Good practices and strategic recommendations on exchange formats applicable to the Registry Service for Libraries and Related Organizations (SERBER) in a context of open and linked data

Report

Madrid, 4 September 2012

A report by Infor@rea commissioned by the Office of Library Co-ordination of the Spanish Ministry of Education, Culture and Sport

Written by
Miquel Térmens Graells
Miquel Centelles Velilla
(Dept. of Library and Information Science. University of Barcelona)
Coordinated by
Elisa García-Morales (Infor@rea)
Domingo Arroyo Fernández (Office of Library Co-ordination. Spanish Ministry of Education, Culture and Sport)

Madrid, 4 September 2012

Contents

Executive Summary 5

1. Background and objectives of publication of SERBER open data in accordance with ISO 2146:2010 7
   - 1.1 Background 7
   - 1.2 Objectives 7
   - 1.3 Formats 9
   - 1.4 Strategic proposal 9
2. The regulatory framework applicable to open data publication in the public sector 10
3. International recommendations and national formats for publishing open data in the public sector 12
   - 3.1 What formats meet the definition of open standards? 12
   - 3.2 What open formats are suitable for the purposes of the SERBER project? 14
4. A review of solutions implemented by other registry services and directories of libraries and related organizations 17
5. Description and comparison of alternatives 24
6. Strategic recommendations for the future: from open data publication to linked data publication 25
7. References 26
8. Glossary 28

Executive Summary

The Office of Library Co-ordination of the Spanish Ministry of Education, Culture and Sport (MECD) commissioned the company Infor@rea to draw up a study of the current situation and best practices regarding exchange formats for data to be published in the Registry Service for Libraries and Related Organizations (hereinafter SERBER). This study falls within the framework of the design and implementation of a strategic model for SERBER based on the ISO 2146:2010 standard.

Section 1 presents the conditions that the project imposes on the data publishing formats and the sources of information used for the study. The conditions include the need to deal in the long term with the data related to the five types of object of the ISO 2146:2010 conceptual model: in addition to the main class "Registry Object", the subclasses Parties (persons and organizations), Activities, Collections, and Services. The data publishing formats are also limited to those within the open data framework. This section also outlines the three basic questions that the study seeks to answer: what are open standards, what specific formats meet the definition of open standards, and what open formats are suitable for the SERBER project?

Section 2 defines and delimits the open standard concept according to the European and national regulations on interoperability and reutilization of data in the public sector. These regulations also provide details of the circumstances in which non-open standards can be applied and the criteria for selecting standards.

Based on recommendations of national and international specialists, Section 3 considers the specific formats that meet the definition of open standards and the open formats that are suitable for the SERBER project. The three reference formats selected for the SERBER project are CSV, XML and RDF (with its various forms of serialization). RDF allows the incorporation of linked data formats, which requires thorough preliminary planning, including analysis and selection of appropriate vocabularies to describe data.

Section 4 presents the results of an exhaustive search and analysis of registry services and directories of libraries and related organizations worldwide that publish data using exchange formats. Ten cases of different types were identified and classified into three groups: registries of libraries and related organizations that apply ISO 2146:2010 and whose main object of description in the conceptual model are parties (persons and organizations); fully operational collection service registries that apply ISO 2146:2010 and whose main object of description in the conceptual model are collections; and finally directories that are designed on the basis of models different to ISO 2146:2010 and therefore do not apply its conceptual model.

The last two sections outline the recommendations for publishing open data of the SERBER project, expressed in two successive stages: Stage 1 (Section 5), in which the data will be published in two structured formats, CSV and XML; and the Stage 2 (Section 6), marked by the strategic transition to linked data in the context of the Semantic Web. Section 6 presents the specific steps to be carried out and the immediate benefits that would be obtained by the MECD.

1. Background and objectives of publication of SERBER open data in accordance with ISO 2146:2010

Within the framework of the design and implementation of a strategic model for the Registry of Libraries and Related Organizations (hereinafter SERBER) according to the ISO 2146:2010 standard, in collaboration with the Ministry of Education, Culture and Sport (MECD), the need for a study of the current situation and best practices regarding data exchange formats was detected.

1.1 Background

According to Royal Decree 257/2012 of 27 January developing the basic organic structure of the Spanish Ministry of Education, Culture and Sport, Art. 10 s), the Office of Library Co-ordination has the function of "obtaining, processing and using library data". The products of this activity include the various directories of individuals and entities that the Office uses for its operational management. The public part of these data constitutes the Directory of Spanish libraries, currently available at <http://directoriobibliotecas.mcu.es/portada.html>.

In 2009 the project of creating a new, unified directory was initiated with the following conditions:

1. Compliance with ISO/DIS 2146:2010. Under this standard, we adopt the variant "registry of libraries and related organizations" characterized by the preeminent position of the object Group (organizations and professionals) in the conceptual model of *SERBER*, which forms the nucleus for the other objects: activities, collections and services (*LIBRARY ISO DIRECTORY project*, pp. 5 and 22).

2. Adaptation to the MARC 21 standard to enable the export and import of information in the MARCXML schema. This standard defines the exchange of information for structuring and identifying data so that it can be recognized and manipulated by a computer (*LIBRARY ISO DIRECTORY project*, p. 5).

3. Interoperability and integration with other information systems and, especially, with the single point of access, in which the priority format for authority, bibliographic and holding records is MARC 21 in some form of coding (ISO2709 or MARCXML).

1.2 Objectives

The possibilities of data download include all elements linked to the five types of object in the ISO 2146:2010 conceptual model. These are, in addition to the main class "Registry Object", the subclasses Parties (persons and organizations), Activities, Collections, and Services. However, at this stage of the project information on activities and collections of libraries and related organizations and statistics on the types of object are excluded. The ISO 2146:2010 conceptual

model will allow all these elements that are initially covered to be incorporated in later stages.

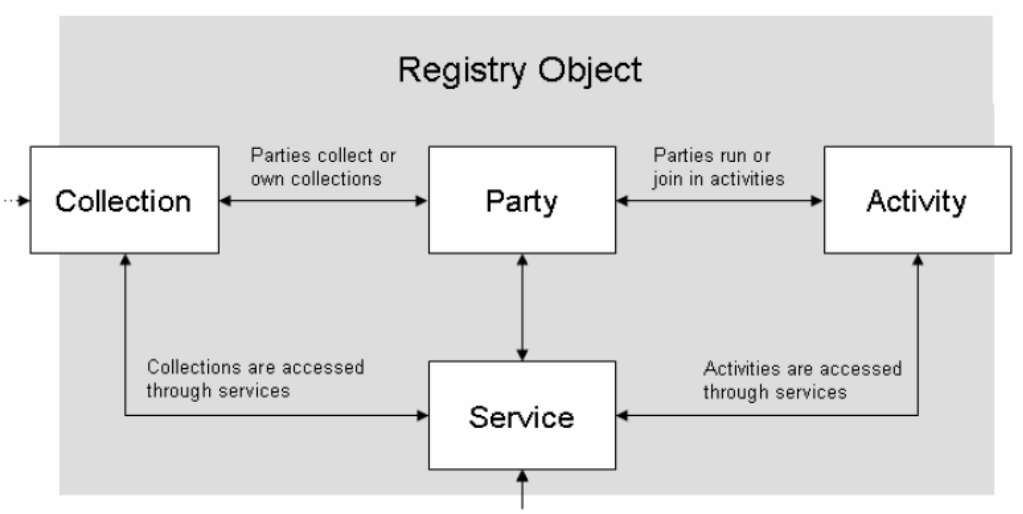

Diagram: Registry object and subclasses of the ISO 2146 conceptual model

The identification of intended uses for the data published in the registry service is a key aspect here, because it largely determines the functional models of data management and therefore the formats to be applied. The uses differ according to the types of client that are established. *SERBER* has three types of client:

- Internal users of the Ministry of Culture, Education and Sport. The directories of organizations and professionals that the Office of Library Co-ordination maintains in the areas of library cooperation and coordination will use SERBER as the single, reliable source of data.
- Providers of data on libraries and related organizations (related third-party users), and in particular the Library Services of the Spanish Autonomous Communities.
- External users unrelated to the management of the directory or the provision of its data.

In the first two cases, the goals are predictable and therefore so are the functional requirements and the developments and formats that they require. In the third case, the functional requirements are less predictable and to some extent it is important for this to be so. In the context of *SERBER*, we must consider goals such as the following:

- Providing authority files for databases and bibliographic catalogues.
- Facilitating the (re)utilization of complete records and data in applications linked to several devices, and especially for information services based on geolocation.
- Promoting the dissemination of collections and services through the use of quantitative and qualitative data.

1.3 Formats

In short, *SERBER* faces the challenge of identifying and implementing suitable data publishing formats for expressing information of a registry service for libraries and related organizations that can adapt to known and unknown needs of external users in the present and in the future.

In this report we limit the identification of data publishing formats to the framework of open data, defined as a philosophy and practice requiring certain data to be freely accessible to everyone, with no technical or legal constraints. In theory, all formats are acceptable; however, the more structured and enriched the data is, the more likely it is that it will be reutilized and that new applications that process it automatically will be designed. Put another way: not all formats are equally valuable for a specific open data project.

Our report considers three basic questions that must be answered in the order of listing:

- What are open standards?
- What specific formats meet the definition of open standards?
- What open formats are suitable for the purposes of the SERBER project?

Fortunately, we have some benchmarks for the internal and external contexts of the project: the European and national regulatory framework on interoperability and reutilization of public sector information; international and national recommendations on open data publication in the public sector; and (some) solutions applied by other registry services and directories of libraries and related organizations.

This report does not cover the nature of the computer applications to be used for data mining, which can be open source or proprietary. It is important to note that the open standards model prioritized by the national regulatory framework (as discussed in Section 2) is clearly aligned with open source software in the public sector. By contrast, private consumers are free to make decisions on the type of applications they use for data mining.

1.4 Strategic proposal

Before concluding this introductory section, it is important to note that a strategic goal of the *SERBER* project is to provide a reference model for publishing open data on libraries, professionals and similar institutions in Spain in the national portal of access to the public information catalogue, datos.gob.es, and in other similar portals at regional level. The following elements are related to this objective:

- The data provided must be completely reliable and up-to-date. The quality is guaranteed by a major effort to coordinate information in

collaboration with the regions, the National Statistics Office (INE) and other organizations and working groups.

- The data provided should be exclusive and it is therefore recommended to establish a strategic alliance and, if possible, to integrate the information with the data service currently present in datos.gob.es: the Directory of Spanish Libraries and Periodical Libraries.

2. The regulatory framework applicable to open data publication in the public sector

The answer to the question "What are open standards?" in the context of a data publishing project of the public administration must take into account the national regulatory framework on reutilization and interoperability of public sector information. In the regulatory provisions governing them, the two concepts are highly interrelated, as we shall see below.

In the field of information reutilization, the applicable regulations are Law 37/2007 of 16 November on reutilization of public sector information and Royal Decree 1495/2011 of 24 October developing Law 37/2007 of 16 November on reutilization of public sector information for the national public sector".[1] Article 2.2 of the Royal Decree states that reusable documents shall be made available to the public

> "in a structured and usable way for interested parties and preferably in raw form, in formats corresponding to open standards that can be processed and accessed automatically"

The concept "open standard" is defined in Law 11/2007 of 22 June on electronic access of citizens to public services, Annex, Point k, as follows:

> "One that meets the following conditions:
>
> - it is public and its use is available free of charge or at a cost that does not involve difficulty of access,
> - its use and application are not conditional on the payment of an intellectual or industrial property right."

A greater specification of the definition and scope of the term "open standard" is found in Royal Decree 4/2010 of 8 January regulating the National Interoperability Framework in the field of e-government".[2] The Annex to this Royal Decree defines "cost that does not involve difficulty of access" as follows:

> "Price of the standard that is linked to the cost of distribution rather than its value and therefore does not prevent its possession or use."

[1] As stated in its preamble, this Law, "which incorporates into our legal system Directive 2003/98/EC of 17 November 2003 of the European Parliament and of the Council on the reutilization of public sector information, establishes the general legal framework for the reutilization of such information."

[2] This Royal Decree develops Law 11/2007 of 22 June on electronic access of citizens to public services.

Article 11 of this regulation governs the standards applicable to technical interoperability. This article establishes that e-government documents and services that public law bodies make available to citizens or other public authorities will be available at least prioritarily through open standards (Article 11.1). Standards widely used by citizens, that is, used by almost all natural persons, legal persons and entities without legal personality that relate or are likely to relate with the Spanish public administration, are considered complementary to open standards by the Royal Decree.

This prioritization of standards has two purposes: independence in the choice of alternative technologies by citizens and public administrations and adaptability to technological progress.

Article 11.2 establishes that the exclusive use of a non-open standard without offering an alternative based on an open standard will be limited to those circumstances in which there is no open standard available that satisfies the functionality offered by the non-open standard in question and only until an open standard becomes available.

Article 11.3 establishes the criteria for selecting standards and drawing up the list of standards. These criteria include the linking of standards to certain types of normalization: open standards and standards whose cost does not involve difficulty of access; standards and technical specifications in the terms established by Directive 98/34/EC of the European Parliament and the Council; and formalized specifications. It also includes other criteria such as:

- adaptation of the standard to the required needs and functionalities;
- conditions concerning their development, use and implementation, availability of complete documentation, publication, and governance of the standard;
- conditions concerning the maturity, support and adoption of the standard by the market, its potential for reutilization, its multi-platform and multi-channel applicability and its implementation under various application development models.

As can be seen, the regulatory framework in which the development of *SERBER* is situated establishes clear criteria for selecting data publishing formats that must meet open standard requirements. In Section 3, this legal requirement is aligned with the recommendations of international experts.

Summary of aspects to be taken into account in the SERBER Project, based on the European and national regulatory framework of interoperability and reutilization of data in the public sector:

- Definition and scope of open standards in the context of European and national regulations on interoperability and reutilization of data.
- Circumstances of application of non-open standards.
- Criteria of choice between different standards.

3. International recommendations and national formats for publishing open data in the public sector

In this section we will answer two basic questions:

1. What formats meet the definition of open standards?
2. What open formats are suitable for the purposes of the SERBER project?

3.1 What formats meet the definition of open standards?

Before identifying and presenting the international and national recommendations on open data publishing in the public sector, we must consider how ISO 2146:2010 deals with them. Though the Introduction of ISO 2146:2010 (p. v) describes the data element directory as an object-oriented model that can be converted into machine-readable formats such as XML, it states that it does not intend to prescribe a specific format for data exchange between systems. Further, it states that while the international standard is not intended to replace existing standards for the exchange of registry objects between systems, an XML schema version of the data element directory can be used for this purpose. In Section 4 we will return to this question to present XML schemas developed and published by ISO 2146:2010 application profiles following the above recommendations.

The initiatives arising from the public sector include, first, the guide *Publishing Open Government Data* (Bennett and Harvey, 2009) drawn up and published at the request of the W3C, which makes the following recommendations for the selection of suitable formats:

- "The primary format for human-readable data is (X)HTML.
- Raw data is more likely to be produced using formats customized to the specific data, the tools used, or industry standards. The W3C has pioneered XML and RDF, which allow for excellent manipulation and standardized tool sets. RDF and XML files can be accessed like databases, using SPARQL, XQuery, JavaScript and many other computer languages.
- When possible, use established open standards, and tools that allow easy and efficient production and publishing of the data.
- Also keep in mind the power of linked data."

These recommendations are corroborated and, in some respects supplemented, by the five-star classification proposed by Berners-Lee (2010) in regard to the degree of implementation of open linked data:

* make your stuff available on the web (whatever format)

**	make it available as structured data (e.g. Excel instead of image scan of a table)
***	non-proprietary format (e.g. CSV instead of Excel)
****	use URLs to identify things, so that people can point at your stuff
*****	link your data to other people's data to provide context

On the basis of the five-star classification for the degree of implementation of open linked data, Cyganiak (2011) specifies some of these criteria, establishing preferences of specific formats for the public sector.

First, he recommends publishing in formats that can be automatically processed, because it allows other organizations to process, analyse and display the information according to their specific interests, and (why not?) to create new services and new ideas. He uses this criterion to establish a gradation of formats:

- Good: MS Excel, CSV, XML, JSON, Microdata
- Not so good: simple websites, MS Word
- Bad: PDF
- Really bad: charts or maps without numbers

Second, he recommends using open formats because they are suitable for tools and applications that people can use directly. These open formats include the following: CSV, KML, RSS, XML, JSON and RDF.

In the Spanish context, the various initiatives aimed at facilitating the implementation of regulations on interoperability and reutilization of data in the public sector must take into account the *Application Guide of Royal Decree 1495/2011 developing Law 37/2007 on reutilization of public sector information*. The section "How is data prepared and presented in the best format?" (P. 24), cites the following formats:

- Plain text
- XML
- CSV

- HTML (HyperText Markup Language)
- RDF/XML (Resource Description Framework / eXtensible Markup Language)
- RDF/N3 (Resource Description Framework / Notation 3)
- TURTLE (Terse RDF Triple Language)
- JSON (JavaScript Object Notation)

3.2 What open formats are suitable for the purposes of the SERBER project?

The regulatory framework for publishing open data in the public sector in Spain and the international recommendations establish a list of data publishing formats, as noted in the previous section. Owing to the characteristics of the SERBER project, in which it is essential to represent highly structured data, the list of preferred formats is reduced to three: CSV, XML[3] and RDF (with its various serialization formats).

CSV (*Comma-Separated Values*): A file format for representing data (textual and numerical) in the form of a table in which the columns are separated by commas (or semicolons) and the rows by line breaks. Text of the standard: *Common Format and MIME Type for Comma-Separated Values (CSV) Files (RFC 4180)* <http://tools.ietf.org/html/rfc4180>.

XML (*eXtensible Markup Language*): A markup language that defines a set of rules for coding documents in a format that is both legible and processable. It stems from the SGML language and allows the grammar of specific languages to be defined (just as HTML is in turn a language defined by SGML) in order to structure large documents. It is defined in *Extensible Markup Language (XML) 1.0 (Fifth Edition)* <http://www.w3.org/TR/xml> produced by the W3C and several related specifications.

RDF (*Resource Description Framework*): A family of specifications of the W3C originally designed as a metadata model. It has come to be used as a general method for the conceptual or modeled description of information that is implemented in web resources using a variety of syntax formats. Its representation model based on subject-predicate-object triplets is a cornerstone of the Semantic Web.

The transition from CSV and XML to RDF marks a qualitative leap from levels 1-3 to levels 4-5 in the classification of Berners-Lee (2010). Put another way: it marks a transition from datasets that are only related to each other ("islands") to datasets that can go beyond the framework in which they were defined and created and enter into relation with all data available on the web, with which they initially had no affinity. The publication of SERBER in CSV or XML allows

[3] The alternative, JSON, is gaining ground but is still less widely used than XML.

limited querying and the possible operations are mainly limited to the data included in the register. However, the application of the linked data philosophy allows the data to interrelate with other data available on the web, so that the directory becomes a component of a huge database in which the results of the queries and other operations are not limited to the information available in the registry.

The benefits of changing to the Semantic Web for data custodians and users are unimaginable. The downside is that adapting the registry service to the linked data model requires the application of four basic principles. The first two place us at level 4 in the classification of Berners-Lee (2010):

> "1. Use URIs as names for things.
> 2. Use HTTP URIs so that people can look up those names."

The last two principles complete level 5:

> "3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
> 4. Include links to other URIs, so that they can discover more things."

In order to apply these principles, one must make decisions and design processes and tools for the data to be published ("how one shapes and structures data to fit neatly in the Web", in the words of Heath and Bizer [2010]), and for the forms of publication. Particularly critical is the selection and use of vocabularies to describe data. RDF provides a generic, abstract data model to describe resources using *subject-predicate-object* triplets. However, it does not provide terms proper to a given domain to describe object classes of the ISO 2146:2010 model (Parts, Services, Collections, Activities) or how they are interrelated. This function is performed by vocabularies to describe data. Generally, it is preferable to reutilize one or more existing vocabularies in combination to represent all objects and relationships between objects in the conceptual model.

Heath and Bizer (2011, p. 84) recommend the following criteria for selecting vocabularies to reutilize:

1. "Usage and uptake. Is the vocabulary in widespread usage? Will using this vocabulary make a data set more or less accessible to existing Linked Data applications?
2. Maintenance and governance. Is the vocabulary actively maintained according to a clear governance process? When, and on what basis, are updates made?
3. Coverage. Does the vocabulary cover enough of the data set to justify adopting its terms and ontological commitments?
4. Expressivity. Is the degree of expressivity in the vocabulary appropriate to the data set and application scenario? Is it too expressive, or not expressive enough?"

By consulting directories of vocabularies, and especially *Linked Open Vocabularies (LOV)*, one can find vocabularies that meet the above criteria and that cover the objects and relationships between objects in the data model of *SERBER*:

AD - Address Schema
<http://labs.mondeca.com/dataset/lov/details/vocabulary_ad.html>
DIR - Directory Schema <http://schemas.talis.com/2005/dir/schema#>
RDAFRBR - FRBR Entities for RDA
<http://rdvocab.info/uri/schema/FRBRentitiesRDA>
FOAF - Friend of a Friend vocabulary <http://xmlns.com/foaf/0.1>
MADS - Metadata Authority Description Schema
<http://www.loc.gov/mads/rdf/v1>
SCHEMA - The schema.org terms in RDFS+OWL <http://schema.org>
SCSV - Schema.org CSV <http://vocab.deri.ie/scsv>
VCARD - An Ontology for vCards
<http://www.w3.org/2006/vcard/ns-2006.html>
WHOIS - Who's who description vocabulary
<http://www.kanzaki.com/ns/whois#>
Open Metadata Registry <http://metadataregistry.org>

As will be seen in Section 4, we have no examples of fully operational registry services or directories of libraries that have implemented linked data publishing formats. The most ambitious project is undoubtedly *lobid-organisations* <http://lobid.org/organisation>, linked to the North Rhine-Westphalian Library Service Centre (hbz) Linked Open Data Service. Although not yet fully operational, it is an interesting reference for *SERBER* and is widely documented through *Linked-Open-Data-Wiki des hbz*. The project, which started in 2010, has required a major investment of financial and human resources and agreement between the different bodies involved. Indeed, the *Application Guide of Royal Decree 1495/2011 developing Law 37/2007 on reutilization of public sector information* warns of the need to increase efforts and resources in order to achieve an increase in the level of implementation of linked open data according to the five-star classification.

Summary of aspects to be taken into account in SERBER according to the recommendations of international specialists:

- Several applications should be applied rather than a single one.
- Proprietary formats should be excluded in favour of open formats that can be created and handled by any software and that are free of legal restrictions.
- Structured data formats or ones that can be processed automatically in several types of application should be preferred over "fixed" data formats such as PDF.
- The formats most favoured in the sources consulted and the ones that are appropriate for the context of the directory of libraries are CSV, XML and RDF.
- Though it should be possible to use the data in any context, the formats should guarantee compatibility with specialized applications for registry services and document management.
- Without excluding the use of standards that facilitate exchange of registry objects between systems, an XML schema version of the directory of data elements of ISO 2146:2010 can be used.
- The incorporation of linked data formats requires a thorough preliminary planning to analyse and select, for example, suitable vocabularies to describe data.

4. A review of solutions implemented by other registry services and directories of libraries and related organizations

By using different sources and query strategies, we conducted an exhaustive search for registry services and directories of libraries and related organizations worldwide that publish data through exchange formats. We identified ten cases, which can be classified into three groups:

- Group A: Registries of libraries and related organizations that apply ISO 2146:2010 and have Parties (persons and organizations) as the main object of description in the conceptual model. Objects related to organizations and professionals are placed in a subordinate position. The standard refers to them expressly in "Annex C".
- Group B: Collection service registries, which apply ISO 2146:2010 and have Collections as the main object of description in the conceptual model. The standard refers to them expressly in "Annex D".
- Group C: Directories that are designed from models other than ISO 2146:2010 and therefore do not apply its conceptual model.

In groups A and B we find the ISO 2146:2010 application profiles.

The following are samples of each of the three groups.

Group A.

This group includes the WorldCat Registry <http://www.oclc.org/registry>, linked to the Online Computer Library Center (OCLC).

This is the ISO 2146:2010 application profile whose conceptual model is closest to *SERBER*. It is defined as "a free Web tool that provides a single location from which any library can view data that describes its institutional identity and services, and update it in OCLC Service Configuration. Free Web services available through the WorldCat Affiliate site distribute Registry data across the Web, enhancing Web discovery of libraries' rich content and services." (*WorldCat Registry at a glance* <http://www.oclc.org/registry/about/default.htm>).

Among the features of the registry service is the possibility of downloading records to a local or network drive as XML files for import into structured data applications and for maintenance and review offline. To this end, an XML schema has been created for the OCLC profile under the name *WorldCat Registry Schema* <http://www.oclc.org/developer/documentation/registry/xml-schemas> and can be used to generate XML records for each object in the directory. The WorldCat Registry Schema has a modular structure divided into a basic schema, *institution.xsd*, covering the data elements identifying the objects, and 11 additional components that can be mixed and combined according to the nature of the institution to be represented. Records in XML format can be downloaded by selecting "Download this Profile as XML" accompanying the description of each object in the directory. This can be seen in the Bank of Spain Library <http://www.worldcat.org/registry/Institutions/111783>.

Group B

This group includes three registry services linked to the Global Registries Initiative <http://www.globalregistries.org/index.html>: the Australian National Data Service <http://ands.org.au>; the Information Environment Service Registry <http://iesr.ac.uk> and the Ockham Digital Library Services Registry <http://www.ockham.org>.[4]

[4] During the writing of this report it has not been impossible to consult this repository.

The Australian National Data Service <http://ands.org.au> is an infrastructure for helping Australian researchers to publish, discover, access and use data resulting from research. The ISO 2146:2010 application profile prioritizes objects of collections and services, as observed in the conceptual model <http://www.ands.org.au/guides/rif-cs-awareness.html>. The description of the objects and their relationships is expressed through the RIF-CS XML schema (Registry Interchange Format - Collections and Services) <http://ands.org.au/resource/rif-cs.html>, which was developed as a data format for facilitating exchange of collection metadata among the participating institutions and for sending collection metadata to the central register (the ANDS Collection Registry). More specifically, the schema was designed in order to expose collection metadata to the central registry via an OAI-PMH data provider.

The ANDS has a modular structure divided into six components corresponding to the five objects of the conceptual model (the Registry Objects Schema, the Activity Schema, the Collection Schema, the Party Schema and the Service Schema) and an additional component dedicated to type definitions.

There are two distinct environments of access to this registry service: as a user contributing records and data through *ANDS online services* <https://services.ands.org.au/home/login.php?logout=logout&page=%2Fhome%2Forca%2Fuser%2Fcollection_add.php>; or as an end user through *Research Data Australia* <http://services.ands.org.au/home/orca/rda/index.php>. In the first case, the RIF-CS records can be viewed or downloaded for each object described. See, for example, the case of the University of Sydney Library <https://services.ands.org.au/home/orca/view.php?key=http%3A%2F%2Fnla.gov.au%2Fnla.party-531570>. In contrast, in the search environment for the end user (Research Data Australia) the records of parties (individuals and organizations) do not allow downloading of data in exchange formats. Downloads are only possible if the user accesses the record in the repository that provides the data. For example, in the Trove repository <http://trove.nla.gov.au/people/783443?c=people>, data of the records can be downloaded in formats such as EndNote, XML and BibTeX.

Information Environment Service Registry <http://iesr.ac.uk>. According to *IESR: Metadata* <http://iesr.ac.uk/metadata/>, this repository contains information about collections, services and parties; and more specifically on

- Collections of information resources, the associated services that provide access to the collections, and the parties that own the collections and/or administer the services.
- Transactional services.

The conceptual model for this profile is available at *IESR: Application Profile* <http://iesr.ac.uk/profile/index.html>. Note that the metadata for collections is based on the RSLP Collection Description <http://www.ukoln.ac.uk/metadata/rslp/schema>.

At present, records cannot be downloaded or viewed with exchange formats; however, according to the article "Linked Data 29 July 2011"

<http://iesr.ac.uk/news/>, IESR is preparing the change to linked data, and as a first step has converted the metadata into RDF triples using RDFa microformats embedded in the web pages.

Group C

Outside the sphere of application of ISO 2146:2010, we find national directories linked to agencies that assign bibliographic identification codes (especially for library loan and exchange services). The two most notable examples are the directories of agencies that assign MARC organization codes (formerly known as NUC codes) and the International Standard Identifier for Libraries and Related Organizations (ISIL).

In the case of the directories linked to MARC codes,[5] only the UK agency (United Kingdom MARC Organization Codes <http://www.bl.uk/bibliographic/marcagency.html>) offers downloading of the complete directory in PDF format.

A greater commitment to data exchange is observed in the directories linked to the code assignment agencies of the International Standard Identifier for Libraries and Related Organizations (ISIL). The ISIL Registration Authority <http://biblstandard.dk/isil/> brings together 26 of these directories created and maintained by national agencies, of which five offer some form of downloading and viewing of records in data exchange formats. These are listed below:

Germany. The directory *ISIL- und Sigelverzeichnis online* <http://dispatch.opac.d-nb.de/DB=1.2> administered by the Staatsbibliothek zu Berlin – Zeitschriftendatenbank. This directory includes all types of codes assigned to German libraries by the Staatsbibliothek zu Berlin (ISIL and MARC codes). These codes are used to identify libraries and related organizations in collective catalogues, interlibrary loan systems, etc. Records can be downloaded in tagged format using e-mail and single records or sets of records can be viewed on screen. The record fields included in the download can be selected.

Austria. The directory *Adressen-, ISIL- und Sigelverzeichnis* <http://aleph20-prod-acc.obvsg.at/F?CON_LNG=ger&func=find-b-0&local_base=acc09> is administered by the Österreichische Bibliothekenverbund und Service GmbH <http://www.obvsg.at>. Individual records or sets of records can be downloaded by e-mail or to a disk drive. The format is tagged and may include all fields of the record or a selection. If the records are downloaded to a disk drive, the extension of the file is .sav, which corresponds to an SPSS application. The character encoding is in ASCII for both types of download: ISO 8859-1 (Roman character sets) for e-mail download and Unicode/UTF-8 (non-Roman character sets) for disk drive download.

France. The directory *Répertoire des centres de ressources du SUDOC* <http://www.sudoc.abes.fr> is administered by the Agence de l'Enseignement Supérieur Bibliographique <http://www.abes.fr>. It allows download by e-mail and on-screen display of individual records or sets of records. The format is tagged and may include all fields of the record or a selection.

[5] The only service close to a directory of directories is the Library of Congress code agency, *Search the MARC Organization Codes Database* <http://www.loc.gov/marc/organizations>.

Italy. The directory *Anagrafe biblioteche italiane* <http://anagrafe.iccu.sbn.it/index.html> is administered by the Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche. Currently, it does not allow users to view or download records in data exchange formats. According to *Anagrafe delle biblioteche italiane* <http://www.iccu.sbn.it/opencms/opencms/it/main/attivita/naz/pagina_78.html;jsessionid=2DB1B10DDB6E0E569403CE646204C535>, this option applies only to the import and export of data for organizations that contribute to the maintenance of the directory. The format applied is XML, whose current schema can be found at *Anagrafe delle biblioteche italiane - Formato di scambio XML* <http://www.iccu.sbn.it/opencms/opencms/it/main/attivita/naz/pagina_365.html>.

New Zealand. The *Directory of New Zealand libraries* <http://directory.natlib.govt.nz/library-symbols-web/Home.html> is administered by the National Library of New Zealand Te Puna Mātauranga o Aotearoa <http://www.natlib.govt.nz/en/services/6docsupply.html#sect1>. It allows downloading of the full directory and of the latest updates to the directory in PDF and MS Word.

Switzerland. The directory *HelveticArchives* is administered by the Swiss National Library <https://www.helveticarchives.ch/suchinfo.aspx>. It allows individual records or sets of records to be viewed and downloaded in tagged PDF format.

The project *lobid-organisations* <http://lobid.org/organisation>, to which we referred in Section 2 of this report, deserves special mention. It is a service of *lobid* <http://lobid.org> ("Linking Open Bibliographic Data"), which describes itself as an international directory of libraries and related organizations that follow the principles of linked data. The conceptual model for this record does not follow the dictates of ISO 2146:2010, though it has profound structural similarities. A prototype of the technical infrastructure for the publication, description and updating of bibliographic data was designed between April and August 2010. At this time it was detected that, in addition to their activities, services and collections, libraries and related organizations needed to be described in the framework of the open data infrastructure. Following the principles of linked data, two processes were planned:

- Identification of objects using appropriate URIs for linked data (HTTP URIs). The URIs are based on the International Standard Identifier for Libraries and Related Organizations (ISIL), which may also act as a MARC Organization Code.
- Provision of useful information for the URIs consulted by means of standards. Initially, data were incorporated from two sources: the address database for German libraries and the MARC Organization Codes Database. For the maintenance and updating of the data, the administrators ruled out the options of receiving them from the libraries and institutions, centralizing data transfer to the registry and providing editing options to third parties. The solution chosen was to add data directly from the institutional webs through an RDF description model embedded in the HTML content: RDFa.

As can be seen, the solutions implemented show fundamental differences from those of the *SERBER* project:

- A different conceptual model.
- A different model for collection, updating and maintenance of the data.

However, this project serves as an example of the processes that must be carried out and the decisions that must be made in order to change from publishing data for external consumption to implementing the principles of linked data. These processes and decisions involve in particular the following aspects:

- The system of identification of all components of the conceptual model (objects and relationships between objects) through appropriate URIs.
- The modelling of objects and relationships between objects in the form of ontology.
- The selection and application of vocabularies to describe data.

With regard to the vocabularies to describe data applied, the administrators highlight the problems that arise: in the trial version (*Howto – Describing libraries, their collections and services in RDF* <https://wiki1.hbz-nrw.de/display/SEM/Howto+-+Describing+libraries%2C+their+collections+and+services+in+RDF>, as many as five different vocabularies have to be combined in order to represent the metadata of all objects and relationships of the conceptual model:

- *"*vcard *for representing address and contact information.*
- Good Relations *for representing information about services such as opening hours.*
- DC Terms
- *For collection description, the* Dublin Core Collections Application Profile, *including the* Collection Description Terms *and* DCMI Type Vocabulary.
- Organization Ontology *for representing hierarchical relations between organizations and their units."*

An example of the type of registry for downloading linked data is <http://lobid.org/organisation/DE-605/about>, which applies the formats RDFa, Turtle and RDF/XML.

Summary of the aspects to be taken into account in SERBER based on the experiences of other registry services and directories of libraries and related organizations:

- Registry services linked to ISO 2146:2010 show a preference for the development of application profiles of the standard based on XML schemas, following the recommendation of the standard.
- In the experiences linked to library coding systems, the solutions applied are not very advanced from the viewpoint of the open data philosophy, showing in particular a preference for PDF and ad hoc tagged formats.
- Experiences with linked data formats are still at the development and experimentation stage. The specific case of lobid-organisations is particularly interesting as a model of best practices in the design and implementation of processes and in critical decision making for the application of linked data principles in a registry of libraries and related organizations.

5. Description and comparison of alternatives

On the basis of the regulatory framework applicable to open data publishing in the public sector in environments of reutilization and interoperability of information (Section 2), and taking into account the international and national recommendations for open data formats (Section 3) and the solutions applied by registry services and directories of libraries and related organizations (Section 4), we make the following proposal for action.

We propose the immediate implementation of the following formats:

- CSV (Comma Separated Values). This is a format intended specifically for data organized in the form of a directory or spread sheet and is supported by many applications, including e-mail services such as Thunderbird, Gmail and Hotmail.
- XML (eXtended Markup Language). This format has more features because of its special link to ISO 2146:2010 and because it is a starting point for advanced options in the framework of the Semantic Web and library automation. This general metalanguage will be specified for the directory conditions in two ways:
    - Through the development and publication of an XML schema that will form the SERBER application profile, following the schemas for the WorldCat Registry profile, and as second choice RIF-CS.
    - Through the application of MARC XML.

From the benchmark experiences, some recommendations on the management of the XML schema can be made:

- The schema should be organized in a modular way from a component that incorporates the data items related to object identification.
- Rigor should be used in maintaining the schema, to ensure the incorporation of new objects and relationships as the application profile is extended or modified and to reflect possible changes in data elements and controlled vocabularies, etc.
- The XML schema should be clearly and correctly located on the SERBER web server.
- Value added services of the XML schema should be provided, including mapping with other benchmark application profiles (WorldCat Registry Schema and RIF-CS) and with schemas of other models such as DCMI Collections AP.

Finally, data access should be performed under the following conditions:

- Both viewing and downloading of data should be allowed.
- Downloading and viewing of both single records and sets of records should be allowed.

- Selection of data specific to a single record or a set of records should be allowed.
- Syndication of data updating in the various formats should be allowed.

The application of the RDF format in one of its forms of serialization would place the *SERBER* project within the framework of the Semantic Web by allowing data linking. However, this step requires further analysis and will be dealt with in the next section.

6. Strategic recommendations for the future: from open data publication to linked data publication

After ensuring the publication of open data as described in Section 5, the transition to the publication of linked data requires four basic data preparation processes.

- First, the conceptual model of the ISO 2146:2010 application profile on which SERBER will be based must be specified. This conceptual model must fully define and represent the objects and the relationships between the objects. This point is dealt with in the information and functional analysis model defined in the framework of this project [RDF]
- Second, the system for identifying the types of object represented in the conceptual model through appropriate URIs for linked data, i.e. HTTP URIs, must be established. In the case of parties, and more specifically libraries and related organizations, the basis of the URIs could be one of the international identification systems, especially the identifiers assigned by the ISIL agency, or the MARC organization codes assigned by the Library of Congress.[6] Additionally, the system of identification using URIs should be established for the other objects represented in SERBER: natural persons, activities, collections and services.
- Third, vocabularies to describe data for representing objects and relationships that are consistent with the conceptual model of the registry must be identified and selected. In this step, it is important to take into account the general rule of reutilization of existing vocabularies and the selection criteria proposed by Heath and Bizer (2011, p. 84) and presented in Section 3. It is especially advisable to consider the previous experience of the lobid-organisations project <http://lobid.org/organisation> presented in Section 5.
- Fourth, a coding scheme for the registry data must be designed in an ontology syntax, preferably RDFS. Also, the correct publication and maintenance of the schema must be guaranteed.

Having done this, one can move forward in a new dimension of data enrichment that includes designing the functional processes for creating links within SERBER

[6] A local alternative would be to use the NIDEN code assigned to libraries by the INE. The problem is that this identifier is not stable because the INE sometimes changes the codes that have been assigned.

and between SERBER and other data sets, and/or selecting and implementing one or several "recipes" for publishing linked data:

- Serving linked data as static RDF/XML files.
- Serving linked data as RDF embedded in HTML files.
- Serving RDF and HTML with client server scripts.
- Serving linked data from relational databases.
- Serving linked data from deposits of RDF triples.
- Serving linked data from existing wrapping applications or website APIs.

Progress toward a linked and enriched data model is considered a strategic objective for the MECD for the following reasons:

- It involves positioning the directory initiative within the latest trends of the Semantic Web and is a boost for such projects in the world of culture.
- There is a "market gap" for the directory project; there are few projects of this type and it may form the core for other initiatives of interest at an Ibero-American or international level.
- It will facilitate worldwide positioning of the directory and its reutilization by smart applications and by the major search engines.
- Within the General Spanish State Administration coordination in the semantic field is a function of the Office of Library Co-ordination.

7. References

Bennett, Daniel; Harvey, Adam. *Publishing Open Government Data: W3C Working Draft 8 September 2009* <http://www.w3.org/TR/gov-data>. [Consulted on 29 June 2012]

Berners-Lee, Tim: "The 5 stars of open linked data". *inkdroid: paper or plastic?*, 4 June 2010. <http://inkdroid.org/journal/2010/06/04/the-5-stars-of-open-linked-data>. [Consulted on 29 June 2012]

Cyganiak, Richard. *How to publish Open Data*. In: Opening Up Government Data. Galway, 8 Nov 2011 <http://www.slideshare.net/cygri/how-to-publish-open-how-to-publish-open-data>. [Consulted on 29 June 2012]

"Directive 98/34/EC of the European Parliament and of the Council of 22 June 1998, laying down a procedure for the provision of information in the field of technical standards and regulations on information society services". *Official Journal of the European Union*, n° L 204, of 21/7/1998, pp. 37-48.

"Directive 2003/98/CE of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information". *Official Journal of the European Union*, n° L 345 of 31/12/2003, pp. 90-96.

Directorio de Bibliotecas Españolas <http://directoriobibliotecas.mcu.es/portada.html>. [Consulted on 29 June 2012]

Directorio de Bibliotecas y Hemerotecas Españolas <http://www.bne.es/es/Servicios/DirectorioBibliotecas>. [Consulted on 29 June 2012]

Guía de aplicación del Real Decreto 1495/2011 por el que se desarrolla la Ley 37/2007 sobre Reutilización de la Información del Sector Público. 1ª edición electrónica. Madrid: Ministerio de Industria, Energía y Turismo; Ministerio de Hacienda y Administraciones Públicas, February 2012. 79 p. <http://www.datos.gob.es/datos/sites/default/files/files/PLANCISP-GRD-07_3_2.doc>. [Consulted on 29 June 2012]

Hausenblas, Michael. "5 ★ Open Data". <http://5stardata.info>. [Consulted on 29 June 2012]

Heath, Tom; Bizer, Christian. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 2011. (Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1), 136 p. <http://linkeddatabook.com/editions/1.0>. [Consulted on 29 June 2012]

ISO 2146:2010: Information and documentation: Registry services for libraries and related organizations. Geneva: ISO, cop. 2010.

"Ley 11/2007, de 22 de junio, de acceso electrónico de los ciudadanos a los Servicios Públicos". *Boletín Oficial del Estado*, nº 150, 23 June 2007, pp. 27150-27166.

"Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público". *Boletín Oficial del Estado*, nº 276, 17 November 2007, pp. 47160-47165.

Linked-Open-Data-Wiki des hbz <https://wiki1.hbz-nrw.de/display/SEM/Home>. [Consulted on 29 June 2012]

Linked Open Vocabularies (LOV) <http://labs.mondeca.com/dataset/lov/index.html>. [Consulted on 29 June 2012]

Linking Open Bibliographic Data <http://lobid.org>. [Consulted on 29 June 2012]

Peset, Fernanda, Ferrer-Sapena, Antonia, Subirats-Coll, Imma: "Open data y linked open data: Su impacto en el área de bibliotecas y documentación". *El Profesional de la Información*, 20(2), 2011, pp. 165-173.

"Real Decreto 4/2010, de 8 de enero, por el que se regula el Esquema Nacional de Interoperabilidad en el ámbito de la Administración Electrónica". *Boletín Oficial del Estado*, nº 25, 29 January 2010, pp. 8139-8156.

"Real Decreto 1495/2011, de 24 de octubre, por el que se desarrolla la Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público, para el ámbito del sector público estatal". *Boletín Oficial del Estado*, nº 269, 8 November 2011, pp. 116296-116307.

Towards the Australian Data Commons: A proposal for an Australian National Data Service. Canberra: the ANDS Technical Working Group, 2007.

8. Glossary

Conceptual model: The process and outcome of the analysis and representation of the concepts and relationships between concepts in a particular domain. The result of the conceptual model is represented by the terms of a data vocabulary.

Data exchange standard format: A specification for storing, accessing, and transmitting data. It provides the syntax for expressing datasets. It differs from the description and the semantic representation of data, which is performed by the data vocabulary. In the context of this report the terms data exchange format, data exchange standard, data exposure format and data publishing format are considered synonyms.

Interoperability: The ability of information systems—and therefore of the procedures supported by them—to share data and enable exchange of information and knowledge with each other.

Linked Data: Refers to a set of best practices for publishing and interlinking structured data for access by both humans and machines via the use of the RDF family of syntaxes (e.g. RDF/XML, N3, Turtle and N-Triples) and HTTP URIs. Linked data can be published by a person or organization behind the firewall or on the public Web. If linked data is published on the public Web, it is generally called Linked Open Data.

National Interoperability Schema: This includes criteria and recommendations on security, standardization and storage of information. It also includes information on the formats and applications that must be taken into account by public administrations to ensure an appropriate level of organizational, semantic and technical interoperability of the data, information and services managed in the exercise of their powers and to prevent discrimination against citizens on the grounds of choice of technology.

Open Data: A philosophy and practice requiring that certain data are freely available to everyone, without restrictions from copyright, patents or other mechanisms of control. To be distinguished from the more formally defined terms open source<b> and open standard<b>, open data emphasizes access to and reutilization of scientific and government data as a means to broaden collaboration, to create government accountability to citizens, and to accelerate the pace of discovery and innovation

Open Government Data: Open data generated in the activities of public administration bodies.

Open standard: A standard that meets the following conditions: it is public and its use is available free of charge or at a cost that does not involve difficulty of access; its use and application are not conditional on the payment of an intellectual or industrial property right.

RDFa: A system for integrating RDF data on websites coded in (X)HTML.

Registry for libraries and related organizations A set of registry objects that have been compiled to support the activity or business of a particular community

Registry object: The main class of the ISO 2146:2010 data model, which includes four different types of object: Activities (operations that occur in time and have one or more outcomes); Collections (sets of physical or digital objects); Parties (persons or organizations); and Services (systems that provide users with some functionality).

Reutilization: The use of documents held by governments and public sector bodies by natural or legal persons for commercial or non-commercial use, provided that such use does not constitute a public administration activity. This concept excludes document exchange between governments and public sector bodies in the exercise of the public functions assigned to them.

Semantic Web: An extension of the current Web that offers an easier way to find, share, reutilize and combine information. It is based on machine-readable information and is built on the capacity of XML technology to define customized tagging schemas and on RDF's flexible approach to representing data. The Semantic Web provides common formats for data exchange compared with the current Web, which is limited to the exchange of documents. It also provides a common language for representing how the data is related to real-world objects, allowing a person or a machine to start from a database and then move through an unending set of databases that are not connected by cables but rather by the fact that they refer to the same thing.

Vocabularies to describe data: Sets of terms describing concepts and relationships between concepts, usually in the context of a particular domain. They include metadata schemas and ontologies (which maximize the contextual relationships between concepts, the semantic restrictions, etc.)

The following sources were used to draw up the glossary:

- *ISO 2146:2010: Information and documentation: Registry services for libraries and related organizations*. Geneva: ISO, 2010.
- "Ley 11/2007, de 22 de junio, de acceso electrónico de los ciudadanos a los Servicios Públicos". *BOE*, nº 150, 23 June 2007, p. 27150-27166.
- "Ley 37/2007, de 16 de noviembre, sobre reutilización de la información del sector público". *BOE*, nº 276, 17 November 2007, p. 47160-47165.
- *Linked Data Glossary: W3C Editor's Draft 26 April 2012* <https://dvcs.w3.org/hg/gld/raw-file/default/glossary/index.html>
- "Real Decreto 4/2010, de 8 de enero, por el que se regula el Esquema Nacional de Interoperabilidad en el ámbito de la Administración Electrónica". *BOE*, nº 25, 29 January 2010, p. 8139-8156.
- *W3C eGovernment (Printable) Glossary* <http://www.w3.org/egov/wiki/Glossary>
- *Webopedia* <http://www.webopedia.com/TERM/S/Semantic_Web.html>
- Zeng, Marcia L.; Qin, Jian. *Metadata.* New York: Neal-Schuman, 2008.