more scarce in regard to the information itself.

It is a great pity that bibliographic repertoires published on paper in its day and still protected by copyright are inaccessible on the Web and in particular to technologies such as Linked Open Data. The paradox is that large retrospective digitization projects are stopped, according to countries and relevant legislation, around 50, 70 or even 80 years after the author's death, thus completely outdated content is easily accessible, while other more valuable and up to date are not available. When talking about the problems of Open Access it's usually forgotten a basic, is not that digital information is accessible or not open, but simply that there is no such a digital information.

This -that should be taken into account in a conference of librarians for whom bibliography should be, more than in any other case, a fundamental discipline-, makes us forget that often the best critical editions with manuscripts collated variants and preliminary studies, etc.., are not accessible on the Web. Of course, there are admirable projects, more and more frequent, which applies techniques of digital editions of historical texts significantly, but they are a minority. Because of its extraordinary quality we cite *Mark Twain Papers & Project*[8] as an example of a digital edition of a relevant author.

Fortunately, the Polymath Virtual Library has tried to, at least, count with scholars with the aim of presenting an overview of the state of the bibliographic studies for each author (polymath) in particular. Moreover, in general, a minimum of information sources has been established that should be consulted in all cases: *Diccionario Biográfico de la Real Academia de la Historia*[9], *Fichero de Autoridades de la Biblioteca Nacional*[10], *Virtual International Authority File*[11] (that includes authority

[6] Available at *Biblioteca Saavedra Fajardo de Pensamiento Político Hispánico* [http://goo.gl/NHiVW]. The transcript of the text does not appear to be comprehensive.

[7] Ferrater, Mora J. *Diccionario de Filosofía*. Madrid: Alianza, 1979. This is the definitive edition of Ferrater Mora, then an expanded edition appeared in the Chair Ferrater Mora under the direction of Professor Jose Maria Terricabras that, fortunately, identified with an asterisk which added by him.

[8] http://bancroft.berkeley.edu/MTP/

[9] http://www.rah.es/diccBiografico.htm

[10] http://catalogo.bne.es/uhtbin/authoritybrowse.cgi

files from Biblioteca Nacional de España and Biblioteca Nacional de Portugal), *Library of Congress Authorities*[12], and prints already mentioned, *Hombres y documentos de la Filosofía Española y Diccionario de Filosofía.* Logically, other sources are used as *Stanford Encyclopedia of Philosophy*[13], *The Catholic Encyclopedia*[14], *Enciclopedia católica*[15], *Jewish virtual Library*[16], *Islamic philosophy Online*[17], *Proyecto Filosofía en Español* [18], *Biblioteca Miralles*[19], *Biblioteca Saavedra y Fajardo*[20], *Biblioteca Virtual Miguel de Cervantes*[21], *Proyecto Sarmiento*[22], among others.

From authority records to bibliographic records

Among the general sources VIAF has been included preferentially. From it the authors have been found in a number of authoritative sources of information such as the Library of Congress, the British Library or the Deutsche Bibliothek. What constitutes a great advantage of VIAF is its ability to navigate from authority records clusters to each of the headings constituent from different authority files that contribute to VIAF. This allows to prioritize the choice of headlines by source: the National Library of Spain for Spanish authors, the National Library of Portugal for Portuguese authors, that of Brazil[23] for Brazilians and so on. And, as a common bond, those from the Library of Congress. It is essential that, from a query it can set a new and very important and universal feature consisting in building up a digital aggregate with all bibliographic records linked to authority records. Thus an author's name and its variants can be located, which can carry out a comprehensive enrichment through associated bibliographic records, made in most cases by National Bibliographic Agencies.

VIAF can connect at this time to nearly twenty authority files through which it gives access to a large number of library catalogs. It is very important to note that only in some cases fields 100 are linked to 6XX fields of bibliographic records. If this functionality were always present there would provide also access to works about an

---

[11] http://viaf.org/

[12] http://authorities.loc.gov/

[13] http://plato.stanford.edu/

[14] http://www.newadvent.org/cathen/

[15] http://ec.aciprensa.com/

[16] http://www.jewishvirtuallibrary.org/

[17] http://www.muslimphilosophy.com/

[18] http://www.filosofia.org

[19] http://www.bibliotecamiralles.org/escritores.html

[20] http://saavedrafajardo.um.es/Biblioteca/IndicesW.nsf/Inicio?OpenForm&m=2

[21] www.cervantesvirtual.com

[22] http://www.proyectosarmiento.com.ar/

[23] Not include yet in VIAF

author[24], and then VIAF could be used as a library resource hub not only from an author but also about works by the same author.

Information sources for bibliographic records

Just as we have defined a number of sources for authors, the same mechanisms has been established for bibliographic sources and therefore have been selected the *Catálogo Colectivo de Patrimonio Bibliográfico*[25], the *Heritage of the Printed Book Database* (HPB) from the Consortium of European Research Libraries (CERL)[26], the *Novum Regestrum*[27] and *WorldCat*. Special mention should be made to Menéndez Pelayo Virtual Library[28]. This virtual library continues the bibliographic work and compilation initiated by Menéndez Pelayo in its work *La ciencia española*[29] devoted to leading Spanish thinkers and their influence, and composed as a reply to the famous M. Masson[30] question *Que doit-on à l'Espagne? Et depuis deux siècles, depuis quatre, depuis dix, qu'a-t-elle fait pour l'Europe?*.

In addition, we must mention two large-scale projects being undertaken in parallel in time, but with different dimensions that affect the Polymath Virtual Library not only in its content but also from the point of view of functional requirements. These projects are Hispana[31] and Europeana[32]. About Hispana there are few papers detailing their creation and evolution, which is surprising given the characteristics and dimensions of the project, although there are numerous presentations[33]. When writing these pages [April 30, 2011] Hispana collects 3,181,786 digital objects from 146 Spanish repositories.

---

[24] It is particularly interesting the access to the Library and Archives in Canada [http://www.collectionscanada.gc.ca/] through VIAF, although for the purposes and objectives of the Polymath Virtual Library has not been found, at least for the moment, none of the authors who form part of the project, but it would be possible if the methodology is extended.

[25] http://www.mcu.es/bibliotecas/MC/CCPB/index.html

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

[27] http://goo.gl/fqtHA. See also *Novum Regestrum: el Catálogo Colectivo del Patrimonio Bibliográfico Iberoamericano* de Xavier Agenjo Bullón, Francisca Hernández Carrascal, *Boletín de la ANABAD*, ISSN 0210-4164, Tomo 44, Nº 4, 1994 , 127-142 http://dialnet.unirioja.es/servlet/articulo?codigo=50938

[28] http://www.larramendi.es/i18n/bvmpelayo/inicio.cmd

[29] Menéndez Pelayo, Marcelino. *La ciencia española*. In: *Biblioteca Virtual Menéndez Pelayo*. http://goo.gl/iB5m

[30] Nicolas Masson de Morvilliers (1740-1789)

[31] http://hispana.mcu.es

[32] http://europeana.eu

[33] *Hispana y las iniciativas del Ministerio de Cultura* María Antonio Carrato Mena *Jornada de Difusión de EuropeanaLocal*, 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 EuropeanaLocal*. 17 nov 2010 URI: http://hdl.handle.net/10421/4783. More recent: Carrato, María Antonia. Hispana. *I Seminario Internacional de la Biblioteca de Galicia* [http://goo.gl/6xIqq]

About Europeana there are obviously a lot more documentation, but for the purpose of this communication it is important to note the new data model *Definition of the Europeana Data Model Elements*, version 5.2.1[34], updated on March 7, 2011, and the *Functional specification for Europeana the Danube release*, published on August 31, 2010[35].

It is not by coincidence that the developments carried out for the Polymath Virtual Library and for the Fundación Ignacio Larramendi are so strongly associated with the functional specifications of Europeana. The reason is this: in a completely majority Spain participates in Europeana through Hispana (with 1,367,808 digital objects at great distances of the Biblioteca Virtual Cervantes [36] that provides only 19,062, although this second project is much more publicized). It happens that both Hispana and most content providers harvested by Hispana[37] are running in the ILS DIGIBIB, currently at version 6.0, but that will be in version 7.0 when this communication is given in Puerto Rico.

Nor should be overlooked the fact that the company DIGIBIS[38], which carries out the ILS DIGIBIB, as well as DIGIARCH for archives and an OAI compliant repository called OAsIs, is a company owned by Fundación Ignacio Larramendi, and that the Polymath Virtual Library is systematically used as a *testbed* for its developments. Therefore (and if so is desired by the clients) many of the features described here significantly affect some thirty major Spanish digital libraries and more than a hundred, if we consider the digital collections collectively gathered by the Biblioteca Virtual de Patrimonio Bibliográfico and the Biblioteca Virtual de Prensa Histórica. [39]

When this communication be read in Puerto Rico it will have entered into production the Danube Phase of Europeana, scheduled for May 1, 2011. Naturally, we cannot be sure that this will be definitely so, but anyway, some delay is completely independent of what has been planned in the Polymath Virtual Library, since Europeana's new semantic features have been truly decisive to develop a new semantic structure of data. For these features, both Europeana and Hispana are two primary sources of information to the

[34] http://goo.gl/ojlL

[35] http://goo.gl/P6jme

[36] http://www.cervantesvirtual.com/

[37] Biblioteca Virtual de Prensa Histórica, Galiciana: Biblioteca Digital de Galicia, Biblioteca Virtual de Andalucía, Gredos (Universidad de Salamanca, Spain), Biblioteca Digital de Madrid, Biblioteca Digital de Castilla-La Mancha, Centro de Documentación de 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, Catálogo Colectivo de la Red de Bibliotecas de los Archivos Estatales, 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.

[38] http://www.digibis.com

[39] Can be seen especially in Biblioteca Virtual de Patrimonio Bibliográfico subdomain *Iberoamérica en las colecciones de la BVPB* [http://goo.gl/8Jyfa]. Also, it can be useful the full text search in Biblioteca Virtual de Prensa Histórica [http://prensahistorica.mcu.es].

Polymath Virtual Library and two methodological examples that also feed-back, as is the case of Hispana.

For now, Europeana has released an API[40] based on OpenSearch that can be integrated within a Web site search and display records retrieved from Europeana. This API, which works very effectively and has been incorporated already into the web interface of the Polymath Virtual Library, automatically launches a search from an initial query in Europeana database. That is, the same query on a database of several thousand records, is also run on the Europeana database, which has around twenty million digital objects (17,901,019).

2. Digital Aggregates

According to the established methodology for the Polymath Virtual Library, the core components of the information system are called 'digital Aggregates', which even in its style and layout are very close to traditional encyclopedia's entries, and also they try to reach the level of complexity and completeness of some entries in Wikipedia (or DBpedia, as it is discussed below).

The terms aggregate and aggregation have different uses depending on the environment to which apply and can generate some uncertainty in the reader. Thus, in the process of harvesting metadata, harvesters that also can be harvested are called aggregators, as in Europeana. However, the concept of digital aggregate used in the Polymath Virtual Library is much more in line with the definition that makes Open Archives Initiative Object Reuse and Exchange (OAI-ORE)[41], for which an aggregate is a resource itself consisting of a set of interrelated resources. A digital aggregation would then be the set of data and digital information resources gathered around a particular author, and its core is an authority record.

The process of authority control carry out by the Fundación Ignacio Larramendi is much broader than the one that traditionally takes place in large libraries or bibliographic agencies. We will not insist on the function of identifying the entity person but on the tasks of contextualization, in the way defined by FRAD[42]. Thus, not only unusual data in authority files are recorded, but also records are linked to other authority data resources and other sources and data available on the Web. In this sense, sources of data obtained (recorded in 670) are not only a justification for the choice of names and their variants, but also track sources of bibliographic information about a particular selected author.

In a similar way, as VIAF does, each authority record is defined, following the MARC 21 Format, for its main heading (1XX), its variants of name (4XX), its relations with other names (5XX) and for the heading equivalence in other languages (7XX). The latter aspect is of great importance for the Polymath Virtual Library since all authors are selected not only for its importance within the Iberoamerican culture, but for the contribution of Iberoamerican culture to universal culture, which is why it is absolutely need to have the equivalents in other languages. In that way it is possible to follow the

[40] http://www.version1.europeana.eu/web/api

[41] http://www.openarchives.org/ore/

[42] http://www.ifla.org/files/cataloguing/frad/frad_2009-es.pdf

transmission of texts worldwide, as well as the transmission and impact of ideas. But the same situation exists in Spain where, together with Castilian or Spanish, other languages such as Catalan, Valencian, Galician and Basque coexists.

Multilingual headings

We want to draw attention to the need for the MARC 21 Format of having an specific element for coding language of the headings. Unfortunately, the Discussion Paper 2001-DP05 Multilingual Authority Records in the MARC 21 Authority Format[43], seems not to have ended in a proposal, at least for now. 7XX fields allow to record equivalent headings in different languages, but have no way to indicate the language of the heading, which is especially necessary in multilingual authority files. Other mechanisms also have limitations: the 008/08 position only codes English and / or French as language of the catalog and the $ 040 b allows only to set the language of cataloging that often differs from the language of the heading.

This situation is reasonably overcome in MADS and certainly in the proposed MADS / RDF[44,] but do not forget that most of the authority records are in MARC Format only and therefore it will be difficult to move immediately this type of data to the Linked Open Data environment. So VIAF has inherited this problem and while it is possible to recognize the institution, and country of origin of a record, it cannot be known, except by inference, in which language is expressed a heading. This is especially important in cases of headings produced in the same linguistic area that can be very similar or indistinguishable (eg, *Geografía* in Spanish and *Geografia* [no accents] in Italian or Catalan).

Variants of personal names

Searching for information in other authority files not only has the intention of establishing linguistic equivalence relations between headings, but also to enlarge the number of variants of the names and relationships with other names. This procedure, appropriate to the behavior of ILS DIGIBIB, allows importing bibliographic records properly linked with the established heading and also controls the generation of not authorized entries from the 4XX fields. The process itself is slower in a beginning, at the time of the creation of the authority record, but is much more productive in the medium term and offers much higher quality. Also it must be noted that the original bibliographic project is much better accomplished if the sources are analyzed starting from the core of authority records.

This procedure is especially important when we consider that this is a group of authors ranging from Seneca (Córdoba, 4 BC-Rome, 65 AD) to Martín de Riquer (Barcelona, 1914 -) and many of them have used different names depending on the language in which they have written, not to mention that these names have been translated into several languages along with their works (eg, Averroes is the Latin name of Ibn Rus). To have therefore a large number of variants of names adequate to resources to be

[43] http://www.loc.gov/marc/marbi/2001/2001-dp05.html

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

imported, requires for each ingest of a resource fewer changes to authority records or files.

3. Authority records

The update No. 12 in MARC 21 format, published in October 2010, consolidates a number of fields, following the philosophy of Resource Description and Access, which are essential for the aggregation of information about persons. We refer, of course, at 3XX fields. From a historical[45] standpoint, the elements that initially were intended to differentiate between homonyms, gradually acquired a constitutive value in itself. Thus, different subfields of field 100 which use were recommended only to distinguish authors with the same name, became mandatory, most notably the $d, because immediately contextualized an author from a chronological point of view, that helped many times to interact with the work and undo attribution errors. Until the adoption of RDA by MARC 21[46] there were no specific items for individual attributes so expressive as the birth place. Just these attributes, refer not to the *name* of the person but the *person* itself and help in giving MARC records an essentially encyclopedic look, almost dbpedic, if we may use this neologism.

For the purpose of the Polymath Virtual Library the release of RDA strengthened their work lines. First, in the retrieval of information about the authors, following user tasks defined in FRAD each as contextualization and search. And secondly, structuring its information resources in the form modeled in RDA as Work-Expression-Manifestation-Item[47]. For this reason, subsequent new versions of DIGIBIB immediately implemented the MARC 21 updates to the extent that the ILS DIGIBIB paced its new release to MARC 21 updates.

For each author, as far as possible, it is included, in addition to the canonical form of the name, its variants and equivalents in other languages, the author's biography (678), dates of birth and death (046), birth and death places (370) and any other place of activity (370), occupation (372 and 374), the language used (377) or gender (375). These person's attributes ultimately aim to support semantic search and navigation between data. They allow the Polymath Virtual Library website to offers the ability to search authors by the dates they were active, the places associated with their biography, their professions and occupations, not only by their works, which is relatively common in Web OPACs.

Authority records in other metadata schemes

Of equal interest to the Polymath Virtual Library is the information of corporate bodies, persons and families from archives. The need for convergence of archives, libraries and

[45] Agenjo Bullón, Xavier y Palá Gasós, Pilar. *El fichero de autoridades del Catálogo Colectivo del Patrimonio Bibliográfico.* En: *Boletín de la ANABAD*, ISSN 0210-4164, Tomo 37, Nº 4, 1987 , págs. 593-606. [http://dialnet.unirioja.es/servlet/articulo?codigo=801041]

[46] http://www.loc.gov/marc/RDAinMARC29.html

[47] 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]

museums data is well taken and some examples of excellent quality can be accessed. They often arise as unified query interfaces which contribute to strengthening the informative capacity of such data. Very recently DIGIBIS has carried out a work of this kind with the project of the Junta de Castilla-La Mancha dedicated to Archivo de la Imagen[48] (photographs) and the Archivo de la palabra (sound recordings)[49]. For all this, a systematic mapping[50] of MARC 21 to EAD and EAC-CPF has been done, as well as ISAD (G) and ISAAR (CPF).

Similarly, and as it was expected, the Polymath Virtual Library has already required linking bibliographic and archival materials relating to, for example, Agustín de Montiano y Luyando (1697-1794), historian, critic and play writer, who has passed into collective memory as the first director of the Real Academia de la Historia, founded in 1735, since his papers takes part of the Larramendi Virtual Libraries[51].

4. Linking data

Polymath Virtual Library is a content provider of Europeana, harvested by Hispana that is a national aggregator in the Europeana network, as it has been already mentioned, and uses its database as a source of information. Therefore, it has raised the strategic objective of adopting the model proposed by Europeana Data Model (EDM)[52] and Linked Open Data technologies. Both EDM and Linked Open Data (LOD) are leading to new actions in a extremely rapid way and are providing vast amounts of data as in the case of VIAF[53] or DBpedia[54]. Since the Linked Data proposal by Tim Berners-Lee in 2006 to present, data structured semantically and reusable have been published at a rate surely not seen on the Web before, probably with the exception of library MARC bibliographic records.

For everything related to the library contribution to LOD should be noted that, within the W3C, the Linked Data Library Incubator Group (LLD) [55] was created in 2010. This group is responsible for preparing several reports and deliverables on the status of ontologies for archives, libraries and museums; vocabularies and datasets of interest to these areas and use cases that library community is carrying out. To follow the activity

[48] http://clip.jccm.es/archivo_de_la_imagen/es/micrositios/inicio.cmd

[49] http://clip.jccm.es/archivo_de_la_palabra/es/micrositios/inicio.cmd

[50] http://www.sedic.es/p_boletin_novedades_mensual.asp#Digibis

[51] In fact it was an ancestor of Ignacio de Larramendi y Montiano (1921-2001) founder, among many other initiatives, of the Larramendi Virtual Libraries or DIGIBIS. See also, 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]

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

[53] http://ckan.net/package/viaf

[54] http://dbpedia.org/About

[55] http://www.w3.org/2005/Incubator/lld/

of this group and Linking Open Data[56], as appropriate, will inform to anyone interested in.

For the Polymath Virtual Library Linked Open Data[57], and annexed technologies, represents an opportunity to fully realize the approach of the 'digital aggregates' as nodes of Linked Data. In fact, the Polymath Virtual Library is a use case reviewed by the LLD. It is not only about capturing information, but linking it to appropriate nodes of the LOD and its resources, according to rules of the game where it is expected that each fulfills its role. Each resource available in a LOD dataset is identified by a URI with special characteristics[58]. This URI drags a network of relations, attributes, to other resources that define, according to the model used (ontology or RDF Schema), entities and properties of a given functional area. But being that important, what really adds value is that it allows to identify that an entity in one environment is the same or very similar to another entity defined in another environment. Thus, Maimonides was a philosopher for DBpedia and the name of a person in VIAF. The process consists of linking such data through URIs, opening them for reuse and, through appropriate tools[59], analyze them and deduce other data from their relations and inference rules, and update the original data to serve them again, once modified and expanded. And so on, iteratively.

DBpedia and VIAF

In 2010 the Polymath Virtual Library began the process of linking authority records with similar existing resources in LOD. Obviously, sources of information on the methodology outlined here were the candidate sources for obtaining links. So, VIAF resource URIs are registered in field 024 of the MARC21 records. The RDF resource Maimonides (rdf: about = "viaf/100185495"), groups together established headings for this author from 7 different authority files, with 255 variants of names, but also has links to the content of DBpedia (http://dbpedia.org/resource/Maimonides). If the librarian is experienced and has expertise in this area, ie a senior cataloger, he can deduce from the net of VIAF relationships up to 12 different headings for Maimonides[60], from various countries and languages. Clearly, the process of acquiring these data manually is very accurate but slow, and soon was seen the suitability to automate it, weighing the advantages and disadvantages of doing so. Undoubtedly, the number of authors who make up the Polymath Virtual Library, around 1,000, would be done manually, but productivity in time and data could be increased in an automated way. In addition, once tools were designed and developed could be incorporated as specific features to DIGIBIB. It is expected that upon reading this paper, 13-18 August 2011, these tools make up a module in DIGIBIB 7.0 and later.

[56] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[57] http://www.w3.org/2005/Incubator/lld/wiki/Use_Case_Polymath_Virtual_Library

[58] http://www.w3.org/TR/2007/WD-cooluris-20071217/

[59] http://www.w3.org/wiki/SemanticWebTools

[60] Ibn Maymun, Musà or, the much more familiar to the orthodox Jews, Rambam

These types of links have also been carried out using SKOS. The results indicate that a particular concept is the same or very similar to other concepts in other vocabularies. Through the subject headings of bibliographic records, and linking concepts with other vocabularies also available in SKOS, it can also get a very wide number of works about an author. In addition, the process of finding information resources can be done automatically, by covering the network of relationships between concepts in different vocabularies and being able to retrieve information resources linked to these concepts.

At present, the Polymath Virtual Library is working, in tests, with the List of Subject Headings for Public Libraries of Spain which was set up by the Dirección General del Libro y Bibliotecas and was enhanced by crossing it with the authority file of the Biblioteca Nacional de España[61] and its equivalence with Library of Congress Subject Headings[62]. Because this process is analogous those made to link SKOS concepts of RAMEAU[63] with LCSH and the Schlagwortnormdatei (SWD)[64] it will be feasible to establish such conceptual relationships that certainly will not possess a big granularity, but treated and crossed with a determined number of authors could offer a clear net of links between large sets of information significantly.

Time and space

Two other elements play an essential role in the aggregation and linking data relating to an author, such as chronological and geographical data. Indeed, an author lives and dies on certain dates and works, have social relationships with their contemporaries during certain periods, writes books that are published, translated and edited in very specific dates[65]. The possibility of linking authors data by this new chronological category, not only by name or subject, precise the work set of an author, although it is certain that the scattering or inconsistency will be visible. The codification of time, present in many elements of the authority or bibliographic records, often need attributes to express probability, inaccuracies, estimations, etc., so frequently in descriptions of people and resources, not mention events. Such data are essential for the generation of timelines or to perform operations and deductions with dates, as can be seen, for example, in WorldCat Identities[66].

---

[61] *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.

[62] http://id.loc.gov

[63] http://www.cs.vu.nl/STITCH/rameau/

[64] http://www.d-nb.de/standardisierung/normdateien/swd.htm

[65] There is now a new format that will attempt to standardize the many different ways in which different chronologies have been established and even numbers with those that have been represented, if not through the alphabet, the latin alphabet, the arabic alphabet, etc. Extended Time/Date Format ETDF [http://www.loc.gov/standards/datetime/]

[66] http://www.worldcat.org/identities/

Similar work is being done in aggregating and linking geographic data. In fact, within Linked Open Data exists a major initiative for this type of data, GeoNames[67], who is also the only case known to date that has been replicated in Spain by the transformation of geographic codes from Instituto Geográfico Nacional[68] to RDF and LOD. Again, there, the aggregation process can be enlarged. Authors have born and died in a particular place, have been associated with their contemporaries in certain places, perhaps schools, universities, religious orders, or military, scientific and cultural societies or political groups located or related to a specific location. Therefore, and through them, is possible to establish new and appropriate aggregations and links.

For linking to GeoLinkedData and GeoNames data the process was done in an automated way, converting the selected data in an structure valid for its ingestion in DIGIBIB. These data have been mapped to MARC, to update automatically all subfields in 752 and 151. That way, we have obtained not only the names available in these vocabularies, but also the geographic coordinates. Certainly there must be further manual work to solve problems that may arise to differentiate jurisdictions from geographic locations and geographic locations of old jurisdiction. Similarly, as discussed below, one thing is to manage *persons* and other *persons' names*. It is clear that management jurisdictions, geographic names and geographic locations is supposed to have different data structures.

5. Using LOD

VIAF website can be queried and results can be visualized in MARC XML or RDF. Through its API[69] VIAF can be searched via SRU and OpenSearch, and results can be downloaded in MARC XML, RDF and a variety of formats and schemas. VIAF[70] is also available as Linked Open Data. Following the recommendation to publish data in LOD[71] often the way of accessing and querying datasets is by SPARQL Endpoint[72] or by SPARQL clients[73] that can be used to query RDF files. This means that DBpedia[74] or VIAF can be searched via SPARQL[75] and therefore queries can make a logical sweep

---

[67] http://www.geonames.org/ontology/documentation.html

[68] GeoLinkedData has been launched with the publication of various information sources from the Instituto Geográfico Nacional, making it available as RDF knowledge bases according to the principles of Linked Data [http://geo.linkeddata.es/web/guest;jsessionid=6E6A9C1E1DAF40F81005E8F4DA3A27E1]

[69] http://www.oclc.org/developer/documentation/virtual-international-authority-file-viaf/using-api

[70] http://ckan.net/package/viaf

[71] *How to Publish Linked Data on the Web*. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

Best Practice Recipes for Publishing RDF Vocabularies. W3C Working Group Note 28 August 2008. http://www.w3.org/TR/swbp-vocab-pub/

[72] http://www.w3.org/wiki/SparqlEndpoints

73 http://www.w3.org/wiki/SparqlImplementations

[74] http://ckan.net/package/dbpedia

[75] SPARQL Query Language for RDF. [http://www.w3.org/TR/rdf-sparql-query/]. Now W3C is working in SPARQL 1.1 Federated Query. *This specification defines the syntax and semantics of a SPARQL 1.1*

against the data and retrieve much more related information than in the ways outlined above. Using SPARQL it is possible to define a more complex search, a possible example in DBpedia, but not in VIAF, is 'philosophers influenced by Maimonides'. Therefore, the Polymath Virtual Library is taking steps again, under the bibliographic plan, to take advantage of certain bibliographic information that is pouring in the form of LOD datasets[76]. And doing so from the consumer as well as the producer point of view.

So, the first step was the implementation of Europeana Data Model (EDM) in response to the changes that Europeana has planned for Danube phase. It is possible that Europeana could treat data more or less automatically to transform the information received from content providers to the new EDM, but considering that these providers are already several thousand, working with over 25 languages and as many vocabularies, makes us believe that those treatment only will affect to an undetermined percentage of data, with results impossible to verify and surely with a great risk of making false equivalences and relations. Whether such a procedure can be practical and able to resolve the linkage between one part of the data, cannot be extended to all, and cannot be enhanced without the intervention of content providers themselves. In a typical case, geographical names are often really jurisdiction names that could not coincide with the current geographical coordinates and thus match Castile of the fifteenth century to current Castile is only an approximation. Anyway, the Polymath Virtual Library has begun a process of publishing their data internally linked in LOD.

Mapping MARC 21 bibliographic records to Europeana Semantic Elements 3.3.1 and to Europeana Data Model is of relative difficulty. But, if only this process takes place, without linking this data with external sources or vocabularies, it will be very difficult that Europeana can make an accurate aggregation of data (here, in the sense of OAI-ORE). The methodology of incorporating VIAF URIs in authority records has proved to be very useful since the generation of EDM data can offer a reference to an external vocabulary for the class *Agent*. The second linking process was carried out through LCSH available in *id.loc.gov*. Thus, the Polymath Virtual Library can publish their subject headings in SKOS, but linked to LCSH. Doing this, the property *dc: subject* will be associated with a *skos:Concept*, linked to LCSH and indirectly to RAMEAU[77] and SWD[78], as a minimum. In other words, subject headings available in a semantic structure and associated with the same heading expressed in English, French and German. This shows clearly the advantage of linking different datasets with one or more specific vocabularies.

In addition to this, which is paradoxical, the same procedure may be applicable for interconnecting national data. One effect of using LCSH as one of the main sources to justify subject headings in Spanish libraries is that through this relationship (in field

---

Federated Query extension for executing queries distributed over different SPARQL endpoints. [http://www.w3.org/2009/sparql/docs/fed/service].

[76] http://ckan.net/package. See also the compilation done by the LLD http://ckan.net/group/lld or DOIs as Linked Data http://inkdroid.org/journal/2011/04/25/dois-as-linked-data/ or Linked Periodicals Data http://periodicals.dataincubator.org/.html

[77] http://ckan.net/package/stitch-rameau

[78] http://ckan.net/package/dnb-gemeinsame-normdatei

670) is possible to link different Spanish subject headings lists, that should not be forgot, may be in several languages. This is a quite beneficial side effect of LOD, even in an indirect way, it will allow linking data between Spanish library catalogs and authority files that have remained fairly isolated, much more than desirable and comprehensible.

Here again we must point out another benefit of Europeana in improving information systems of a country. The Spanish Ministry of Culture that maintains Hispana has also begun the process of its adaptation to EDM and, as mentioned, one of the first activities has been the conversion of the List of Subject Headings for Public Libraries (LEM) to SKOS, linking entries, when possible, with LCSH. Logically, the Polymath Virtual Library has completed its linking processes linking their subject headings with LEM. It is hoped that through LEM dataset also other Spanish subject headings lists, in Galician, Catalan, Valencian and Basque, could be linked. In April 30, 2011 the Biblioteca Virtual de Patrimonio Bibliográfico from the Spanish Ministry of Cultura, implemented the ILS DIGIBIB 7.0 ability to export records according to EDM 5.2.1.

These processes will converge both in the full implementation of EDM and in the publication of the Polymath Virtual Library datasets in LOD. In this sense, the publication of data as LOD should be slow, corresponding with the progress of the bibliographic plan so that only data well established, with a reasonable degree of liability, and sufficiently linked[79] were published. Just as the bibliographic project of the Polymath Virtual Library involves a critical selection of sources, the librarianship project for its publication as LOD implies a similar selection of datasets. There can be no better environment that IFLA Congress to remember that trees of metadata schemes, ontologies and ontology alignment cannot make us to lose sight of the library forest, of the bibliographic objectives of datasets that are going to be published in LOD.

6. Automatic processes

As consumers of information Polymath Virtual Library is developing various applications for data capture, integration and dissemination. We already mentioned data reuse from GeoNames or LEM. Furthermore, in the time of writing this communication, other applications are being designed to capture LOD datasets using semi-automated procedures. This process may involve the capture of some or all triples associated with one or more authors, or obtaining only the URI of the resource. This data is stored into an intermediate repository on which search, selection an updating operations can be performed. That way, once data has been verified can be ingested in the bibliographic database or in the authority file, being possible to choose all the attributes of a resource or only its URI, and update the base records.

The same procedure will be carried out with descriptions from bibliographic and information resources. This process will start soon since it has the added problem of the correct identification of works. If the identification of names of persons in authority files has inherited a number of problems such as lack of enough data for contextualization, in the case of works that lack of data may be even greater. So the absence of uniform titles and author-uniform title entries, or relationship between the titles of works and its translations made especially inefficient automated or semi-automated processes.

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

Moreover, not all sources of authority or bibliographic data are available on LOD, although this is changing every day, so the Polymath Virtual Library uses SRU servers on which can launch specific search profiles. These profiles can be configured to accommodate queries according to elements sought. The advantage of the design made is that the results can be stored into an OAI repository that performs regular harvesting of metadata from SRU servers. We are aware of the great possibilities that this system can provide to create a national or even international harvester. The tremendous progress of Europeana to date has been based on OAI metadata harvesting and a similar system, which could be called *Americanae*, can be initiated. The main advantage of this procedure would be the selective metadata harvesting and we hope to obtain good results in both Spanish and international resources.

7. Encyclopedia of authorities

This model has shifted from authority records for *personal names* to the records for *persons*, which is not a nominal issue, but long range. In fact, some of the cataloging problems to be solved in the near future will be to combine *person's* attributes with *personal name's* attributes. A look at some ontologies and datasets as VIAF or DBpedia shows different definitions of classes and attributes that both perform to establish person names.

If the center of the Linked Open Data cloud diagram is DBpedia, that is Wikipedia, it would seem normal that the presentation of our encyclopedic shape authority records recalled, or have a similar structure, to that of Wikipedia, but using MARC format as a basis. Disregard the huge amount (billion) of MARC records in the world would be absurd given also the great versatility and granularity of MARC 21 could easily converted to XML structures. It is possible, of course, that any system could get data from any dataset available in LOD, but do not forget that this is not only about linking data, but also about maintaining and updating data In this sense structures from DBpedia, VIAF, MADS, MARC 21 and EAC-CPF can be arranged to mashup information, but this may be insufficient, since linking data only and by itself could lead to uncontrolled mistakes, inaccuracies or redundancies. That is, we need a new data structure for *persons* and *names of persons* that can support management and maintenance of data.

The structure of authority records may be a combination of schemes based on the MARC format, which also would share the analysis of MADS / RDF, EAC-CPF[80] and include some of the attributes of DBpedia[81]. It is very possible that the influence being exerted by DBpedia becomes necessary to expand the number of attributes that define a person. A glance at the DBpedia ontology shows some properties of the class person of real interest to the Polymath Virtual Library as *'Influence*', *'influencedBy'* or *'philosophicalSchool'* among others.

[80] Although it has been not mentioned throughout this communication, we have to cite the recent publication of CIDOC-ICOM Linked Open Data Recommendation for Museums CIDOC-ICOM Linked Open Data Recommendation for Museums. [http://www.cidoc-crm.org/URIs_and_Linked_Open_Data.html]

[81] http://mappings.dbpedia.org/server/ontology/classes/Person

This aspect can be seen better in the sub-project (or, better, subset) of the Polymath Virtual Library called Escuela de Salamanca. This collective name includes a set of Spanish and Portuguese academics (mostly theologians, jurists and economists) who participated directly in the renaissance of thinking in the sixteenth and early seventeenth centuries that follows the discovery of the New World, with roots in the intellectual and pedagogical work of Francisco de Vitoria, to cite one case (and an eminent one), at the Universidad de Salamanca. The ultimate goal is simply to follow the influence of these precursors of some fields of law, politics and especially economy[82]. The Modern Age was a significant change in the concept of man in society and is precisely the Escuela de Salamanca which addresses these issues from new approaches. That way, Francisco de Vitoria, Domingo de Soto, Martín de Azpilcueta, Tomás de Mercado and Francisco Suárez attempted to reconcile the Thomistic doctrine with the new social and economic order. Thus we find a group of authors that led to economic science[83], that concerned about the moral legitimacy of the conquest, and developed very innovative theories[84].

Summarizing, we see that the authors of the Polymath Virtual Library, authors of several thousand works, which had been located in many libraries, using the procedure of linking works with other authors through VIAF; that later had been expanded, establishing virtual navigation among the subjects of these books, using vocabularies or lists of headings such as LEM, LCSH, RAMEU or SWD, are now interconnected and aggregated spatially and chronologically, forming a network easily visible. It has established a wide relationship through the data, especially metadata referring not so much to the name of the person but to the person himself and, no doubt, that corporate bodies can be treated as persons. We will now add specificities of works because -in the majority of cases- works through its expressions will be manifested thanks to printers, editors and booksellers (and libraries where items are kept or digitized). And all those entities will now have as well their representation in both space and time. Thus, following the Escuela de Salamanca, we will see how far the Convento de San Esteban (40º 57' 38"N - 5º 39' 47" O) played a key role for teachers and students, and even did an editorial and printing work, which certainly continues to the present. That term from the time of Library Science Schools, invisible colleges, become clearly visible.

Redefining ILS functionalities

Since then, the possibilities of providing linked data will lead to a complete redefinition of the functions of ILS, namely DIGIBIB. First, as mentioned, to allow incorporation of data elements or attributes not found in the MARC format for its updating, management and export as LOD. Second, allowing ingestion of complete or selected datasets; this means having the ability to update all or part of a record. For example, it is possible, as in the Polymath Virtual Library to select and ingest only the URI of a resource. Thirdly, it will be necessary to modify query interfaces allowing discovery and online mashup

[82] 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*. Paper given on 2011, April, 14th in *X Jornadas de la Asociación de Hispanismo Filosófico: Crisis de la modernidad y filosofías ibéricas*, held in Santiago de Compostela, 2011, April, 13-15. [http://goo.gl/uFrXx]

[83] Schumpeter, Joseph A. *History of Economic Analysis*. New York: Oxford University Press, 1954.

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

from other sources through APIs based on SRU / OpenSearch or even in a combination with SPARQL. At the bottom of this list, not exhaustive, it should be possible to manage other types of information not widely used as the descriptions of the LOD datasets as a whole, within which we must mention archive finding aids and museum records.

8. Availability of datasets

It should be noted that besides the availability of the datasets it is very important, as it always has been, the critical selection of these datasets, which ultimately are nothing but bibliographic information resources with a new structure. Critical analysis of the quality of data from the intellectual and theoretical point of view remains fundamental. Apart of what is required to register datasets according CKAN[85] or VoID[86], there is clearly a need for information about, not only the ontology or data structure, but also the number of entries and relationships, the frequency of updates, the way that updates are managed, and the sources used in any data fusion. We can say that probably the most immediate experience of linking data from external sources, as indeed is already well known by librarians and bibliographers, is in the range of assertions about the same thing, as we have said, sometimes are inaccurate, sometimes erroneous, redundant, and in others occasions there are no trustworthiness about the source.

The semantic web and LOD propose greater accuracy in navigation, but it can be seen, despite having few sources now available in LOD, that it is easy to find things defined in different ways, with very different attributes and that linking per se could not be enough to improve the accuracy of navigation. It must be noted at this point what has been said about the need to transform in LOD datasets a huge amount of high quality bibliographic sources that remain in paper and unstructured as they are not in the public domain. We believe this is a responsibility that libraries and librarians cannot delay, although mass digitization projects that converge in Europeana and Hispana are making a good progress. *Das Ding im sich, das Ding am sich* y *das Ding bei sich,*, as elucidated by nineteenth century German idealism.

9. Conclusions

All work described here has gravitated around two projects that at the time of closing the last review, April 30, 2011, are still open. On the one hand, the implementation of Europeana Danube phase has just begun. On the other hand, the outcome of the trial implementation of RDA in a wide range of American libraries is expected to become public, may be in next ALA meeting. If we add that conclusions of the working group W3C Library Linked Data Incubator Group are not yet finished, it is obvious that our results are just a set of actions in parallel with the development and implementation of this new paradigm. Perhaps there lies the greatest interest of our contribution, since the Polymath Virtual Library has taken steps similar to the implementation of RDA and participates actively in Europeana and it is a LLD's use case. So our actual work, accessible and available through the Web for those who want to see it, can be considered as an additional contribution to acchieve that information (and not only bibliographic information) reaches maximum accessibility and, above all, meet the

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

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

objectives proposed by Tim Berners-Lee when enunciated Linked Open Data. This is the ultimate meaning of this communication.