Vol. 4, n. 1 (2013)

JLIS.it

Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia,
archivistica e scienza dell'informazione

Università di Firenze
Dipartimento SAGAS
Storia, Archeologia, Geografia, Arte e Spettacolo

JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione
Pubblicazione periodica semestrale (esce in gennaio e in luglio)
ISSN: 2038-5366 (print) – ISSN: 2038-1026 (online)
Website: http://jlis.it – Email: info@jlis.it

ABBONAMENTO 2013:
Italia € 50,00; Estero € 60,00.

SUBSCRIPTION 2013:
Italy € 50,00; Overseas € 60,00.

Ledizioni – LediPublishing – Via Alamanni 11
20141 - Milano - Italia
Tel. +39-0245071824 – Fax +39-0242108107 – IVA/VAT: IT04627080965

JLIS.it è edita dall'Università di Firenze, Dipartimento di Storia, Archeologia, Geografia, Arte e Spettacolo (SAGAS).

Proposte di pubblicazione & Peer-Review: Le submission, compiute tramite il sito web della rivista, verranno inizialmente esaminate da un editor e, superata la prima valutazione, saranno inviate a due revisori per il processo di peer review, al termine del quale verrà notificata l'accettazione o meno del contributo, o l'eventuale richiesta di modifiche.

JLIS.it is published by the University of Florence, Department of studies on the Antiquities, Middle Age, the Renaissance and Linguistics.

Submission & Peer-Review: Papers submitted via the journal website will be checked by one of the editors: if these pass the first step, the papers will be sent to two reviewers for the peer-review process. After this, the author will be notified on the acceptance of his paper, or will be given suggestions on how to improve it.

Global Interoperability and Linked Data in Libraries

Special issue

Editor: Mauro Guerrini
Proceedings by Gianfranco Crupi and Ginevra Peruginelli

University of Florence, Aula Magna, June 18-19, 2012

seminar

GLOBAL INTEROPERABILITY
AND LINKED DATA IN LIBRARIES

This issue collects the papers delivered at the Seminar Global Interoperability and Linked Data in Libraries, held on 18th and 19th June, 2012 at the University of Florence. With respect to the conference program (http://www.linkedheritage.org/linkeddataseminar), however, not all the conference papers are present (those of Paola Mazzucchi, Roberto Moriondo and Romano Nanni are missing), while the issue includes three more contributions (Barbara Tillett, Mauro Guerrini and Tiziana Possemato, Gianfranco Crupi). Moreover, papers are presented in a different order with reference to the conference program.

The printing of this volume was partly financed by the FIRB 2009 Project (RBNE07C4R9_004).

Table of Contents
Vol. 4, n. 1 (Gennaio/January 2013)

Introduction

Linked Data as a new Paradigm of Data Interconnection

Publishing Value Vocabularies and Standard as Linked Data

Towards a Web of Data: Applications and Experiences

Introduction

Opening and welcome

Alberto Tesi

On the behalf of the University of Florence it is a great pleasure to welcome you all to the international seminary Global interoperability and linked data. First of all, I would like to thank the organizers, the many supporting institutions - Istituto centrale per il catalogo unico delle biblioteche italiane (ICCU), Biblioteca nazionale centrale di Firenze (BNCF), Casalini Libri, Comune di Firenze, Conferenza dei rettori delle università italiane (CRUI), Associazione italiana biblioteche (AIB), Istituto di teoria e tecniche dell'informazione giuridica del Consiglio nazionale delle ricerche (ITTIG-CNR), Fondazione Rinascimento digitale – and the chair, professor and colleague Mauro Guerrini, for giving us the opportunity to discuss about the very important and very timely theme of open data. Just few weeks ago I was representing the Italian rectors to a meeting of EUA (European University Association) in Bruxelles where the focus was on open science, which of course is something more general than open data and open access. I was impressed to hear about the story of Tim Gowers, a mathematician which posed on his blog an unsolved problem. He was contacted by many other researchers and received almost one thousand substantive contributions which finally made it possible to solve the problem in only one month. It is clear that to make such processes a reliable way to approach scientific developments, it is necessary that the scientific and aca-

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-7441

demic community will overcome several obstacles. I believe the most important ones are probably coming from inside the universities, such as the criteria for credit and promotion of researchers. On the other hand, it is also clear that these processes are potentially able to provide an enormous increase of knowledge and to allow for solving many demanding problems in all the scientific areas. In any case it is very important that open science, and hence open data, would become a more central theme of the scientific community. I wish that today you will exploit this seminary also to discuss about fundamental issues, such as infrastructures for connecting and linking data and procedures to certify quality of data, which I believe are at the basis of the development of a real open data. Let me conclude this brief welcome by thanking very much again the organizers, the speakers, all the colleagues and students that are attending this seminary. I really hope that you will enjoy the program lectures as well as your staying at the University of Florence.

Alberto Tesi, Rettore, Università degli Studi di Firenze.
rettore@unifi.it

Tesi, A. ”Opening and welcome”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #7441. DOI: 10.4403/jlis.it-7441. Web.

Published: 2013-01-15

Global interoperability and linked data in libraries: ICCU international committment

Rosa Caffo

The digitization of cultural heritage issues even new challenges for accessing and sharing the content on the web. The Central Institute for the Union Catalogue of the Italian Libraries (ICCU), which is involved in the digitization of cultural heritage, participates in the European debate on open data to develop new approaches for the valorization of cultural heritage within some major European projects such as Europeana and Linked Heritage, and national initiatives as CulturaItalia, the Italian culture portal.

This conference is organized in cooperation with the Linked Heritage project, funded through the ICT Policy Support Programme, and coordinated by ICCU. Linked Heritage involves 20 European countries and dozens of European cultural institutions in each sector (archives, libraries, museums). Its main objective is to provide 3 million data to Europeana; the first ones will be online in September 2012. In addition, the Linked Heritage partners are working on three areas considered crucial for the development of cultural content on the web: multilingualism - with the aim to supply cultural institutions with a technology platform that facilitates the production of multilingual metadata that can be easily found in Europeana and in

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-8726

the web; the participation of the private sector (publishers in particular) in the context of content aggregation - traditionally developed in the public sector - with the aim to spread in the web even content protected by copyright; finally, the linked (open) data of cultural domain (LOD). This last topic is now considered crucial for all cultural institutions in Europe who want to publish their cultural resources on the web; on the other hand, concrete experiences of application are not yet widespread. Europeana has recently produced a video on linked open data in the cultural sphere in order to promote their use and reuse; the video has also been translated into Italian and is available on CulturaItalia. Linked Heritage partners are currently working on some resources already sent to the Europeana to experience possible applications of linked open data in the cultural field and to help cultural institutions to publish their data as LOD.

The possibility of connecting information offers many advantages because publishing raw data online and making connections with other data of the same type allow content providers to enrich the knowledge, improve their visibility on the web and obtain multiple benefits:

Less duplication of information. For example, those who create a dataset can directly connect it to other existing datasets created by other entities. Less work and updated data at the same time.

Higher profile. The Linked Open Data helps generate meaningful links between web pages. This helps users to discover content, which in turn raises the institutions' profile and drives the traffic to their websites.

Authority. DBpedia, FreeBase, and Project Gutenberg are often cited as sources of authoritative metadata. Cultural institutions can become authorities for information on cultural heritage.

New audience. When users analyze data and use it to create applications such as APIs and mashups, they offer your content to a new audience that couldn't be otherwise reached.

Better user experience. By providing users high quality, contextually relevant information, their viewing experience is improved and they will therefore be more likely to visit again the website of the institute.

Efficient use of resources. sharing data coming mainly from public investment means that it can be used more efficiently because users can help enrich those resources. This also means direct reuse in areas such as education, scientific research and cultural tourism.

Even CulturaItalia, the Italian culture portal promoted by the Ministry for Cultural Heritage and Activities (MiBAC) and managed by ICCU, is moving in the direction of linked open data: his system is also designed to transform the metadata provided by cultural institutions in LOD and automatically send them to Europeana. The content providers that wish to join CulturaItalia sign a license that illustrates the content (metadata) to be made available in the portal, the activities to be carried out for the interoperability of databases, and the frequency of updates. This agreement also includes the possibility to send data to Europeana under the terms of the Data Exchange Agreement (DEA) based on the open license Creative Commons 0 (licensed under public domain) aimed at the re-use of data as Linked Open Data. CulturaItalia released a new license in line with the Data Exchange Agreement of Europeana and started a plan to renegotiate the existing agreements between CulturaItalia and its content providers. The answers are not immediate because the Linked Open Data are not yet so widespread and the aggregators must in turn ask their content providers if they agree. The positive feedbacks are mainly from the library world: the libraries that participates in Internet Culturale, the portal of the Italian libraries whose data are included in CulturaItalia, were the first.

Rosa Caffo, ICCU.
rosa.caffo@beniculturali.it

Caffo, R. ”Global interoperability and linked data in libraries: ICCU international committment”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #8726. DOI: 10.4403/jlis.it-8726. Web.

Published: 2013-01-15

Introduction to the Seminar Global interoperability and linked data in libraries

Mauro Guerrini

Why a linked data Seminar? Linked data provide a new language in the world of global communication: from the public administration, banks, insurance companies, to archives, libraries and museums. In technical terms, linked data refers to a set of best practices for exposing, sharing, and connecting structured, readable, understandable and, above all, usable data. Through these mechanisms, data are linked to other external datasets, which, in their turn, can refer to other datasets. All these data create an increasingly vast network, consisting of ontologies, or logically defined groups which meaning is explicitly defined by interacting words and tags. "Linked data" and "semantic web" represent related concepts, referring to the same application domain, as linked data is a technology used for the practical realization of the semantic web. The semantic web, or web of data, is the natural evolution of the hypertext web or web of documents. The Semantic web is a container of things, objects, rather than a container of representations of objects, typical of traditional web or web of documents. In other words, the semantic web is a web in which pages have elements that can be automatically interpreted by machines. Linked data brings to the idea of concreteness: the data

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-6307

are correlated to the resource participating in its content nature; the data are an integral part of the resource because this resource would not be represented without their data. The web of data is a web in which data are published and connected with the techniques of linked data; it is a web in which each data is much more visible and, therefore, much more (re)usable and interoperable. In this context it becomes clear the need for an integrated access to data, beyond its diversity, quantity, distribution and properties. This Seminar handles primarily the issue of interoperability and openness of data in the library world, by analyzing standards, experiences and best practices for the adoption of semantic web technologies in this field. The data, produced by libraries and maintained in their catalogs are not on the web, but rather isolated from the web. Here it is appropriate to wonder: "How to modify catalogues and data so that they can be of the web and not only on the web?". Nevertheless, libraries have always produced qualitative data in highly structured bibliographic and authority records, as a response to the need of shared rules disseminated at national and often international levels. Unfortunately, today this data begin to represent nothing more than a huge amount of data, unused and unusable. The adoption of the linked data approach will involve a radical change in the creation of tools of mediation between user and bibliographic universe. All sectors of the library service will definitely be interested in the implementation of linked data. In particular, the web search, bibliographic control (with bibliographic description), access, authority control, classification, data portability and disambiguation will be strongly advantaged. The library world and the world of the internet are both interested in the integration within a global network. Firstly, there is a need to ensure a global visibility and usability of data and, secondly, to exploit information and create increasingly large and significant networks. How would it be possible to achieve the objective pursued?

- first of all, there is a need to transform the textual description in datasets usable for automatic processes and elaborations by machines;

- besides, there is a necessity to make data items uniquely identifiable within the web information context;

- then, it is important that data become compatible with the technologies and web standards;

- as well as it is of big importance to use a cross-interoperable language within the web.

This Seminar aims to highlight the benefits that the linked data can offer to libraries, providing advices on operational procedures by which they can participate in the development of the semantic web. This Seminar intends, therefore, to present an overview of initiatives and evolving techniques at the national and international levels. It also aims at facilitating the discussion in Italy about the benefits of linked data as a new communication language to reduce costs, improve services and deliver greater value to the needs of library users, through more effective and, above all, more shared design of digital information environments. Finally, this Seminar intends to encourage institutions and, in particular, libraries to invest in terms of skills and economic resources. Public institutions and individuals, which chose to open and share their data on the web, have everything to gain from the creation of highly-organized and useful information.

Acknowledgment

I would like to thank the Scientific Committee, in particular Ginevra Peruginelli and Gianfranco Crupi, and the Organizing Committee (represented by Sandra Torre and Antonella Marinaro), Maria Teresa Natale of ICCU who worked on the website, Andrea Turchini, the creator of the logo, as well as translators of reports from English and Italian, and the editorial staff of ”JLIS.it“(which has published the preprint version of the papers starting from 11 June) for their decisive support given. I also thank the co-sponsors of this Seminar, together with the University of Florence: ICCU, National Central Library of Florence, Casalini Libri, the City of Florence, CRUI, IEA, Institute of Legal Information Theory and Techniques of the National Research Council, Digital Renaissance Foundation, AIB; the sponsors: Ex-Libris, JLIS.it, Linked Heritage, AtCult, Itale e Space.

MAURO GUERRINI, Università degli Studi di Firenze.
mauro.guerrini@unifi.it

Guerrini, M. ”Introduction to the Seminar Global interoperability and linked data in libraries”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6307. DOI: 10.4403/jlis.it-6307. Web.

Published: 2013-01-15

JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

Beyond the Pillars of Hercules: Linked data and cultural heritage

Gianfranco Crupi

The term *linked data* refers to a «set of best practices for publishing and interlinking structured data on the Web. These best practices were introduced by Tim Berners-Lee in his Web architecture note *Linked Data* and have become known as the *Linked Data principles*» (Heath and Bizer).[1] The underlying paradigm is that of the traditional web, the web of hypertext or documents, focused, as we know, on a small but effective number of standards: HTML as a markup language and format for page layouts, formatting and visualization; HTTP, the universal protocol for the transmission of information in hypertext; URI, the only and universal identification system. This "simple" logical architecture is the basis of the underlying principles for publishing and sharing structured data on the web: the use of URIs to identify not only web documents and digital contents, but also objects in the real world and abstract concepts

[1] The principles formulated by Tim Berners-Lee are:

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

.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5509

(partly because URIs work as a means of access to information that describes the entities identified); the adoption of HTTP URIs, allowing URIs to be dereferenced through the HTTP protocol, in a description of the object identified or abstract concept; and finally, the use of a standard mechanism for specifying the existence and significance of the connections between the elements described in the data, provided by RDF, which, through descriptions of the relations between the "things" of the world (people, places or abstract concepts) expressed in qualified links, provides a flexible way of describing them, indicating the relationships they have with other "things" and of explicitly stating the nature of these relationships. Dereferencing means that clients can search for the URI using the HTTP protocol and thus recover a description of the resource (be it an HTML document, a real-world object or an abstract concept) that is identified by the URI; the descriptions of resources that are destined to be elaborated by machines are represented as RDF data. However, when the URIs identify "things" in the real world, in order to avoid any risk of ambiguity, confusing "things" with documents that describe them, the normal procedure is to use different URIs, thus distinguishing in a coherent manner statements about a "thing" from the document that describes it. The technology of linked data is therefore tied to the RDF model, not only because it provides the unique identification of entities on a global scale, but also because it allows for the parallel use of different schemes for the representation of data. However, at this point, we need to take a step back in order to give a theoretical and methodological context to the technology of linked data, in the light of the contributions that have been made to the *Global Interoperability and Linked Data in Libraries seminar*, the proceedings of which will be published here.

The language of the semantic web

In the context of the semantic web, the term semantic does not refer to the semantics of natural language but to the fact that the data can be elaborated by a computer, and that they contain information that allows the computer to process them correctly. Nevertheless, the semantic web has its own language, which is not a spoken language but a language invented to communicate and exchange data and information between human beings, and to be read, interpreted and processed by machines. It is a language with its own grammar, which functions to express the relational nature of the data and their proteiform typology. This grammar, known as RDF, provides the logical structure for managing and expressing the relationships between pieces of information based on the principles of predicate logic, according to which the information is expressed through statements consisting of a basic tripartite (*triple*) syntagmatic model:

1. a *subject*, i.e. any resource, not necessarily accessible via the web, which identifies the "thing" described (*documents*, readable by humans, or *objects*, readable by machines);

2. a *predicate*, that is a specific *property* of the resource or relation used to describe it, identified by a name;

3. an *object*, known as a *value*.

Furthermore, according to the grammar of RDF, every sentence or statement describes the relationship between two entities – for example, between a work and its author (Giuseppe Verdi composed *La Traviata*) – or between an entity and the textual annotations that characterize it (e.g. the words *La Traviata* and the words that indicate the date and place of its first performance: March 6, 1853, Venice, Teatro La Fenice). Nevertheless, as already stated, except for textual annotations, each element in an RDF statement is represented, in its grammar, not by words from spoken language but by strings of characters preceded by the prefix http://, which uniformly identify any resource (URI, Uniform Resource Identifier): from a web address to an e-mail address, from a document to a service, from a file to a program, etc. In the language of the semantic web, the URI also allows the use of the object identified in contexts other than the original and regardless of its textual expression.[2]

Each RDF statement can be expressed by a graph consisting of nodes and arcs that represent the resources, their properties and their respective values. To be published this graph model is encoded in *serialization* formats,[3] which allow the machine to process the model and understand the meaning of the descriptions of resources. More specifically, the identifiers used by RDF are *URI references* (URIref), or identifiers formatted by a URI, to which is added a suffix with Unicode characters, allowing it to express and define

[2] «A URI can be classified as a URL or URN. A URL is a URI that, in addition to identifying a network-homed resource, specifies the means of acting upon or obtaining the representation: either through description of the primary access mechanism, or through network "location". For example, the URL `http://en.wikipedia.org/wiki/Main_Page` identifies a resource, in this case English Wikipedia's home page, whose representation, in the form of the home page's current HTML and related code, as encoded characters, is obtainable via the HyperText Transfer Protocol from a network host whose domain name is www.wikipedia.org. A uniform resource name (URN) is a URI that identifies a resource by name, in a particular namespace. One can use a URN to talk about a resource without implying its location or how to access it. The resource does not need necessarily to be accessible over a network. For example, the URN urn:isbn:0-395-36341-1 is a URI that specifies the identifier system, i.e. international standard book number (ISBN), as well as the unique reference within that system and allows one to talk about a book, but the URI doesn't suggest where and how to obtain an actual copy of it»(Uniform Resource Identifier, in Wikipedia. L'enciclopedia libera, http://it.wikipedia.org/wiki/Uniform_Resource_Identifier, 04-12-2003; last modified 04-08-2012).

[3] "Serialization" means the process of converting a data structure into a format that can be stored and then regenerated in the same or in another computing environment.

the relationships between any things. Although the objects, which represent the values associated with the predicates, can be expressed as strings of characters (known as literals), the use of URIref allows applications to distinguish the properties that may be identified with the same literal name and which may in turn be treated as resources, allowing their additional information to be associated.

«A URI address - thanks to the way in which it is formed - contains in itself, at least implicitly, a quote. URI type addresses used for properties and classes lead the reader to definitions documented in an official manner. Thus it is the web itself that supplies the data language with its dictionary» (Baker).Tom Baker rightly insists on the linguistic nature that informs the entire system, a key to understanding the functioning of linked data and their many applications, especially in the context of cultural heritage and, in particular, libraries. In fact, it is precisely this linguistic dimension that explains the construction of multiple phrases concerning the same subject, or phrases that, in accordance with the principle of inference, generate new ones, giving rise to a network of assertions, and thus to a set of relations (according to a model derived from the logic of relational databases), which extends the semantic network of the areas of origin of the data, expressed in the individual statements.

The assimilation of the principle of combinatoriality, according to which a limited number of smaller units can be combined to form an unlimited number of larger units, thus facilitates the production of messages that contain higher levels of relational complexity and at the same time granularity relative to the domain to which the individual objects belong. It is therefore the syntagms – segments of sentences that may consist of one or more words, that constitute the statements – and the syntactic functions they assume in the sentence that encourage and facilitate the integration of data from different sources, thereby generating new connections between nodes, thanks to ontological rules based on the meaning of the properties and resources described. It goes without saying that the information potential of syntagms lies in the relationship between the predicate contained in the message, conveyed by the sentence, and the entity to which the predicate refers.

If this simple but structured linguistic system is to work correctly «a technological infrastructure must be used in which concepts are identified uniquely and in which software agents recognize these objects and realize associations and equivalences among them» (Guerrini and Possemato). This technological infrastructure consists of a set of shared tools for terminology control and semantic disambiguation, which allow one to uniformly describe data and to express their formal semantics: it is essentially a question of languages, meta-languages, controlled vocabularies and ontologies.

Languages, meta-languages, controlled vocabularies and ontologies

We are referring above all to that family of languages for representing knowledge, designed to create ontologies and intended to be processed and interpreted by machines, called the Web Ontology Language (OWL), developed by the W3C (World Wide Web Consortium).[4] With OWL, one can define and express ontologies, that is, logical structures in which the semantics of a specific domain of knowledge are encoded, which explain what we know of it through classes, relationships between classes and individuals belonging

[4] The acronym OWL, instead of the more correct WOL, was adopted by the Working Group of the W3C because it was easier to remember, partly because of its homophony in English with the name of the bird.

to classes; an automatically processable knowledge,[5] that allows for the implementation of inferential and deductive processes. In short, the purpose of OWL is the description of knowledge bases, the development of inferences about them and their integration with the content of web pages, creating a language that allows greater and better data integration between communities that describe their domains.

It is known that ontologies have a conceptual framework similar to that of a thesaurus, except that they may provide a greater number of relations, thus generating a complex network of connections between concepts, which can also be displayed graphically. Furthermore, their specific characteristic is the ability to express concepts in a non-ambiguous manner and therefore with a high level of semantic precision. «The work of harmonizing the ontologies and descriptive diagrams is entrusted to software agents which, having a representation of knowledge and rules of deduction expressed in a interoperable language, act to harmonize different kinds of knowledge.» (Signore).

Then there is the family of formal languages used to represent thesauri, classification schemes, taxonomies, subject heading systems and other types of controlled structured vocabularies that make up the Simplified Knowledge Organisation System (SKOS).[6] Once again it is an RDF application, which allows for the defining of

[5] «The modelling of reality in forms that can be analysed in accordance with fixed rules is also called *formal ontology*. In our context, the term clearly has a more applicative sense, and some philosophers turn up their noses when you use the same word to indicate it. However, there are some similarities between the two meanings: if we manage to model the structure of reality more faithfully, we will also be able to build more effective systems of knowledge organization» (Gnoli, Marino, and Rosati, p. 44–45).

[6] SKOS is a data model developed by the W3C Semantic Web Deployment Working Group (SWDWG) and adopted by many national libraries for their controlled vocabularies.

semantic relationships between concepts and that can be used as an interchange format.[7] Its flexibility allows for interaction with other tools and vocabularies used in the semantic web, such as GeoNames[8] (a geographical database that provides tools to translate geographical locations into the data that represent them: latitude, longitude, height, population, post code etc.) or Friend of a Friend (FOAF),[9] which uses the logic and philosophy of the social network to encode personal data as well as the personal relations and contacts that people establish and maintain within groups and communities into standard formats.[10]

[7] An example of a thesaurus built according to a SKOS framework is that created to support archive indexing in the UK, UKAT (United Kingdom Archival Thesaurus): http://www.ukat.org.uk. See also the ongoing project at the Biblioteca Nazionale Centrale in Florence; cf. note 27 on page 41.

[8] http://www.geonames.org.

[9] http://www.foaf-project.org.

[10] Among the converging technologies of the semantic web is that formed by topic maps, an ISO standard, which, like RDF, is «a technology based on the concept of identity. It uses symbols that represent things identifiable on the web (even if they cannot be recovered from it) in order to make statements about them.» (Topic Maps, in Wikipedia. L'enciclopedia libera, http://it.wikipedia.org/wiki/Topic_Maps 26.04.2007; last modified: 10 mar 2012). Topic maps «provide functionality made up of indexes, glossaries and thesauri, thus creating powerful mechanisms for navigating among vast collections of interconnected digital resources, where this type of interconnection does not necessarily need to be physical but may only be conceptual. This is due to the leap of abstraction that is made: these maps are not positioned on the same level as the document or resource, but are superimposable, positioned at a higher level and form a common semantic superstate to the objects to which they refer and which are "mapped". In this way, several maps can be applied to the same information or the same map may be applied to different groups of information, allowing a high level of flexibility and customization. The proposed structure is reticular and multi-layered, using a scheme that lends itself much more to the system of scientific research and ways of organizing thought, overcoming the limits of linear and tree structures imposed by the storage formats of computer media» (Meschini p. 62).

Linked data

This then is a summary of the technological and conceptual context of reference to linked data that, through RDF and the use of URIs as universal identifiers of things, put entities coming from different and ever-new data sources in natural relations and integrate them. A process – made possible by reference to shared vocabularies (that thus make the definitions of the words recoverable) and by the fact that terms from different vocabularies are connected to each other through links between the vocabularies themselves – about the choice of which there are no preliminary constraints on the part of data editors. And this on the assumption that the data are properly structured (*conditio sine qua non* of their re-usability) and are self-describing, which means that if an application finds data described with an unknown vocabulary, the application can dereference the URI that identifies the terms of the vocabulary in order to find their definition, thus allowing client applications to discover all the relevant meta-information required to integrate data from different sources. In short, the reusability of data is requested by the self-descriptive nature of linked data, in the sense that each property used to describe the relationship between two things is itself described using the same data format that describes the data (Hodson).

In the linguistic articulation of the RDF model, the logic of the links is to break the self-referentiality of the data, multiplying the relationships with other data sources that, for example, provide context information about the identity of a person or the place where he or she lives. In addition, the fact that they point to different URIs to refer to the same thing in the real world or the same abstract concept, makes it possible to document and express the polysemy and the plurality of viewpoints that exist around them. The promise of the web, modelled on the logic of linked data, is not only to allow client applications to discover new sources of data, following RDF links at run time, but also to help them to integrate data derived from these sources (Coyle, *Linked Data Tools. Connecting on the Web*).

In fact, information from a variety of sources can be easily combined by merging them into a single graph consisting of two sets of triples. However, since RDF provides only a general, abstract data model for the description of resources, integration, from the semantic point of view, occurs mainly through mapping operations, using taxonomies, vocabularies and ontologies expressed – as stated earlier – in languages and knowledge representation schemes such as OWL, SKOS and RDFS (RDF Vocabulary Description Language, better known as RDF Schema). These satisfy the need to express taxonomies, thesauri and subjects (SKOS) and to provide vocabularies to describe conceptual models, in terms of classes and their properties, as well as the subsumption relations between terms (RDFS, OWL).

Linked data and the bibliographic universe

Linked data therefore appears as an application of the principles of the web aimed at a new, more flexible data publishing paradigm. The result is a global data space – the data web – based on open standards and made up of an incalculable number of RDF statements from the most disparate sources and covering an enormous range of topics. This is the source of the success that linked data technology is beginning to have in every area of social interaction on the web and, more specifically, in the field of cultural heritage and scientific communication.[11]

[11] There are numerous examples of applications and case studies covering a wide variety of sectors Gangemi; Agnoloni et al.; Moriondo; Menduni, Vannuccini, and Innocenti.

In particular, libraries are discovering that they can integrate the structured information in their catalogues with information from other catalogues and from third parties (such as, for example, DBpedia[12]), and make it easier to access their data based through the use of web standards. The problem is that in order to be visible to the user the library catalogue must cease to be detached entity, a separate database, a "silo" isolated from the web, but must be integrated into the web, queryable from it, able to speak and to understand the language of the web, namely the language of the web users who "live" and operate on it as if it were their natural habitat, and where new players present themselves, competing to populate the universe of information mediation and to redraw the geography of knowledge and places giving access to knowledge.

The transformation of the catalogue into a system that is integrated with the technology used for research and for the creation of new ideas is possible if it emerges from that self-referential dimension that in many ways has always characterized it, to meet the needs of users, who are not necessarily limited to the elective users of the traditional catalogue, but who normally use the web as their primary source of information. This involves the development of an alternative way to use and exploit bibliographic data, able to respond more closely to the way the web operates and the rules of expanded social relations, which has embraced the philosophy of open access to sources of knowledge and, above all, to data, to their ever-changing variety, to data that are themselves relationships, which are the structural connection between things and whose combinations continuously generate new knowledge.

The key word in this process is "interoperability",[13] not merely tech-

[12] DBpedia is a collaborative project to extract and reuse semantically structured information from Wikipedia and make this information available on the web and reusable by software and applications.

[13] «Thanks to the actions of the Digital Agenda for Europe, the Guidelines for

nological but also semantic, cultural: one might say, that which arises from the encounter of different digital communities and ethnic groups, with their languages, their traditions, their different way of classifying and representing the things of the world. The world of libraries is very familiar with the concept of interoperability because it has analysed it and practiced it in recent decades. These days the problem is how to make bibliographic data useable on the web, «using the computing power that exists today as well as the computational capabilities provided by the web itself» (Coyle, "Linked Data: an evolution"). The technology offered by linked data is an opportunity of extraordinary importance, although not the only one possible. «But we cannot move into the rich and dynamic information environment of the 21st century with data that is based on 19th century principles» ("Linked Data: an evolution").

Thus, interoperability means – in this specific case – making data accessible and available, so that they can be processed by machines to allow their integration and their reuse in different applications. The pilot schemes of the Bibliothèque Nationale de France,[14] the Li-

semantic interoperability through linked open data, *Linee guida per l'interoperabilità semantica attraverso i Linked Open Data* were published. They provide a reference framework for the production of open data that is interoperable between public administrations, thus making data management in the public sector accessible and transparent» (Martini).

Martini, along with Graham Bell ("Commercial and cultural sectors: potential for data collaboration?"), underlines how within the European project Linked Heritage interesting models of interoperability are developing between metadata from the public and private sectors, which generate new services and undoubted benefits to the community of users.

[14] The Bibliothèque Nationale de France with its project data.bnf.fr provides access, through a single web interface, to digital documents in its possession and descriptive data from its various catalogues and other sources. The interoperability between the BNF's different catalogue and documentary sources and between them and those from external data sets is ensured by the adoption of the standards of the semantic web and by their expression according to the conceptual model of FRBR *Présentation*

brary of Congress,[15] the Sveriges Nationalbibliotek,[16] the Bayerische Staatsbibliothek,[17] the British Library[18] and the OCLC[19] are clear indications that the world of libraries (as well as that of archives and museums) is entering the world of the semantic web, introducing into it a solid tradition of theories and practices based on bibliographic control and control of the authority of data, as well as on sensitivity and the ability to manage information, catalogue knowledge, and create new semantic connections between documents. They are thus providing added value through the syndetic structure of the catalogues, indexical tools, the language of semantic indexing

générale du projet data.bnf.fr; Wenz.

[15] The Library of Congress has launched a project to make available, in the form of linked data and without restrictions on use, its controlled vocabularies, including a first core of classes taken from the LCC (Library of Congress Classification) (Library of Congress, LC Linked Data Service. Authorities and Vocabularies, http://id.loc.gov; Ford).

[16] The National Library of Sweden, which as early as 2008 made the Union Catalogue of Swedish libraries (LIBRIS) available in linked data mode, is now actively involved in the creation of the Open National Bibliography (Malmsten).

[17] At the Deutsche Nationalbibliothek and the Hochschulbibliothekszentrums des Landes Nordrhein-Westfalen, and the North-Rhine-Westphalian Library Service Centre (HBZ), a linked open data service has been set up (known as Culturegraph) that generates a single and specific identifier for all types of resources in the possession of German libraries with the aim of creating a catalogue of open metadata; cf. p. 42–43.

[18] The British Library is developing a version of the British National Bibliography (BNB) in the form of open linked according to a conceptual model that has been effectively represented in graphic form (http://talis-systems.com/wp-content/uploads/2011/07/British-Library-Data-Model-v1.01.pdf). The initial offering includes monographs and serial publications (British Library, Free data services, http://www.bl.uk/bibliographic/datafree.html (Hodson).

[19] The OCLC has recently made available over a million linked data resources (approximately 80 million linked data triples) regarding the most widely held works in WorldCat, chosen according to the number of localizations (at least 250) of each document. The project http://www.oclc.org/us/en/news/releases/2012/201252.htm is illustrated in a video, Linked Data for Libraries http://youtu.be/fWfEYcnk8Z8, which also serves as a concise and useful introduction to the technology of linked data.

and classification.[20] This is a sensitivity that today is translated into the design of new digital contexts and logical spaces of interaction between users and the universe of documents and services, enabling intuitive access to and easy retrieval of contents.[21] This is why it is vital that the data structured and controlled by libraries are present on the web and accessible with new tools that are compatible with web technologies and standards.

The linked data will create new services based largely on the wealth of knowledge and practices that are an integral part of the tradition of libraries, archives and museums, which have always strived to convert information into quality data and metadata. If fully harnessed, the opportunities offered by this new way of publishing data on the web, made up of linked data, will bring about a radical transformation of the relationship between the user and the bibliographic universe:

- the integration of one's own data with those of other institutions not only increases their informative potential but renders them more complete, more usable and reusable, even in contexts very different from the original;

- the explanatory clarity of the language used on the web makes the language of the library and the semantic tools it adopts for the classification and organization of knowledge less obscure and therefore more comprehensible to the user;

[20] Also worth noting is the project being launched at the Vatican Library to develop specific application profiles for managing various typologies of metadata, designed to allow access via the web to digital collections of ancient manuscripts and books (Manoni).

[21] Among the most interesting experiences from the point of view of the creation of innovative tools for the enhancement of cultural heritage are: the ITACH@ project (Innovative Technologies And Cultural Heritage Aggregation), which has created a platform for the creation and publication of linked data (Possemato), and the discovery platform developed by ExLibris (Kaschte).

- the aggregation and connection with other web resources, even if structured according to different standards, allows for the infinite extension of the context information for each item of data;
- the encounter with other segments of the web increases the number of tools available for terminological control, increasing the accuracy and relevance of information sources, whose recognized authority is the fundamental distinguishing criterion for conferring legitimacy and validity to the data;
- bringing local data out of the "deep web" and making them open and universally accessible, means offering minority cultures a democratic opportunity for visibility;
- the integration of cataloguing data in the semantic web implies enriching the catalogues and the potential to offer new services based on the technology and language of the web;
- furthermore, «the recent accord - known as schema.org - between the major search engines (Google, Yahoo, Bing and the Russian Yandex) to encode data on normal HTML pages (HTML5) in RDF language can (or should) also be an interesting opportunity for libraries. With this encoding – which looks like a very simple extension of the HTML tags of the web pages, but is based on the RDF language – the search engines are able to understand the structure and nature of a given document. With encoding based on schema.org our catalogues, thanks to the structured data they contain, can be "semantic objects" able to be interpreted by the major search engines» (Bergamin and Lucarelli).

The quality of a library is measured not so much by the number of documents held as the ability to structure and model the data and make them accessible while maintaining the stratification of contexts, the relationship between the new one that is created and the context of origin, as well as all other documents with which they form semantic relationships, whether implicit or explicit. That is to say, it must be able to reconstruct the logical and genetic relations between documents, while making them available to new semantic shifts, left totally to the users judgment and choice: in other words, they must know how to exhibit the multiple contexts to which the documents refer. Hence the need to work – as they are currently doing – to make their data uniquely identifiable in the context of the web and to make them available to be read, interpreted and used by machines. The international community of librarians is already acting, creating – as mentioned above – important projects to transform and adapt their catalogues.

The experience of the Bibliothèque Nationale de France leads us to think that the catalogues and bibliographical data of the near future will have a very different form and function than those of today: an encyclopaedia-catalogue, which displays all possible relationships between the data contained within it and those recovered from other sources and that becomes itself elaborated knowledge and a primary tool of reference. A similar effort is being made by national and international organizations (lead by the IFLA) to try to translate bibliographic and classification schemes such as

ISBD,[22] FRBR,[23] RDA,[24] DDC,[25] LC Classification[26] and the Nuovo soggettario italiano[27] into linked data. In each case they are delicate operations that affect the logical architecture of complex documentation and regulatory systems, and that pose significant problems of systemic consistency, particularly as regards linguistic choices and data rights management. The first aim to safeguard and ensure multilingualism and linguistic and cultural diversity[28] with actions (as in the case of ISBD) that are geared towards the adoption of opaque URIs, expressed in figures, since «the declarations [of the URI] contain important information such as metadata name, label, definition, notes used for extending the information or its applica-

[22] The IFLA ISBD Review Group has recently acted with the aim of «improving the portability of bibliographic data in the semantic web and consequently the interoperability of the ISBD standard in connection with other content standards» IFLA p.1; Escolano Rodriguez.

[23] One of the main objectives of the FRBR Review Group is to promote the IFLA standard and take part in the creation of namespaces for all bibliographical standards (including ISBD, FRBR, FRAD, FRSAD) «and in connection with this promote and position the IFLA standards and models in the semantic web» (Action Plan for 2012, http://www.ifla.org/en/node/1959; cfr. Riva).

[24] One of the stated objectives of the Joint Steering Committee for Development of RDA (Resource Description and Access), the new standard that replaces the AACR2 cataloguing, is to make the data «adaptable to new and emerging database structures» Joint Steering Committee for Development of RDA; Danskin; Tillett.

[25] In 2009, the OCLC was already committed to publishing the Dewey Decimal Classification as a controlled vocabulary of linked data. The initiative is still in progress (Mitchell and Panzer).

[26] Cf. note 15 on page 37.

[27] Since November 2010 the Nuovo soggettario from the Biblioteca Nazionale Centrale of Florence has made its metadata available in the RDF/SKOS format, in order to improve their "usability" in the world of Linked data (Bergamin and Lucarelli).

[28] On the efforts being made in the European Community to develop a TMP (Terminology Management Platform), cf. Leroi ("Linked Heritage: a collaborative terminology management platform for a network of multilingual thesauri and controlled vocabularies").

tion, the affiliation (whether it is property or sub-property), the state of acceptance, etc. [...] Using an opaque URI and specifying the language in which you desire to obtain the information, it is possible to collect all declarations in different languages with the same URI [...] An opaque URI would also extend its use to linguistic communities different from the English ones ensuring, at the same time, access to these ontologies in other languages without the necessity of creating independent URIsi» (Escolano Rodriguez). As regards the second – the choices relating to data rights management – these are conditioned both by the level of control that the publisher of the data wishes to exercise, and by their intrinsic nature and typology. In general, they pose a problem of legal interoperability[29] as regards the integration of data from different sources (public and private), which obviously could be attained through the development and harmonisation of national legal frameworks in the field of public data, and the adoption of suitable licensing schemes, which currently fall into two classes: «*Open licence* – This allows any use of the data, especially including commercial use, sometimes with restrictions about attribution and misuse. *Not-open licence* – This restricts uses to non-commercial only, with similar requirements for attribution and misuse. With both classes there are a range of standard licences, e.g. those provided by Creative Commons and GNU, and the option of a specific organisational licence» (McKenna). The German experience is significant in this respect: in the Bavarian and Berlin-Brandenburg library networks an interesting debate is taking place on the legal aspects of open data and, in particular, on the publication of all or part of bibliographic records in the form of open or linked open data. This has led to the decision to publish the most complete records possible, with the exception of URLs

[29] «Legal interoperability could be defined as the possibility of legally mixing data from different sources (including governmental data, data generated by online communities and data held by private parties)» (Morando).

linked to indices supplied by commercial service providers, which cannot be published for reasons of copyright. Nevertheless, there are those within the library community who argue that records in fields that have significant production costs, such as those regarding the semantic indexing of documents, should not be made available free of charge (Messmer). As previously stated, the semantic web is a very heterogeneous information environment that naturally tends towards the hybridization and contamination of contents and data from different sources. On the one hand, this is a limitation for the library world, which needs to pay attention to the quality and authority of information sources, and to defend the legitimacy of their terminology and linguistic tools for the formal control of the data. On the other hand, the integration of data that is selected, structured and homogenous with the often unstructured data from very heterogeneous information environments (scientific research, business, government, community crowd-sourced, etc.), is a challenge that libraries must face, «on pain of death for catalogues, abandoned by users in favour of other information retrievaltools, such as search engines» (Guerrini and Possemato). Although, even in the face of the exponential growth in digital resources, it is undeniable that alongside the objectives of the Linked Open Data project[30] (that is, to render the data accessible in non-proprietary formats, linking to other datasets that serve to disambiguate the content and give them a semantic context) there is a need to guarantee the quality of the data and their sources, particularly with regard to the requirements

[30] Linked Open Data (LOD) promotes the availability of data from public and private, institutional and commercial sources in order for it to be as open as possible to every kind of application and thus reusable in contexts other than the original. Open data is the infrastructure that linked data need to create the network of inferences between the data scattered across the web. Public administration, education, infrastructure and research are just some of the potential areas where access to data can bring benefits and open new opportunities (Bauer and Kaltenböck).

for integrity and authenticity (Lunghi, Cirinnà, and Bellini). The use of persistent identification systems is certainly the most convincing solution (and the linked data "movement" is well aware of this), as it can ensure the long-term usability of the data and their effective interoperability (Brase). This requires the choice of the appropriate technology and the adoption of authoritative certification and accreditation systems (even at a non-institutional level) by the user communities that adopt them. However, because open linked data are becoming a common part of librarians' sphere of scientific tools and professional practices it is necessary that, as has been noted, this new and different method is viewed as an opportunity for libraries and not as an obstacle to their growth: «Linked Data becomes more powerful the more of it there is. Until there is enough linking between collections and imaginative uses of data collections there is a danger librarians will see linked data as simply another metadata standard, rather than the powerful discovery tool it will underpin» (Byrne and Goddard). Michele Barbera has pointed out that to overcome the current limitations in data reuse within the scientific community and in the field of cultural heritage, there needs to be a cultural change in the way we produce, manage and disseminate data, allowing space for the unpredictability that can generate new insights and new ways to exploit the information (Barbera).

Stop hugging your data (Berners-Lee): this was the title of a lecture by Tim Berners-Lee, who a few years ago invited everyone to make their data available and bring them out of the silos in which they were stored and sealed, rather than build better and more efficient silos. We now know that the invitation made sense. Data acquire value as knowledge when they are interconnected with other data, when their interconnection produces explosive web effects. And the Copernican revolution of linked data is the fact that the link, an instrument for connecting documents on the traditional web, in the context of the semantic web acquires a primary semantic role, a predicate function that gives meaning to the data themselves, because it expresses the different types of relationships that they can have. This is a revolution that implies – as we have seen – the division of information into individual atomic components, into fragmented units, that can be recombined with different functions and for different purposes. These principles, which constitute the paradigm of linked data, when applied to the world of cultural heritage, modify (as some exemplary experiences have proved) the cognitive processes that have hitherto governed our relationship with the bibliographic universe and with the tools that have historically mediated the relationship between reader and knowledge (catalogues, records, index systems etc.). This is based on the idea that a vision of the world is possible only if one starts from the awareness that knowledge is a dynamic process, the continuous putting together and taking apart of what we discover and know about the world.

References

Agnoloni, Tommaso, et al. "Schema di annotazione per la dottrina giuridica: il caso di studio della banca dati DoGi Dottrina Giuridica". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5480. (2012). (Cit. on p. 34).

Baker, Thomas. "Designing data for the open world of the Web". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-6308. (2012). (Cit. on p. 29).

Barbera, Michele. "Linked (open) Data at Web Scale: research, social and engineering challenges in the Digital Humanities". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-6333. (2012). (Cit. on p. 44).

Bauer, Florian and Martin Kaltenböck. *Linked Open Data: The Essentials. A Quick Start Guide for Decision Makers*. Vienna: DGS, 2012. http://www.semantic-web.at/LOD-TheEssentials.pdf. (Cit. on p. 43).

Bell, Graham. "Commercial and cultural sectors: potential for data collaboration?" *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5487. (2012). (Cit. on p. 36).

Bergamin, Giovanni and Anna Lucarelli. "Il Nuovo soggettario come servizio per il mondo dei linked data". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5474. (2012). (Cit. on pp. 39, 41).

Berners-Lee, Tim. *Stop hugging your data*. 2009. http://www.youtube.com/watch?v=OM6XIICm\protect_qo. (Cit. on p. 44).

Brase, Jan. "DataCite and Linked data". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5493. (2012). (Cit. on p. 44).

Byrne, Gillian and Lisa Goddard. "The Strongest Link: Libraries and Linked Data". *D-Lib Magazine* 16.11-12. DOI: 10.1045/november2010-byrne. (2010). (Cit. on p. 44).

Coyle, Karen. "Linked Data: an evolution". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5487. (2012). (Cit. on p. 36).

—. *Linked Data Tools. Connecting on the Web*. Chicago: ALA Editions, 2012. (Cit. on p. 34).

Danskin, Alan. "Linked and open data: RDA and bibliographic control". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5463. (2012). (Cit. on p. 41).

Escolano Rodriguez, Elena. "L'adattamento di ISBD al web semantico dei dati bibliografici espressi in linked data". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5484. (2012). (Cit. on pp. 41, 42).

Ford, Kevin. "Library of Congress Classification as Linked Data". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5465. (2012). (Cit. on p. 37).

Gangemi, Aldo. "Semantic technologies and linked data for the Italian PA: the case of data.cnr.it". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5457. (2012). (Cit. on p. 34).

Gnoli, Claudio, Vittorio Marino, and Luca Rosati. *Organizzare la conoscenza. Dalle biblioteche all'architettura dell'informazione per il Web*. Milano: HOPS– Tecniche nuove, 2006. (Cit. on p. 31).

Guerrini, Mauro. "Saluto e introduzione ai lavori (Seminar "Global Interoperability and Linked Data in Libraries", Firenze 18-19 giugno 2012)". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-6307. (2012).

Guerrini, Mauro and Tiziana Possemato. "Linked data: a new alphabet for the semantic web". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-6305. (2012). (Cit. on pp. 30, 43).

Heath, Tom and Christian Bizer. "Linked Data: Evolving the Web into a Global Data Space". *Synthesis Lectures on the Semantic Web: Theory and Technology* 1.1. DOI: 10.2200/S00334ED1V01Y201102WBE001. (2011): 1–136. (Cit. on p. 25).

Hey, Tony, Stuart Tansley, and Kristin Tolle, eds. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft Research, 2009. http://research.microsoft.com/en-us/collaboration/fourthparadigm.

Hodson, Tim. "British Library Data Model: Overview". *Talis Systems*. (2011). <http://talis-systems.com/2011/07/british-library-data-model-overview>. (Cit. on pp. 33, 37).

IFLA. *ISBD: International standard bibliographic description*. Consolidated ed. Berlin: De Gruyter Saur, 2011. (Cit. on p. 41).

Joint Steering Committee for Development of RDA. "Resource Description and Access. Frequently Asked Questions". (2010). <http://www.rda-jsc.org/rdafaq.html>. (Cit. on p. 41).

Kaschte, Axel. "Linked open data on its way into next generation library management and discovery solutions". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5492. (2012). (Cit. on p. 38).

Leroi, Marie-Veronique. "Linked Heritage: a collaborative terminology management platform for a network of multilingual thesauri and controlled vocabularies". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5471. (2012). (Cit. on p. 41).

Lunghi, Maurizio, Chiara Cirinnà, and Emanuele Bellini. "Trust and Persistence for Internet resources". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5494. (2012). (Cit. on p. 44).

Malmsten, Martin. "Cataloguing in the open - the disintegration and distribution of the record". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5512. (2012). (Cit. on p. 37).

Manoni, Paola. "Metadata framework and application profiles in the global structure of catalogs and digitization projects of the Vatican Library". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5516. (2012). (Cit. on p. 38).

Martini, Patrizia. "Standard bibliografici e Linked Data. Verso una collaborazione tra il settore culturale e il settore commerciale". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-8598. (2012). (Cit. on p. 36).

McKenna, Gordon. "Linked Heritage Experience in Linking Heritage Information". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-6304. (2012). (Cit. on p. 42).

Menduni, Giovanni, Gianluca Vannuccini, and Giacomo Innocenti. "Open Data nella Pubblica Amministrazione italiana: l'esperienza del Comune di Firenze". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5500. (2012). (Cit. on p. 34).

Meschini, Federico. "Le mappe topiche. Come imparai a non preoccuparmi e ad amare i metadati". *Bollettino AIB. Rivista italiana di biblioteconomia e scienza dell'informazione* 45.1. (2005). <http://bollettino.aib.it/article/view/5389>. (Cit. on p. 32).

Messmer, Gabriele. "Linking library metadata to the web. The German experiences". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5507. (2012). (Cit. on p. 43).

Mitchell, Joan S. and Michael Panzer. "Dewey linked data: Making connections with old friends and new acquaintances". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5467. (2012). (Cit. on p. 41).

Morando, Federico. "Legal interoperability: making Open Government Data compatible with businesses and communities". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5461. (2012). (Cit. on p. 42).

Moriondo, Roberto. "Open Data e politiche pubbliche". *JLIS.it* 3.2. DOI: 10.4403/jlis.it-6313. (2012). (Cit. on p. 34).

Possemato, Tiziana. “Linked data: re-thinking knowledge. How software producers can accompany this revolution”. *JLIS.it* 3.2. (2012). (Cit. on p. 38).

Présentation générale du projet data.bnf.fr. 2011. (Cit. on p. 36).

Riva, Pat. “FRBR Review Group initiatives and the world of linked data”. *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5482. (2012). (Cit. on p. 41).

Signore, Oreste. “Semantic web: il futuro è già qui?” *JEKPOT The joint between Economy Knowledge Pathos Organization and Technology, 10° Knowledge Management Forum, 24-25 Novembre 2005*. 2005. http://www.w3c.it/papers/km10.pdf. (Cit. on p. 31).

Tillett, Barbara. “RDA and the Semantic Web, Linked Data Environment”. *JLIS.it* 3.2. DOI: 10.4403/jlis.it-6303. (2012). (Cit. on p. 41).

Wenz, Romain. “Linked open data for new library services: the example of data.bnf.fr”. *JLIS.it* 3.2. DOI: 10.4403/jlis.it-5509. (2012). (Cit. on p. 37).

GIANFRANCO CRUPI, Sapienza Università di Roma.
gianfranco.crupi@uniroma1.it

Crupi, G. ”Beyond the Pillars of Hercules: Linked data and cultural heritage”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5509. DOI: 10.4403/jlis.it-5509. Web.

ABSTRACT: The purpose of Linked Data is to develop a total data space (the data web) able to mutually connect and enrich shared databases. Libraries therefore have the opportunity to integrate the structured information of their catalogs with information from other multiple sources and to make them more accessible by building them on web standards. The ability model the data, making them accessible and preserving the contextualization is proposed as a criterion for determining the quality of a library. The article deals with the essential articulation of semantic web and its application in the universe of libraries, and the opportunity to use shared languages, metalanguages, controlled vocabularies and ontologies that are able to meet the need for automatic processing.

KEYWORDS: Cultural heritage; Library linked data; Semantic web

Submitted: 2012-07-05
Accepted: 2012-08-31
Published: 2013-01-15

Linked Data as a new Paradigm of Data Interconnection

JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

Library linked data: an evolution

Karen Coyle

I am particularly pleased to be here in Florence as part of this seminar on library linked data. I say "particularly" because it was here in Florence, I believe about four years ago, at another conference, where I made an attempt to present these new ideas about linked data, but without great success. In the intervening years I have learned much more about this topic, and at the same time the concepts of the semantic web have spread throughout the information communities, including those of the sciences and the cultural heritage institutions. We are here today to continue our support of this evolutionary development, not only for libraries but for all users of the web who are or who could also be library users. My goal today is to introduce certain basic concepts that will help to provide a context for the remainder of this meeting. It is not uncommon when discussing a technical topic like the semantic web to focus on particular details, yet for us here today it is essential that we steer our discussions toward areas that are particularly important for our community, especially in these times. To understand our future we must of course know our past. In the case of libraries, our past is long and I could only give a nod to the centuries of experience and tradition that have brought us here. Many of the presentations that you will hear in these days will give a vision of our future. Therefore, in the few minutes that I have I would like to speak neither of the past

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5443

nor the future, but of the present. With this I hope to provide some context that will allow us to connect our past and our future.

Our world today:

- is increasingly experienced through computers and devices, like cell phones and iPads, that are connected to the net;
- it is enormously interactive; everyone can create (albeit perhaps only a personal Facebook page), can interact, can be seen and heard;
- the world is pluralistic in terms of culture, politics and economics; in the analog world power may be concentrated in 1% of the population, but a blogger who belongs to the 99% could have millions of followers and a significant influence.

Our information resources:

- are either born digital or are being digitized;
- are relatively easily accessible throughout the global network but are also costly to use because they require advanced technology, such as devices, wires and reliable electricity, but also familiarity with this technology;

Today's users:

- expect to do their research and interact with information without prior training, preferably using a single search box;
- interact with the library through software and hardware that is not under the library's control;
- To today's users "access" means "obtain a copy," and "obtain a copy" means that the resource is removed from the organizational context of the library or the database or the web site;

every user has a hard drive full of documents that have no particular organizational context.

Communication today is significantly different from only two decades ago:

- communication is not face-to-face but across distances; if you see two youngsters side-by-side, each sending text messages on a cell phone, there is a good chance they are sending messages to each other;
- communication is becoming faster and shorter; it takes years to write a book and weeks to read it; it takes hours to write a blog post and minutes to read it; it takes less than a minute to write a tweet and only seconds to read it;
- communication today is based on interaction; one can comment on a blog or respond to a Tweet, or even comment on newspaper articles; a text message is a single entry in a continuous communication; today's youngsters would probably be more at home with a Socratic dialog than with the fixed, inactive, printed book;
- at the same time that the printed word is waning in influence, the use of other media, such as photos and videos, is increasing; these are used not only as mass media but today also as individual communication; and not only as entertainment but as the primary means of instruction - instead of the instruction manual that once came in the box with the purchase of software you now find online "how to" in video form. YouTube receives 60 hours of video every minute;
- communication that in the past was informal and un-captured, like a hallway conversation or a classroom discussion, now

may be fixed in a digital form. We have come to treat these formerly informal communications as equal to traditional records, using them in the courtroom or even as the basis for research;

To summarize, the world today is online and interactive; communication that is informal but in digital format now is included in our historical record; the printed word is from another time. Print will not disappear, but it is clear that it is no longer to be considered a modern technology. The web has changed everything. Libraries must confront this change; it is a matter of life or death, existence or disappearance. An institution based on the pre-web civilization cannot be relevant, and we cannot assume that such an institution will continue to exist. So, what is the state of libraries today? Our libraries contain a huge cultural heritage. To organize this cultural heritage and to make it available and useful to the public is a complicated and costly endeavor. But the big problem for libraries today is not just the curation of the past; the present provides a huge challenge. Not only has the number of printed books increased in recent years, while the financial support for libraries has decreased, but as we learn with the example above of YouTube, every minute an untold number of new resources is added to our digital culture, and none of these is under the bibliographic control of the library. Where in past epochs one could consider the library the main source of recorded information, this is no longer true today. This, in itself, is not the problem. We should be pleased with the growth of and use of information and the resulting potential for an informed society representative of an active and vocal populace. The problem, instead, is that libraries are distant from and unconnected to today's primary information resources, which are on the web. The push to move libraries in the direction of linked data is not just a desire to modernize the library catalog; it represents the necessity to transform the library catalog from a separate, closed database to an integration with the technology that people use for research and creation of new ideas. Library data needs to be online where it will interact with existing and future information resources. This means that to be visible to today's user, the library catalog needs to cease to be a separate database; it must become data dispersed throughout the web, fully linked to the web of data. Our job today, as librarians and information scientists, is not to translate library data to linked data; our job is to create a new system for access and use of bibliographic data that is compatible and works within the web. There are two primary aspects of this development. The first is to make bibliographic data usable on the web. Every person who does research, who studies, who writes and cites, needs bibliographic data, some of which can be provided by libraries. With library bibliographic data on the web, everyone online becomes potentially a "library user." The other aspect is the use of online data to improve the libraries' user services. By making connections between bibliographic data and web resources one can, for example, place a book within its historical context or demonstrate the influence of an author on his time. Progress has already been made in some areas, as you will learn from the speakers at this seminar. There are two primary activities that provide the background for the creation of linked data: the first is the development of the metadata elements that one will use for the data, such as "author" or "title"; the second is the gathering of controlled lists of terms that will be used as values, lists like languages, geographic places, and names of persons. Because library metadata standards already define a number of controlled lists of terms, these have been fairly easily converted. The Library of Congress presents its subject headings as linked data, as do the national libraries of France, Germany, Japan, and others. Some linking has been created between them, forming the basis for a future web of subject data that is multilingual and international. Name authority data in linked data form can be found in the Virtual International Authority File, Virtual International Authority File (VIAF). VIAF, which is held at Online Computer Library Catalog (OCLC), receives name authority records from about twenty different major libraries. It clusters the records for the same person and creates an identity for that group. Where possible, a VIAF cluster links to the Wikipedia article for that same entity, and in some cases there is a reciprocal link from Wikipedia to VIAF. Again, this is the beginning of a web of data. There is a certain amount of experimentation in the translation of traditional bibliographic schemes to linked data: in particular, International Standard Bibliographic Description (ISBD), Functional Requirements for Bibliographic Records (FRBR), Functional Requirements for Autority Data (FRAD) and Resource Description & Access (RDA) have been coded using semantic web standards. However, these are not connected to any web-based data, and this is a very important point to make. A key part of the semantic web that differs significantly from metadata practices of the past is that of linking, and in particular linking between metadata elements from different communities. It is only through this linking that we will make the transformation from a closed world of library bibliographic data to the open world of the semantic web. This means that we need to make connections between library data and data that has its origins in other communities and resources, whether these come from scientific research, government data, commercial information, or even data that has been crowd-sourced. If we must understand one key thing about the semantic web it is that it is an information environment that is highly heterogeneous, both in its breadth but also in quality. The closed world of bibliographic control that we have enjoyed up until now will not be part of our future. To conclude: We must ask ourselves if linked data is going to solve all of the libraries' problems, and the obvious answer is: no, of course not. But the bottom line is that we cannot move into the rich and dynamic information environment of the 21st century with data that is based on 19th century principles. It is possible - no, it is probable - that we will need a profound change to library data to meet today's needs. In the end there will be a significant difference between today's library catalog and the access and view of library data that integrates with the web. We must no long create bibliographic data that is intended only for library use. Our users are not limited to those who interrogate the library catalog but are all persons who seek information and create new resources, whoever they are, wherever they are. We must be not only on the web, but of the web. We must use the standards of the web, the structure of the web, and the services and applications of the web. The biggest risk is that we will change, but we will not change enough. The original goal of RDA was radical: it intended to break with the cataloging standards of the past and create a new view of library bibliographic data that was open, flexible and extensible. However, as the work on the standard went forward many in the field questioned our ability to make this change, and the committee retreated to a position of guaranteeing that RDA would integrate well with current library data. Unfortunately, no analysis was done of possible systems solutions for transformation of the data. We have let our past anchor us in place, and to keep us from moving forward. The result is that when we adopt RDA in 2013 it is possible that our data will be nearly indistinguishable from that of our current catalogs. It is not just the machine-readable format of our data that needs to change, but the content of our data. We will not become relevant by recreating ISBD or Machine Readable Cataloguing (MARC) in Resource Description Framework (RDF). The library bibliographic record today is essentially a marked-up text, using natural language to describe resources, and is not suitable for machine actionability. We continue to create headings whose function is directly relevant to the linear catalog and alphabetical order. This is not only no longer useful in today's world but it actually makes it harder for us to exchange our data with communities whose data is structured for machine-applicability. We can no longer view the goal of our data creation to be a library catalog that looks much like the catalog we have today. And we can no longer view our catalog as a destination that is separate from the open web. The time of the library catalog is over, as much in the past as the time of the horse and carriage. Instead of insisting that our data cannot change because it has always been like this, we have to turn our attention to ways that we can re-utilize this data: to the transformation of our data using the computing power that exists today as well as the computational capabilities provided by the web itself. All this said, I want to end with a call to all of you to consider the idea of a library of the web as worth exploring; as one possible future, but not the only one; to be willing to consider that library data will take an entirely different form from what it is today, and that this will not lead to the destruction of the library as we know it but to its evolution for future generations.

Karen Coyle, Digital Libraries Consultant.
kcoyle@kcoyle.net
http://www.kcoyle.net/

Coyle, K. ”Library linked data: an evolution”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5443. DOI: 10.4403/jlis.it-5443. Web.

Abstract: Actually the world is full of hi-tech devices that allow in different ways to stay always connected and up-to-date. The high level of interactivity and interoperability is now a fundamental prerogative in the new millennium communication which no one can be exempted. In particular, librarian universe must reconfigure this new way to exchange and provide information otherwise it will cause the steady decline, and this will cause a huge loss of users who will find different ways to information: such as tablet or smartphone. The linked data can provide significant support for libraries to ensure that they can still be leaders in the future. A brief mention to the studies conducted in this domain.

Keywords: Library linked data; Semantic web

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

Designing data for the open world of the Web

Tom Baker

Resource Description Framework (RDF) is the grammar for a language of data. Uniform Resource Identifier (URI)s are the words of that language. As in natural language, these words (i.e., the URIs) belong to grammatical categories. RDF properties (such as ”isReferencedBy”) function a bit like verbs, RDF classes like nouns. As in natural languages, where utterances are meaningful only if they follow a sentence grammar, RDF statements follow a simple and consistent three-part grammar of subject, predicate, and object. Analogously to paragraphs, RDF statements are aggregated into RDF graphs.

Aside from being words in the language of data, URIs double as footnotes. As footnotes they indicate the maintenance responsibility for words by way of ownership of the domain names under which the URIs were coined, as recorded in the globally managed Domain Name Service (DNS). Inasmuch the URIs of words lead to documentation of official definitions, the web itself provides the language of data with its dictionary. The fifteen elements of Dublin Core have been likened to a ”pidgin” – a lexicon of generic predicates good enough for the sort of rudimentary but serviceable communication that occurs between speakers of different languages. Just as pidgins are inadequate for more subtle or differentiated expression,

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-6308

a healthy ecosystem of RDF vocabularies needs to include more specialized vocabularies for use by social or scholarly communities of discourse among themselves. RDF is a language designed by humans for processing by machines. The RDF language – the grammar together with available RDF vocabularies – does not itself solve the difficulties of human communication any more than the prevalence of English guarantees world understanding. However, RDF does support the process of connecting dots – of creating "knowledge" – by providing a linguistic basis for expressing and linking data. Just as English as a second language provides a basis for communication among non-native English speakers, RDF provides a common second language into which local data formats can be translated and exposed. Just as English is useful without being the best of all possible grammars, RDF happens to be what we currently have – the only general-purpose language for data with any traction. But just as English grammar follows deep linguistic structures determined by the human capacity for language, it is likely that RDF, if re-invented, would end up strongly resembling what we currently have. Aside from supporting data interchange in the here and now, RDF provides a response to the ongoing and inevitable obsolescence of computer applications and customized data formats by expressing knowledge using a well-understood grammar and citing publicly documented vocabularies and resource URIs. In this sense, it supports data that does not require additional out-of-band information for its interpretation, i.e., data that "speaks for itself". This assumes, of course, that our cultural memory institutions will deploy robust methods for preserving the parts of the Web where the underlying RDF vocabularies and resource identifiers are documented. We are in the midst of a rapid shift from a world in which information was predominantly print-based to one in which it is predominantly digital. The scale and speed of transformation virtually guarantees that any computer applications and user interfaces we use today will at some point, probably soon, be superseded. Data that cannot speak for itself will be more vulnerable to becoming irrelevant.

Not only is data expected to be linkable in the present, but we hope they will be remain intelligible in the future. In 2012, to put information into ad-hoc data formats in the absence of well-defined interpretations as RDF triples is like making statements without grammar. Creating data without URIs is like writing without proper footnotes. This is okay for information with a short shelf life – i.e., most information – but information of lasting cultural significance deserves better. Cultural memory institutions live by the ethos of scholarship, by which things like good grammar and proper footnotes should really matter. The language of RDF represents the application of that ethos to data itself.

TOM BAKER, Dublin Core Metadata Initiative.
thomas.baker@bi.fhg.de

Baker, T. ”Designing data for the open world of the Web”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6308. DOI: 10.4403/jlis.it-6308. Web.

ABSTRACT: The domain name system of the world wide web provides a managed space of globally unique identifiers for web pages – Uniform Resource Identifiers, or URIs. URIs can also be used to name things – specifically, to name things in the world (”people,” a ”books,” or ”Nelson Mandela”); to name concepts used to describe those things (”Renaissance Sculpture” or ”Lyme Disease”); and to name relationships between things (this book ”was translated by” that person). Because URIs, used as names, are globally unique, they serve to anchor the strands in ”webs of meaning” (”semantic web”). Each strand of the web is a statement following a grammar, the Resource Description Framework (RDF), which uses URIs as its words. Each RDF statement expresses a simple idea – ”Dante wrote L'Inferno” or ”Dante was born in Florence” – which, taken together, can express complex webs of relationships. Expressing data as statements makes it easy to integrate data across many different sources (”linked data”). The opportunity for cultural heritage lies in translating the traditions of resource description into the language of URIs so that its descriptions of Works, Items, Subject Headings, and People can serve as central hubs in growing webs of linked data.

KEYWORDS: Semantic web; Library linked data; URI

Submitted: 2012-06-10
Accepted: 2012-08-31
Published: 2013-01-15

Linked data: a new alphabet for the semantic web

Mauro Guerrini, Tiziana Possemato

What is linked data

The term linked data is entering into common vocabulary and, as most interests us in this instance, into the specific terminology of library and information science. The concept is complex; we can summarize it as that set of best practices required for publishing and connecting structured data on the web for use by a machine. It is an expression used to describe a method of exposing, sharing and connecting data via Uniform Resource Identifiers (URIs) on the web. With linked data, in other words, we refer to data published on the web in a format readable, interpretable and, most of all, useable by machine, whose meaning is explicitly defined by a string of words and markers. In this way we constitute a linked data network (hence linked data) belonging to a domain (which constitutes the initial context), connected in turn to other external data sets (that is, those outside of the domain), in a context of increasingly extended relationships. Next is presented the Linked Open Data cloud (LOD), which collects the open data sets available on the web, and the paradigm of its exponential growth occurring in a very brief period of time which demonstrates the level of interest that linked data has garnered in organizations and institutions of different types.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-6305

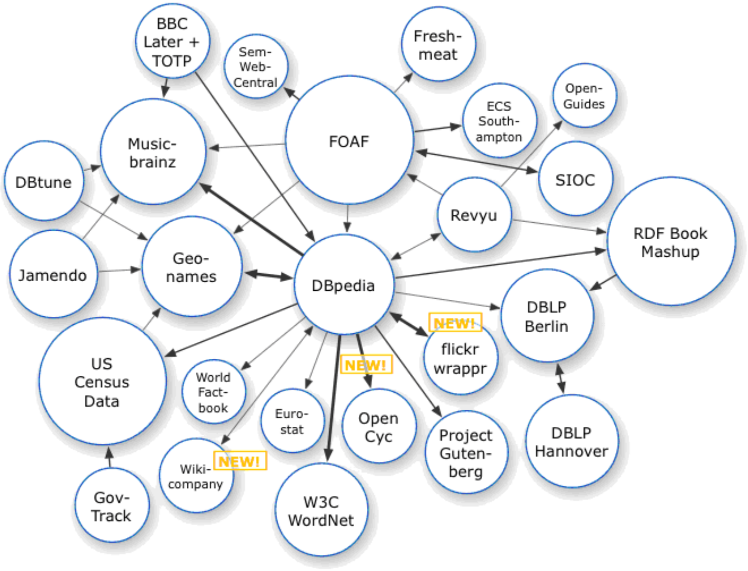


Figure 1: Diagram of the linked open data cloud (LOD) in 2007.

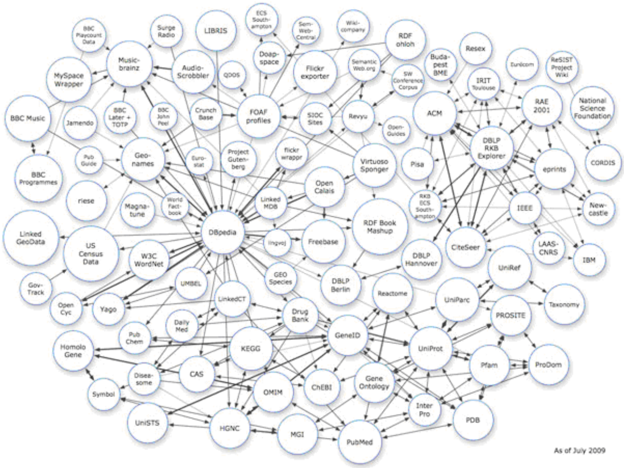


Figure 2: Diagram of the linked open data cloud (LOD) in 2009.

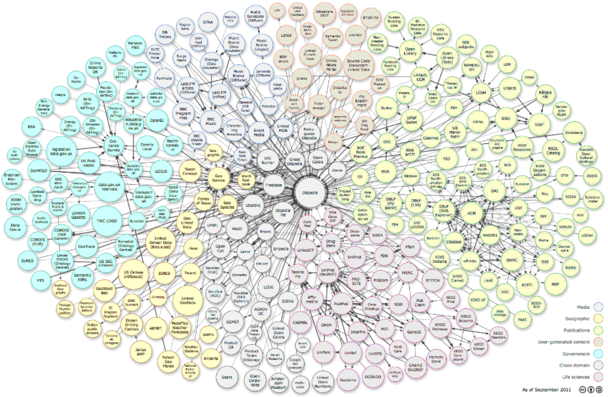


Figure 3: Diagram of the linked open data cloud (LOD) in 2011.

The concept of linked data is closely related to the semantic web, although the semantic web cannot be reduced to the mere technicality of linked data, but requires, for its construction, that certain important rules be respected whose ultimate goal is the creation of a layer of content accessible to automated processes. Linked data make explicit the meanings and connections implicitly contained (or in some cases, absent) in web resources (data, pages, programs, etc.). The two terms – linked data and semantic web – relate to the same semantic field and area of application. Linked data is a technology used to realize the semantic web. To better understand the concept we are aided by the definition that Tim Berners-Lee, inventor of the world wide web (www), provides for semantic web: ″A web of things in the world, described by data on the web″. The concept is generic, but it contains important references: the network, the things (the objects related), the data (no longer a record but individual elements, atoms). This differentiates the traditional web (the hypertext web) – constituted of documents, HTML objects, connected via unclassified hyperlinks – from the web constituted of ″real things″ (existing entities) described via data. A more precise image begins to emerge:

- the hypertextual web or web of documents as a flat, linear, representation of objects; the concrete nature of the semantic web is in opposition to the abstract nature of the traditional web;

- the semantic web or web of data as a container of things, of objects, rather than as a container of representations of objects: an idea of concreteness, in the sense that the data relate to the resource and participate in its nature, that is, they are an integral part of it, as the resource would not be representable without this data.

The semantic web was not born, therefore, to replace the traditional web, but rather to extend its potential, realizing what Tim Berners-Lee describes as a world in which ”the daily mechanisms of commerce, of bureaucracy, and of our everyday lives will be managed by machines that interact with other machines, leaving to human beings the task of providing them with inspiration and intuition” (Berners-Lee and Fischetti).
The web of data is, therefore, the natural evolution of the web of documents. Let us try to identify the distinctive features of each of them, comparing their characteristics:

- web of documents (hypertextual web):
    - analogy with a global filesystem, an expression of extreme richness but also particularly monolithic;
    - flat description of objects and documents; documents as primary objects of description;
    - network of relationships between objects made up of relationships between documents which are neither inherent in the objects themselves, nor form part of their structure; links between documents; in consequence:
        - semantics of the content and of the links between documents is empirical, associated with the objects, and thus not part of the object itself, created by a human agent;
        - low degree of structure in the objects;
        - objects represented on the web designed for human consumption, not machine-interpretable or reusable.

The hypertextual web is simple in structure, and has sparse connections between the data. It can be imagined as an enormous notebook, in which information is noted in a linear fashion, that is, with little structure and few relationships, and in which documents are readable and useable only by humans.

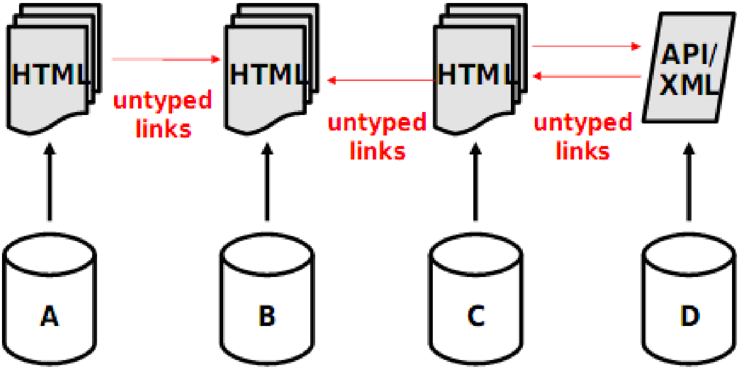


Figure 4: Representation of the web of documents, 17th International World Wide Web Conference W3C Track @ WWW2008, Beijing, China 23-24 April 2008 - Linked data: principles and state of the art.

- web of data (semantic web):
  - analogy with a global database conceived as a relational database, consisting of individual objects richly related to each other, which in turn form larger entities;
  - articulated description of the object, a description which itself becomes an object in the web, because it is reusable; things (or descriptions of things) as primary objects of description;
  - network of relationships between objects inherent in the objects themselves; links between things (including documents); in consequence:

* semantics of the content and of the links is explicit, expressive;
* high degree of structure in (the descriptions of) things;
* entities designed for machines first, human beings second.

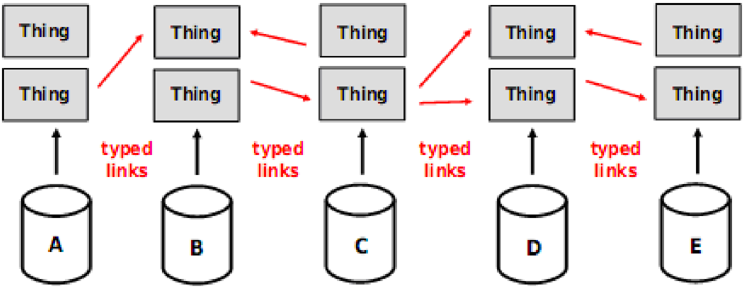


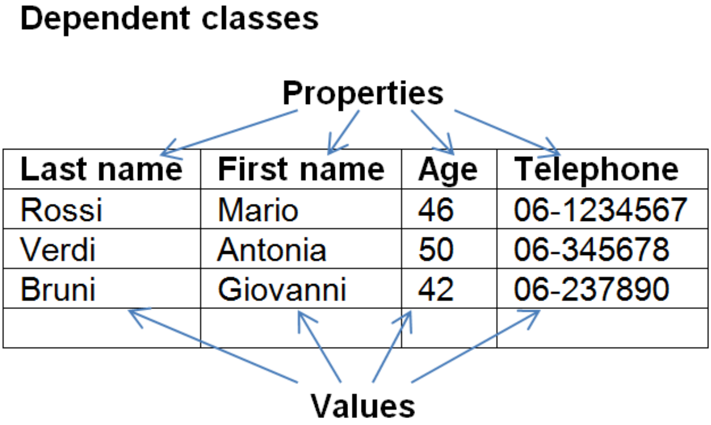
Figure 5: Representation of the web of data 17th International World Wide Web Conference W3C Track @ WWW2008, Beijing, China 23-24 April 2008 - Linked data: principles and state of the art.

The comparison with relational databases is a basic concept in the literature on this topic. We can read on the site of the W3C:

> ”The semantic web and relational databases. The semantic web data model is very directly connected with the model of relational databases. A relational database consists of tables, which consists of rows, or records. Each record consists of a set of fields. The record is nothing but the content of its fields, just as an RDF node is nothing but the connections: the property values. The mapping is very direct
>
> - a record is an RDF node;

- the field (column) name is RDF propertyType; and
- the record field (table cell) is a value.”

A strong point of the semantic web has always been the expression, on the web, of a large quantity of information in the relational database formulated in a machine-processable format. The serialization format RDF – with its syntax XML – is a format suitable for expressing the information in relational databases. The analogy is appropriate as the central point of linked data is precisely the “predicates” that express the types of relationships through which ontologies and networks can be represented.



Last name	First name	Age	Telephone
Rossi	Mario	46	06-1234567
Verdi	Antonia	50	06-345678
Bruni	Giovanni	42	06-237890

Figure 6: Representation of a relational database.

The atomization of the structure of information expresses the characteristics of the web of data; one no longer has a monolithic object, rather a set of individual data points, minimal particles – atoms – that can be reaggregated in different ways and for different purposes;

each attribute of the object has a value in itself, and participates in its nature, through expressive, self-explanatory, relationships. The entities constituted by the ensemble of atoms are assembled into a set of structured data, each individually independent, but able to be logically combined with other data to produce new entities. Having given the image of the notebook to illustrate the web of documents, we can now take the image of the mechanism (reminiscent of Ranganathan), in which every element, independent in itself, can be combined and reused in an infinite variety of solutions. The web of data is, therefore, a global network of statements (or sentences) connected through qualified and self-expressive links which become a collection of knowledge, which is readable and understandable by a machine, only secondarily for a person.

Linked data: the world of the internet and the role of libraries, archives and museums

Why is the world of networked information so interested in the legacy data produced by libraries, archives and museums? Why are libraries, archives and museums equally interested in linked data? The interest is actually reciprocal. Libraries have always produced quality data in highly-structured bibliographic and authority records, according to shared and widely disseminated rules, a vast quantity of data. The world of libraries and the world of the internet are both interested in integration into the net; the former to ensure the visibility and usability of its data, the latter to exploit information and create increasingly large and significant networks. The quantity and quality of the information that populates the net are two aspects which are often inversely proportional: much information is of poor quality. The increase in networked information (through publication methods that are increasingly widely-known and used, such as for example, self-publishing, social networks) is not, in fact, always synonymous with quality. The exponential growth and use of information available on the net does not coincide with increasing trustworthiness of the records either: their degree of reliability is low. Users must select from the sea of information retrieved to arrive at a credible record. On which criterion to base the selection? The authoritativeness of the source becomes the key factor, the selection takes place at the outset, preferring to select a resource on the basis of the authoritativeness of its creator, instead of later on, choosing uncritically on the basis of the ranking of the records that appear on the page. The quality of the source, the certainty of the provenance become, therefore, crucial elements in the searcher's exploratory process. The role of libraries, archives and museums thus becomes relevant, due to their tradition of attention to the quality of the information they produce. Libraries, archives, museums assume, thus, the role of generators of quality information for the net. It is for this reason that their data are sought after.

Legacy metadata in libraries: still functional?

The history of library catalogues demonstrates early widespread use of metadata, understood as information serving as a surrogate for the resource. The evolution of data into ever more structured and detailed records coincided with the renewed centrality of the catalogue on which every service of the library is based, the proliferation of formats of bibliographic resources and the central role of automation in library systems. The main characteristics of metadata are its:

1. nature: it is created, formed from the resource;

2. aim: to describe an object;
3. use: it must be structured in such a way as to be processable (that is, useable) by a machine, a computer.

Libraries have long had the stable and consistent objective of sharing information through metadata, and have always accorded importance to its quality. Are the metadata used up to this point still functional? Do they respond to the requirements of current information usage? Is it enough to expose on the web the data that libraries have produced over the centuries? Is this exposure (for example, in MARC format) comprehensible and useable outside of a strictly library context? Does this not risk being a niche exposure, restricted to a narrow environment, in a closed and highly professionalized domain?

The catalogue of the future: of the web and not only on the web

We note that the data produced by libraries – the catalogues –, whose creation required the development of standards, professional competencies and financing, are not on the web, but isolated from the web. Catalogues are not, in fact, integrated into the web, they are not searchable, even though the web is the place in which most users work, play, operate and create other information. The question, therefore, is: "How to modify catalogues and data so that they can be of the web and not only on the web?". It is exactly the philosophy that underlies linked data technology that can offer an interesting starting point for achieving this strategic goal, on pain of death for catalogues, abandoned by users in favour of other information retrieval tools, such as search engines. It is a fundamental transition:

the inevitable adoption of linked data will bring about a new revolution, even more radical than that of the 1970s, which saw the passage from the card catalogue to the automated catalogue and then on to the computerized catalogue, a revolution which crowned the role that information technology has assumed in the management of communication processes and, therefore, as concerns us more closely, in the creation of mediation tools between the bibliographic universe and the user. On the record, the report of the Library of Congress Working Group on the Future of Bibliographic Control, gives sound guidance in achieving this goal; the change implies:

1. the transformation of textual description into a set of data usable for automatic processing by machines;
2. the need to render data elements uniquely identifiable within the information context of the web;
3. the need for data to be compatible with the technologies and standards of the web;
4. the need, in short, to use a language that is in reality interoperable across the web.

The concept of unique identification of objects is of particular interest: the object identified, characterized as being the same thing regardless of its textual expression (having, thus, the same meaning) should have a unique identifier, so as to be useable in diverse contexts (libraries, publishers, booksellers, distributors, producers of online biographies ...), as well as through the use of different textual values.

Tim Berners-Lee identified four rules for the creation of linked data on the web:

1. use URIs (Uniform Resource Identifiers) to identify things (objects): URI is a system of global identification, thus valid for

all resources contained on the entire web. URI is a keystone of web architecture, inasmuch as it constitutes a mechanism of resource identification common to the whole web. Each resource on the web (a site, a page within a site, a document, any object) must be identified by a URI to be found by other systems, used, linked, etc.;

2. use HTTP URIs so that these things can be looked up by people and user agents (browsers, software . . .): the schema used to construct a URI is declared in the URI itself prior to the colon (:); for example, http://weather.example.com/. HTTP uses HyperText Transfer Protocol as its protocol, which is precisely the schema prescribed for the semantic web;

3. when someone looks up a URI, provide useful information, using the standards (RDF, SPARQL (a query language devised for linked data)): it is necessary to define the context and the characteristics of the resources, through the attribution of the resource itself to a class, the identification of its properties and the assignment of values;

4. include links to other URIs, so that they can discover more things: the more the data are linked, the more they can be used for enrichment and the deduction of information.

Linked data: RDF (Resource Description Framework)

Producing linked data means, therefore, expressing the meaning of information, making it shareable among different applications and useable by applications other than those for which it was originally created. The data model used to structure linked data is RDF, a flexible standard proposed by the W3C to characterize semantically both resources and the relationships which hold between them. We have defined the reality of the web as a global network of statements (or sentences) linked via qualified links. The RDF model codifies the data in the form of statements comprised of:

1. subject: the portion of the sentence that identifies the thing that is described;
2. predicate: the property of the thing specified by the sentence;
3. object: the value of the property of the thing (the RDF triple).

Examples:
Alberto Moravia is the author of La noia
Bompiani published Il nome della rosa
Alberto Moravia is the pseudonym of Alberto Pincherle

Each element of the triple, Tim Berners-Lee reminds us, can, or rather, must, technically, be represented via URI. The more URI are used the more the information is reusable; this is not required and elements of the triple can be expressed even in textual format. The statements, or triples, are expressed in RDF in the form of graphs (nodes and arcs) which represent the resources, their properties and their respective values.

The triples are encoded via an XML-based syntax (RDF/XML) to make them readable, interpretable and understandable by machine, which can be the one for which the data was created (the native system) or a system other than (external to) the one for which it was originated. This is the most important characteristic, which opens the data to the global information community.

Let us observe the following assertions:

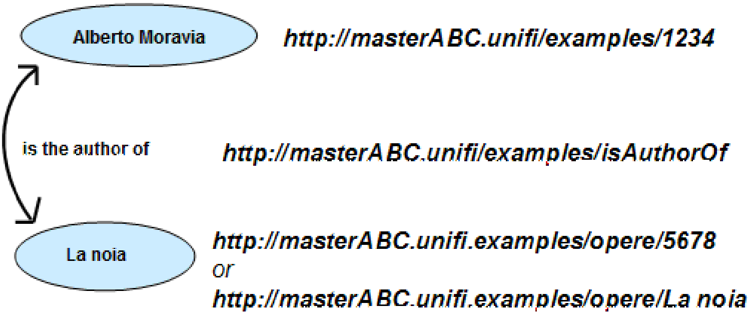

Alberto Moravia
http://masterABC.unifi/examples/1234
is the author of
http://masterABC.unifi/examples/isAuthorOf
La noia
http://masterABC.unifi.examples/opere/5678
or
http://masterABC.unifi.examples/opere/La noia

Figure 7: Representation of a triple (nodes and arcs) in RDF.

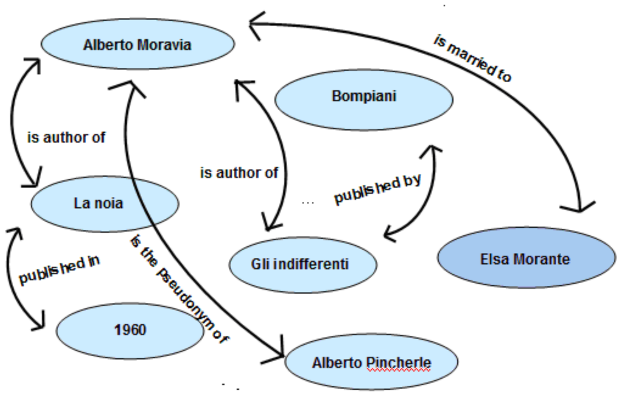
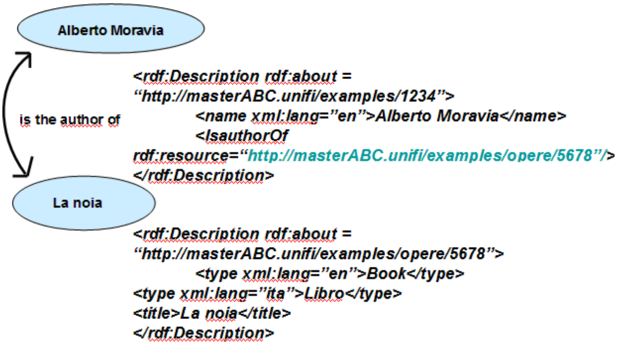


Figure 8: Representation of a network of assertions or triples.



```
<rdf:Description rdf:about =
“http://masterABC.unifi/examples/1234”>
        <name xml:lang=”en”>Alberto Moravia</name>
        <IsauthorOf
rdf:resource=“http://masterABC.unifi/examples/opere/5678”/>
</rdf:Description>

<rdf:Description rdf:about =
“http://masterABC.unifi/examples/opere/5678”>
        <type xml:lang=”en”>Book</type>
<type xml:lang=”ita”>Libro</type>
<title>La noia</title>
</rdf:Description>
```

Figure 9: Representation of a triple in RDF/XML.

Marco is the son of Gianni
Susanna is the daughter of Gianni
Gianni is the son of Chiara

From these simple assertions it is possible to recover at least three others, even though not made explicit with triples:

Marco and Gianni are male
Susanna is female
Chiara is the grandmother of Marco and Susanna

and we could deduce even more, for example:

Marco and Susanna are grandchildren of Chiara
Marco is the brother of Susanna
Susanna is the sister of Marco

This mechanism, termed inference – the process through which, from a proposition accepted as true, one can pass to a second proposition whose truth-value is inferred from the content of the first – is the principle governing the engines that are behind the semantic web, which infer knowledge via paths. Each new statement, expressed in the form of triples and, therefore, in graphs, becomes in turn the generator of new information; the more the spheres of belonging of these statements (data sets) grow and intersect, the more the semantic network present and available on the web is enriched and becomes categorized information. The mechanism of inference is well-known in logic and mathematics (inferential calculus) and is widely used in computer applications. It acquires a particular flavour when applied to the library world; the mechanism explains, in fact, the relationships present in bibliographic data but not always evident, and of which we became fully conscious with the theoretical systematization accomplished by FRBR: a systematization of concepts existing in cataloguing tradition, at least from Cutter onwards, and made increasingly explicit.

For this mechanism to work, a technological infrastructure must be used in which concepts are identified uniquely and in which software agents recognize these objects and realize associations and equivalences among them, through reference to ontologies, formal representations, shared and explicit to specific domains of knowledge. Ontologies permit the representation of entities through the description of their characteristics and the identification of the relationships holding among them, and thus of the semantics that links such entities, used primarily to realize categorizations and deductive reasoning. Examples of vocabularies and ontologies widely-known in the library world are:

FOAF (Friend Of A Friend) an ontology used to describe persons, their activities, their relationships with other persons or things, very useful in structuring authority files in linked data;

SKOS (Simple Knowledge Organization System) a family of formal languages created to represent thesauri, classification schemes, taxonomies, subject headings systems and every type of controlled vocabulary.

IFLA is concentrating on publishing its own standards in RDF with the creation of vocabularies and ontologies for FRBR, FRAD, FRSAD and ISBD, published in the Open Metadata Registry (previously the NSDL Registry), a space created by the W3C to support developers and users of controlled vocabularies, hosting ontologies from different fields, among which are the vocabularies for RDA (Resource Description and Access), the new cataloguing standard that replaces AACR2 (Anglo-American Cataloguing Rules, 2nd edition) created by the Anglo-American library community, expanded with reference to the European context (France in particular) and offered to the international bibliographic and library community.
Ontologies are necessary, therefore, to create and publish a dataset, which expresses a domain of belonging representing a kind of collection of resources (or graphs), having some characteristic in common, and identified via dereferenceable URI. Examples of datasets available on the web are:

Dbpedia dataset containing data extracted from Wikipedia;

LinkedMDB dataset on the world of cinema;

VIAF Virtual International Authority File.

Let us try to elaborate possible inferences combining data present in these datasets:

Eduardo De Filippo was alive between 1900 and 1984 (from VIAF)
Eduardo De Filippo is the author of Filumena Marturano (from VIAF)
Eduardo De Filippo was born in Naples (from Dbpedia)
Naples is the capital of the Region of Campania (from Dbpedia)
Questi fantasmi is a film directed by Eduardo De Filippo (from linked MDB)

Massimo Troisi is the director of Ricomincio da tre (from Dbpedia)
Massimo Troisi was born in Naples (from Dbpedia)
Ricomincio da tre is a film from 1981 (from linked MDB)
Scusate il ritardo is a film directed by Massimo Troisi (from linked MDB)

If we wanted to create a dataset relating to celebrities from Campania who have distinguished themselves in literature and cinema we could use the triples above, extracted from various data sets, to feed into our set and infer in this way new information: Eduardo De Filippo and Massimo Troisi are 20th century celebrities from Campania, literary authors and filmmakers.

Open Linked Data Project

How accessible are these datasets, and what are the ways to make them truly usable for the wider community? Each institution could produce its own linked data, as defined by the criteria and rules mentioned above, but not make them open for use on the web. For a dataset to be open (and therefore not subject to commercial licenses or use restrictions) it must be published as defined by the Open Linked Data Project, which provides for the conversion of existing datasets or the production of new ones, according to linked data principles, but with open licenses. The project, kicked off initially with the participation of small organizations, and researchers and developers in universities, has, over time, gained numerous adherents among larger, more authoritative organizations and institutions, among them the BBC, Thomson Reuters and the Library of Congress. This level of adherence and dissemination among respected, recognized and prevalent circles has resulted in the remarkable growth and expansion of the project, facilitated by its open nature: anyone can participate by publishing a set of data that respects the principles of linked data and creating cross-links (interlinking) with other existing datasets.

Library Linked Data Project

The W3C Library Linked Data Incubator Group was founded to support and favour the development and growth of the interoperability of library, archival and museum data on the web. It followed the principles of linked data and the semantic web, and the group's work was carried out in strict collaboration with the actors in these areas. Interesting use cases for the writing of the Final report[1] of the Incubator Group were provided by the projects supported by organizations, small, medium, or the large national libraries. The Final report began with the analysis of ongoing projects and defined an overall picture, it can be summarized as follows:

- analysis of the benefits possible from the application of the principles of linked data in the library sector;

- discussion of open issues with particular reference to traditional data;

- analysis and enumeration of linked data projects and initiatives in the library sector;

- discussion of issues relating to legal rights and to publication;

- making of recommendations for next steps in the process of applying the principles of linked data to the sector.

[1] Available at: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/.

Life cycle of linked data

What are the steps that an organization must take to process its own data and result in its publication as linked data? A good methodological reference is provided by Boris Villazón-Terrazas ("Methodological guidelines for publishing linked data"), which reproduces the life cycle for the production of linked data in 7 steps:

1. identification of the data sources;
2. generation of the ontology model, with the adoption of existing ontologies, expressed in OWL, Web Ontology Language, or RDF(S) or with the creation (more complex) of new ontologies;
3. generation of data in RDF format, through various available mapping languages, also in relation to the original format of the data. In this phase the most delicate operation is the creation of URI, as these are the key to aligning heterogeneous resources drawn from different sources;
4. publication of the RDF data;
5. data cleaning, to identify eventual and possible conversion errors and make the data qualitatively useable;
6. linking the RDF data with other existing data sets, with the identification of datasets of interest that can become linking targets, identifying relationships between individual data, validating the relationships thus identified;
7. make concrete the use of the data, through various steps, among which the publication of the resulting dataset on the CKAN Registry (Comprehensive Knowledge Archive Network), a registry for the publication of open data and packages, which makes their discovery, sharing and reuse possible.

The 5 stars of open linked data

A dataset obtained with the 7 steps suggested by Boris Villazón-Terrazas can then be evaluated via a ratings system defined by Tim Berners-Lee to assign a score to sites that expose data on the web, termed the 5 stars 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.

The assessment of the open linked data produced must be carried out considering, therefore, five fundamental aspects:

1. one's own data being available on the web (in whatever format);

2. the material put on the web is available as structured data (for example, in excel instead of as a scanned image of a table);

3. having chosen non proprietary formats (for example, in csv instead of excel);

4. having used URL to identify the objects, so that users can point to these objects;

5. one's own data is linked to data produced by others so as to define a context.

Tim Berners-Lee's indications for the assessment of open linked data were followed by a series of recommendations, suggestions and ways to establish ever more precise norms and rules for evaluation, to arrive at a standard as participatory and shared as possible.

References

Berners-Lee, Tim and Mark Fischetti. *Weaving the web: the original design and ultimate destiny of the world wide web by its inventor*. New York: HarperCollins, 2000. (Cit. on p. 71).

Villazón-Terrazas, Boris and Oscar Corcho. "Methodological guidelines for publishing linked data". *Una Profesión, un futuro : actas de las XII Jornadas Españolas de Documentación : Málaga 25, 26 y 27 de mayo de 2011*. Madrid: Federación Española de Sociedades de Archivística, Biblioteconomía y Documentación, 2011. (Cit. on p. 87).

MAURO GUERRINI, Università degli Studi di Firenze.
mauro.guerrini@unifi.it

TIZIANA POSSEMATO, @Cult.
tiziana.possemato@atcult.it

Guerrini, M., T. Possemato. "Linked data: a new alphabet for the semantic web". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6305. DOI: 10.4403/jlis.it-6305. Web.

ABSTRACT: The paper defines the linked data as a set of best practices that are used to publish data on the web using a machine; the technology (or mode of realization) of linked data is associated with the concept of the semantic web. It is the area of the semantic web, or web of data, as defined by Tim Berners-Lee "A web of things in the world, described by data on the web". The paper highlights the continuities and differences between semantic web and web traditional, or web documents. The analysis of linked data takes place within the world of libraries, archives and museums, traditionally committed to high standards for structuring and sharing of data. The data, in fact, assume the role of generating quality information for the network. The production of linked data requires compliance with rules and the use of specific technologies and languages, especially in the case of publication of linked data in open mode. The production cycle of linked data may be the track, or a guideline, for institutions that wish to join projects to publish their data. Data quality is assessed through a rating system designed by Tim Berners-Lee.

KEYWORDS: Library linked data; RDF; Semantic web

ACKNOWLEDGMENT: Paper presented at the Conference ”I nuovi alfabeti delle biblioteca. Viaggio al centro di un’istituzione della conoscenza nell’era dei bit: dal cambiamento di paradigma ai linguaggi del cambiamento”, Milan, Italy, Palazzo delle Stelline, 15-16 March 2012. Revised and augmented version.

Submitted: 2012-06-01
Accepted: 2012-08-31
Published: 2013-01-15

JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

Linked (open) data at web scale: research, social and engineering challenges in the digital humanities

Michele Barbera

In the last decade, the amount of data available in the linked data cloud has grown enormously in several domains, including cultrual heritage and digital humanities. However creative reuse of data both within the scholarly community and within the cultural industry is still very limited. The limited creative reuse of data does not only depend on the limitations of existing technologies, but also on several social and cultural habits whose consequences need to be fully addressed and further researched. If linked data is to be exploited at its full potential, a profound cultural shift needs to occur in the way data is produced, managed and disseminated. This is especially true in the cultural heritage and digital humanities domains, where a strong tradition of two-dimensional, paper-like thinking is still predominant. The first section of this paper briefly presents the most pressing technological and engineering challenges to be addressed within the linked data sector. In the second section it is argued that the full exploitation of the linked data sector does not only depends on technological advancement but also on the possibilities enshrined in a radical cultural change in thinking about the semantic web and linked open data visions. The potential effect

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-6333

on the cultural industry and on the scholarly community is also explored. The importance of nurturing a lively business ecosystem and the role of public and open data published by GLAM[1] organizations is the main condition enabling the linked data vision to take off. Despite it is here impossible to provide a comprehensive analysis of potential of the semantic web, its enabling conditions and implications, this paper nonetheless aims at offering a stimulating insight into one of the possible ways of thinking about it.

Technical and engineering challenges

The web of data is characterized by its universal nature, its virtually infinite size, and by the heterogeneity of data. It comes at no surprise that these and many other features have influenced the way in which data producing and data consuming applications are -– or ought to be -– designed. First of all, due to the rapidly growing amount of data available in the linked open data cloud and in enterprise linked data repositories, it is not possible to centralize and compute all the data in a single local repository. The largest existing public repository, Sindice.com[2] holds today around 80 billion triples, which is just a fraction of the LOD Cloud. Sindice.com is based on a large map-reduce cluster (implemented on Apache Hadoop) whose TCO is still beyond the possibilities of most small and medium enterprises (SMEs) and research organizations.

One of the most significant features of the linked data vision is its capacity to find novel ways to exploit unexpected information and links to discover new insights from data. A way of *slicing* large datasets and reduce them down to a manageable size – possibly on demand – is necessary in order not to sacrifice this desirable feature.

[1] Galleries, Libraries, Archives and Museums.

[2] http://sindice.com.

This approach is not entirely new and it is sometimes refereed as dataspaces in the relational database community. Linked dataspaces are essentially a transient materialization (or a *view*) computed on demand of a slice of the originating data graph. Despite some implementations of linked dataspaces do already exist, neither standard specifications nor reference implementations have been defined yet.

Streaming linked data

Streaming linked data has been only partially addressed by the research community[3] (Barbieri and Della Valle; Le-Phuoc et al.; Sequeda and Corcho) and almost entirely neglected by production-grade industrial systems. At a first sight, this may appear marginally relevant but it is instead of paramount importance considering the growing amount of live streaming data produced by sensor networks and sensors embedded in personal mobile devices.

Versioning

From the capacity of versioning RDF graphs comes the possibility of identifying evolutionary atomic changes and to roll them back, in order to revert the graph to a previous state. Although some approaches have already been explored in research, efficient and production-ready industrial implementations in commercially or open source graph stores[4] are still under-developed. In the web of data, data sources are highly heterogeneous. The capacity of slicing and mixing different sources, that have various degree of trust (e.g. think about crowsdourced data vs. authoritative national library

[3] LarKC: The Large Knowledge Collider, http://www.larkc.eu.

[4] The availability of Open Source software of high quality is a very important element in the growth of this field of study.

data) and frequent updates a core feature. Hence, the lack of solid versioning systems is an important gap to fill as soon as possible.

Despite the importance of overcoming such limitations, technological innovation is a necessary but not sufficient condition for exploiting the potential of the semantic web. As the next section shows, this must come along with a critical understanding of the cultural shifts which are inner to the Semantic vision itself.

Social challenges

In a famous blog post, written in 2007, ("Giant Global Graph") introduced the concept of "thinking in the graph" and the notion of a "giant global graph" as opposed to the existing web of documents. The most striking changes whose effects are not yet being fully internalized in the digital publishing sector, are well summarised by Berners-Lee who writes: "The less inviting side of sharing is losing some control. Indeed, at each layer – Net, Web, or Graph – we have ceded some control for greater benefits", and "It is about getting excited about connections, rather than nervous". In order to understand the importance of these changes, in the following, three issue are considered: the AAA principle, the graph nature of the giant global graph and the Open World Assumption (OWA).

One of the pillars underpinning the vision of the semantic web and linked data – which also holds true for the web itself – is the principle known as AAA, which stands for *Anyone can say Anything, Anywhere*. This implies a profound shift in the paradigm dominating the publishing and media industry. In the era of the web of documents – and prior to the advent of the web – the dominant conceptual framework rested upon individuals acting as information producers, publishers (gatekeepers) or distributors. Information consumers (e.g. readers), acting as passive actors, were neither involved in the production nor in the distribution of information. As a result of the cultural and technological revolution enshrined in the web 2.0, readers have become active producers of information. Yet, the dissemination and often the production of information, as well as the gatekeeper role of filtering, was in the hands of publishers or aggregators of information. The web of data– which is here used as a synonym for semantic web and linked data – and the prominent role of the AAA principle in shaping publishing and consumption models,including query federation across multiple repositories, imply shifting the core activity of the information lifecycle from the production to the mesh-up of several heterogeneous and *personalized* data sources. In this novel scenario, individuals and organizations play at the same time the roles of information producers, gatekeepers, and consumers of information in an ever-reconfiguring ecosystem. In the traditional publishing world, both in the public and in the private sector, on the web and in other media, information (and data) are modeled in order to maximize the accessibility and especially the usability for consumers. This is achieved by anticipating scenarios in which information is consumed. By design, we cannot know in advance how data will be used, combined, enriched and repurposed to produce information goods. This is one of the premises that makes linked data so powerful. However, habits are difficult to change especially when they affect not only the production of data but also its consumption (e.g. think about the need for data consuming applications to be able to deal with unexpected data). Additionally, without knowing in advance how data will be consumed, data have to be represented in the most universal way possible. Such a universal representation does not allow any optimization in the information design phase and leaves room for optimizing data usage only in the consumption phase.

Secondly, the graph nature of the giant global graph has profound social implications in the way in which information is produced and consumed. Thanks to the great success of relational database technologies and spreadsheets, people mentally model data in tabular structures. IT developers and programmers, as well as information and data managers, scientists, marketers, educators and other actors involved in the data lifecycle think about data in a purpose-specific context and from an individual point of view.

Thinking in the graph as Berners-Lee puts it, is not an easy task when data is modeled, produced, aggregated or consumed. As my colleague Gradmann often remarks, this is even more difficult in the cultrual heritage and digital humanities communities, where there is a strong tradition of two-dimensional thinking derived from the paper-world. The two-dimensional paper-world approach has often been mimicked rather than revolutionized in the digital world. The problem is once again aggravated by technological constraints. First of all, most of the linked data nowadays published in the Cultural LOD Cloud is semi-automatically generated from legacy tabular data repositories. Secondly, many computer science and information design courses in universities are still mostly based on tabular, relational and tree-like data structures. Another source of misunderstandings is the use of URIs or IRIs to identify at the same time informational and non-informational resources, that is seldom accepted and understood, even within the experts community (see for example the recurring discussion about http-range-14). The issue known as the *HTTP Range 14 problem*, is about what mechanisms should be used to distinguish between statements about web pages and statements about the real world item or concepts the web page *talks about*. Along with some technical and engineering implications, what is interesting is that the distinction between information and non-information resources is not always clear for data publishers and data consumers. As a result, the web of data and many data-consuming applications present inconsistencies which derive from the misinterpretation of this concept. This is not a critical problem *per-se*, but it becomes a serious problem when reasoning is applied to these inconsistencies (e.g. same-as reasoning). Finally, the web of data is built upon the Open World Assumption (OWA)[5], according to which ”the truth-value of a statement is independent of whether or not it is known by any single observer or agent to be true. In other words, a statement cannot be considered false just because there is nothing explicitly stating that it is true. This is another important shift from the relational database world, which is based on the Closed World Assumption. The logical and the technical implications of the OWA are beyond the scope of this presentation, however it is worth spelling out some of its social consequences. The choice of operating under OWA is justified by the fact that open worlds are particularly well suited to deal with incomplete information and exceptions. OWA's features are desirable within a universal systems such as the web. However, they also pose some serious challenges. For instance, some problems are inherently related to closed worlds and most importantly many of the IT tools are designed to work in closed worlds. Furthermore, people are more familiar with thinking in closed worlds rather than in open ones. Once it has been shown which are the intrinsic cultural implications of the vision of the semantic web and linked data, the next section introduces another *problematique* which is essential for advancing a revolutionary twist in the semantic web: the importance of nurturing a dynamic linked data economy.

[5] Open World Assumption, in Wikipedia, http://en.wikipedia.org/wiki/Open_world_assumption.

Nurturing a linked data economy

In the last decade, Europe has made large investments in research over semantic technologies. This has generated brilliant ideas, core scientific knowledge and many prototypal implementations. Unfortunately, the research community has not yet been able to leverage this potential within the industry to build production-ready tools easily usable by end-users. There is not yet a Microsoft Excel, or an Apple ITunes for linked data. Similarly, there is not yet MySPARQL or any Apache HTTPD that can serve streamed linked data. A lively data economy, with a rich ecosystem, is not yet in place. The time has come to invest in innovation in order to be able to transform the enormous knowledge accumulated through research and the large amount of data recently produced/liberated into a virtuous circle able to generate a self-sustaining and evolving ecosystem. Recently a number of game-changing announcements has been made which can be considered as potentially contributing to create a linked data economy: first of all, Google Knowledge Graph, a sort of Closed Enterprise linked data cloud as well as the acquisition by the big G of one of the most important nodes of the LOD cloud, Freebase; secondly, the coalition between the largest search engines to introduce schema.org, a combination of a technology and a set of incentives for web publishers to annotate their content with semantic markup. Finally, large private organizations are approaching the web of data, by evolving their business models or by modifying their production processes to comply with the openess of the linked open data cloud, or by building closed enterprise linked data clouds (e.g. many large pharma are bolding their own enterprise linked data). On the one hand some of these announcements may raise some socio-economic issues related to the risk of endangering the public good and to monopolistic threats (see for example the interesting, if a bit outdated, analysis of the risks related to schema.org[6] published by (Tennison) on her blog.

In Europe, there are numerous small-medium Enterprises which are the major driver for innovation and economic growth. A careful strategy to protect our common knowledge-heritage and the (linked!) public good that is at the same time able to offer the right economic incentives to SMEs, is key in paving the way to a vibrant and sustainable linked data economy. On the other hand, this is clearly a huge leap forward for the web of data, whose economic value may start to unlock thanks to the critical mass (of users, investments, technology, media visibility and demand) mobilized around the leading web companies. In a recent post (Dodds) suggests that the media-hype created by Google's KnowledGraph, – that is still mostly fed by public domain and open knowledge – may represent an opportunity for SMEs which can leverage the same public goods to meet the increasing demand of vertical and custom enterprise linked data clouds.

Conclusions

After having presented some of the technological challenges for a full exploitation of the linked open data web, this paper has argued that such innovation must come along with a new understanding of the cultural changes inner to the web of data. This is, however, not enough. The web of data needs a lively economic environment where to flourish and further develop around its potential. Yes, but how? The policy advices suggested by this paper are well expressed by the European Union in the following quote which concludes this contribution:

[6] http://schema.org.

> The volume of data being digitally stored and exchanged is growing exponentially. [...] Obviously, these data generate the potential for many new types of products and services. The accessibility of public services can be improved for open and linked data, smart traffic and cities can improve mobility, products can report their life cycles, monitoring their provenance and quality, social trends can be recognized and turned into services, and products can come closer to meeting consumers' needs. We foresee a whole new industry implementing services on top of large data streams. The impact of this emerging economic sector - the data economy - may soon outrange the current importance of the software industry. The gist of the matter is to turn large streams of data into added value for the public and private sector. This industry can help to increase the efficiency of processes working with these data, it can provide transparency, support well-informed decision making, and enable new services not possible today (e.g., smart cities, interactive trend analysis or seamless data flows along value creation chains). Clearly, research, engineering, policy making for the Data Economy and the exploitation of the unprecedented wealth of data have become keys to the Future of Europe.[7]

References

Barbieri, Davide and Emanuele Della Valle. "A Proposal for Publishing Data Streams as Linked Data - A Position Paper". *Proceedings of the Linked Data on the Web (LDOW2010) Workshop, co-located with WWW2010*. 2010. (Cit. on p. 93).

Berners-Lee, Tim. "Giant Global Graph". http://dig.csail.mit.edu/breadcrumbs/node/215.

Dodds, Leigh. "Welcome to the Knowledge Graph". http://talis-systems.com/2012/05/welcome-to-the-knowledge-graph/.

[7] http://2012.data-forum.eu/about.

Sequeda, Juan F. and Oscar Corcho. “Linked Stream Data: A Position Paper”. *Proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN09)*. Washington DC, USA, 2009. (Cit. on p. 93).

Tennison, Jeni. “Schema.org and the Responsibility of Monopoly”. http://www.jenitennison.com/blog/node/157.

MICHELE BARBERA, Net7; Spazio Dati.
info@netseven.it

Barbera, M. ”Linked (open) data at web scale: research, social and engineering challenges in the digital humanities”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6333. DOI: 10.4403/jlis.it-6333. Web.

ABSTRACT: The amount of data available in the linked data cloud has grown enormously in the last years in several domains, including cultural heritage and digital humanities. However creative reuse of data both within the scholarly community and within the cultural industry is still very limited. It depends on a mixture of technical and social problems that needs to be addressed in research and within the industry. The paper will explores some of these challenges with a focus on the digital humanities.

KEYWORDS: Library linked data; Semantic web

Submitted: 2012-06-12
Accepted: 2012-08-31
Published: 2013-01-15

Publishing Value Vocabularies and Standard as Linked Data

FRBR Review Group initiatives and the world of linked data

Pat Riva

The FRBR Review Group[1] was created by the IFLA Cataloguing Section[2] in 2003 to ensure the ongoing maintenance of the FRBR conceptual model for bibliographic data described in the Functional Requirements for Bibliographic Records final report published in 1998 (*Functional Requirements for Bibliographic Records: Final Report*). The FRBR family of conceptual models also includes the Functional Requirements for Authority Data (FRAD) approved and published in 2009 (IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR)) and the Functional Requirements for Subject Authority Data (FRSAD) approved in 2010 and published in 2011 (IFLA Working Group on the Functional Requirements for Subject Authority Records (FRSAR)). The revised charge of the FRBR Review Group, approved in 2009, includes responsibility for the maintenance and development of all three models, FRBR, FRAD and FRSAD, as well as for their consolidation.

The Cataloguing Section *Action Plan for 2012*[3] puts particular emphasis on two tasks being carried out under the responsibility of the FRBR Review Group:

[1] http://www.ifla.org/en/frbr-rg.

[2] http://www.ifla.org/en/cataloguing.

[3] http://www.ifla.org/en/node/1959.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5482

- FRBR: Explore the preparation of a consolidated document for IFLA's FRBR family of conceptual models in an entity-relationship formulation
- Promote IFLA standards: Participate in the development of namespaces for all IFLA bibliographic standards, including the ISBD, FRBR, FRAD, and FRSAD and in connection with this promote and position the IFLA standards and models in the semantic web

The strategic importance of namespace creation within the overall IFLA development of standards is thus clearly acknowledged.

Namespaces for the FRBR family of conceptual models

The activities on representing IFLA standards and models in RDF started in 2007 following the joint RDA/DCMI Data Model meeting held April 30-May 1, 2007 at the British Library.[4] One of the members of the FRBR Review Group, Barbara Tillett, attended that meeting in her role as a member of the Joint Steering Committee for Development of RDA[5] and saw that the new push towards linked data was just as relevant for FRBR as it was for RDA. At its next annual meeting during the IFLA General Conference in Durban, the FRBR Review Group agreed that it was both appropriate and important that IFLA take a lead in making its models and standards available in an authoritative form for reuse. RDA is based on the

[4] Data Model Meeting, Outcomes. Available at: http://www.bl.uk/bibliographic/meeting.html.

[5] http://www.rda-jsc.org. The JSC is responsible for maintaining *RDA, Resource Description and Access*.

FRBR and FRAD models, so one option that was under consideration for an RDA namespace was to refer to IFLA FRBR and FRAD namespaces where possible. And so the Review Group decided to start a new project, Namespaces for FRBR entities/elements in RDF, with the specific task being "to define appropriate namespaces for FRBR (entity-relationship) in RDF and other appropriate syntaxes".[6] At the time the Review Group members did not have the technical expertise required, and sought the assistance of a consultant for the project.

For the FRBR Review Group's annual meeting during the 2008 IFLA conference in Québec the consultant, Gordon Dunsire, prepared the document Declaring FRBR entities and relationships in RDF (Dunsire, *Declaring FRBR entities and relationships in RDF*) which identified as issues:

- the need for a stable, branded, web domain to host the IFLA namespaces,
- that the FRBR Review Group needed to commit to the validation and maintenance of the content of the declarations.

Initial declarations of vocabularies for the FRBR entities, FRBR relationships and FRBR user tasks were carried out in the Open Metadata Registry[7] (prior to 2010 this was called the National Science Digital Library Metadata Registry).

The FRBR Review Group was easily able to validate the labels and scope notes that had been extracted from the (English) text of the FRBR final report. IFLA is a multilingual body, and so opaque URIs were chosen, with the expectation that eventually labels and scope

[6] FRBR Review Group. Meeting Report, Durban, August 21, 2007. Available at: http://www.ifla.org/files/cataloguing/frbrrg/meeting_2007.pdf, p. 4.

[7] http://metadataregistry.org.

notes in several languages would be added.[8] However, this was a new type of activity for IFLA and many technical issues remained before the namespaces could actually be published. Around this time other IFLA standards groups, particularly the ISBD Review Group, started expressing an interest in creating their own namespaces (Riva and Willer), and so it became evident that any technical solution for hosting IFLA namespaces had to be designed to be extensible. Gordon Dunsire's report to the FRBR Review Group at the IFLA annual conference in Milan in 2009[9] identified potential requirements for the management of the IFLA namespaces, including making them available in an open environment, providing dereferencing services, and managing them within a common framework. Immediately after the 2009 IFLA congress the base domain iflastandards.info was registered. The same year, at the initiative of the Classification and Indexing Section, the IFLA Professional Committee established an IFLA Namespaces Task Group[10] with Gordon Dunsire as chair. This group proposed the pattern for deriving extended base domains for each distinct namespace.
In http://iflastandards.info/ns/ the ns/ segment indicates the namespaces proper, as opposed to any standards documentation which might be made available through the basic http://iflastandards.info/ site. Adding the segment fr/ gives http://iflastandards.info/ns/fr/, the base for all namespaces relating to the FRBR family of conceptual models. Then, http://iflastandards.info/ns/fr/frbr/frbrer is the base for the entity-relationship (ER) namespace for

[8] Initial translation experiences led to the preparation of the draft guidelines document *Translations of RDF representations of IFLA standards* which was distributed for comment by the ISBD/XML Study Group in April 2012.

[9] FRBR Review Group. Meeting Report, Milan, Italy, August 25 and 26, 2009. Available at: http://www.ifla.org/files/cataloguing/frbrrg/meeting_2009.pdf, p. 3-4.

[10] http://www.ifla.org/en/node/5353.

FRBR. The classes are sequentially numbered with the letter C as a prefix, for example,iflastandards.info/ns/fr/frbr/frbrer/C1001 is the URI for the entity work. Properties use the letter P as a prefix, for example, iflastandards.info/ns/fr/frbr/frbrer/P2001 is the URI for the is realized through relationship.

Similarly, http://iflastandards.info/ns/fr/frad is the base for the FRAD namespace and http://iflastandards.info/ns/fr/frsad is the base for the FRSAD namespace.

Since 2009 the FRBR Review Group has been able to concentrate on its responsibility for developing the content of the namespaces, leaving the technical aspects to the Namespaces Task Group.[11] Our intention throughout has been to reflect the full semantics of the FRBR family models as accurately as possible within the tools available. The original IFLA reports were not written with a view to their expression in RDF, and so such essentials as the domains and ranges of relationships are not always stated explicitly and must be deduced. Turning implicit constraints, such as which relationships are transitive, which form reciprocal pairs (inverse properties), which are equivalent (or symmetric), into explicit statements required some thought.

At its August 16, 2010 meeting in Göteborg, Sweden the FRBR Review Group resolved all the outstanding questions raised in the development of a full ontology for FRBRer. The status of the FRBRer element set[12] was set to "published" in May 2011. It contains 10 classes (entities) and 206 properties (attributes and relationships), the additional semantic constraints are expressed in approximately 2000 triples in TTL (terse triple language).

[11] See the report IFLA namespaces – requirements and options from the IFLA Namespaces Task Group, March 2010, updated February 2011, at: http://www.ifla.org/files/classification-and-indexing/ifla-namespaces-requirements-options-report_corrected.pdf.

[12] http://metadataregistry.org/schema/show/id/5.html.

FRBRer declares the 10 classes (the entities) to be disjoint. This semantic constraint is plainly evident from the FRBR final report text, and in general makes logical sense – all would agree that a person is not a place, nor a work, expression, manifestation or item. Their respective FRBR definitions also indicate that a work is not an expression, which is not a manifestation, which in turn is not an item. This implies that any property whose domain is the manifestation must be disjoint from any property with work or expression as its domain. Therefore specific properties must be declared at each level, resulting, for example, in the following declarations to correspond to the relationships defined in FRBR sections 5.3.1.1, 5.3.2.1, 5.3.4.1, and 5.3.6.1:
http://iflastandards.info/ns/fr/frbr/frbrer/P2057
has part (work);
http://iflastandards.info/ns/fr/frbr/frbrer/P2079
has part (expression);
http://iflastandards.info/ns/fr/frbr/frbrer/P2085
has part (manifestation);
http://iflastandards.info/ns/fr/frbr/frbrer/P2091
has part (item).

This feature of the FRBR model has an impact on the potential for creating formal links between the FRBR properties, particularly those involving FRBR group 1 entities, and conceptually similar properties defined in other namespaces. Unless the FRBRer classes can be mapped to classes in other namespace, the properties cannot be declared as formally equivalent. As many of the ISBD data elements can be aligned intellectually with attributes of group 1 entities in FRBR, and were the historical source of the FRBR attributes, producing a mapping has been articulated as a goal by the ISBD Review Group.[13] However, establishing a formal mapping between the two namespaces is problematic since all properties in the ISBD element

[13] See goal 3.2 of the Cataloguing Section Action plan for 2012 at: http://www.ifla.org/en/node/1959.

set have as their domain the class *resource*[14] which cannot be mapped to any of the FRBRer classes.

In response to suggestions that permitting broader mappings would be useful, the FRBR Review Group has taken some steps towards defining in draft an additional set of unconstrained properties – unconstrained meaning properties without declaration of domains and ranges – which could serve as superproperties of the FRBRer properties. Thus one possible approach to creating linkages with element sets such as ISBD could be aligned at this general level.

Developing namespaces for FRAD and FRSAD followed the work on the FRBRer namespace, and all outstanding issues were resolved at the Review Group's August 19, 2011 meeting in San Juan, Puerto Rico. The FRAD element set[15] consists of 12 classes and 134 properties, with 600 TTL triples to express the semantic constraints. The FRSAD ontology[16] is the smallest, with only 2 classes and 17 properties making up the element set and 60 TTL triples to express the constraints. Both FRAD and FRSAD refer to some elements already defined in FRBRer, these linkages are also expressed as TTL triples. The status of the FRAD and FRSAD namespaces was set to "published" in February 2012.

This work has taken longer than initially imagined, but as of March 2012, all the FRBR family namespaces in the Open Metadata Registry are providing dereferencing services to the individual class and property level (Riva, "Functional Requirements namespaces published"). This enhancement to the OMR was funded by the FRBR Review Group.

[14] The class *resource* is defined as: «An entity, tangible or intangible, that comprises intellectual and/or artistic content and is conceived, produced and/or issued as a unit, forming the basis of a single bibliographic description». http://metadataregistry.org/schemaprop/show/id/2107.html.

[15] http://metadataregistry.org/schema/show/id/24.html.

[16] http://metadataregistry.org/schema/show/id/26.html.

Aggregates Working Group report

The Working Group on Aggregates was formed in 2005 and submitted its final report on September 12, 2011 (IFLA Working Group on Aggregates). It describes three categories of aggregates: aggregate collections of expressions, aggregates resulting from augmentation, and aggregates of parallel expressions, and provides well-chosen examples of each. When the Working Group was first constituted, the expectation was that it might recommend an amendment to FRBR to clarify the treatment of aggregates. However, as the FRBR Review Group has begun working on a consolidation of the models, the report will be considered in the context of the consolidation project.[17]

Consolidation

The FRBR Review Group's priority activity is the production of a consolidated statement of the conceptual models in the FRBR family. Having three documents written over such a long period of time and by different working groups (two of which functioned concurrently) is inconvenient for application development as there is no official statement of the interrelationships between the models. The three reports reflect different approaches and the evolution of thinking over time. This is evident just in the names of the models: FRBR includes the word records in its name, but actually models data; the model initially named FRAR (Functional requirements for authority records) was renamed FRAD (Functional requirements for authority data) to reflect this realisation. Of the three models, FRSAD takes

[17] FRBR Review Group. Meeting Report, San Juan, Puerto Rico, August 15, 2011. Available at: http://www.ifla.org/files/cataloguing/frbrrg/meeting_2011.pdf. Item 6.2 on p. 4.

the most general approach, FRAD the most detailed. Some of the intended interrelationships are obvious, for instance, the entity family defined in FRAD is clearly intended to function as an agent along with the FRBR group 2 entities person and corporate body, and thus any relationship that can involve person or corporate body should be extended to family. By forcing the FRBR Review Group to carefully examine each entity, attribute and relationship defined by the three models to select appropriate domains and ranges for the properties, and to make explicit any implicit constraints and explicitly declare the intended semantics of the properties, the development of the namespaces has definitely laid important groundwork for the consolidation process (Dunsire, "Interoperability and semantics in RDF representations of FRBR, FRAD and FRSAD"; "Representing the FR family in the Semantic Web"). The process of developing the namespaces also highlighted areas which pose particular challenges in the consolidation process. The treatment of subjects is one of these; the concept of "having a subject" is viewed as a relationship in FRBR and FRSAD, but as an attribute of the entity *work* in FRAD ("Subjects in the FR family"). Another complex area is the treatment of names. FRAD defines three interrelated entities name, identifier, controlled access point; FRSAD just defines a single entity nomen, which might be viewed as a superclass of the three FRAD entities. Examining the attributes of these entities highlights the importance of the context of the name use (Doerr, Riva, and Žumer), and this in turn offers some insight into the FRAD definition of person as a bibliographic identity, or a name that a real person uses in a specific context. At its most recent working meeting on April 25, 2012 (Riva, "Report from the FRBR Review Group mid-year meeting"), the FRBR Review Group concluded that the aim of the consolidation process should be to define a coherent model of the bibliographic universe. This is to be done using the three existing models, as well as the Aggregates report and insights from the development of the namespaces and FRBRoo, but that a fresh approach to certain issues can be taken when warranted. Some of the conclusions reached so far include the intention to retain the FRSAD general model of the subject relationship, with no detailed typology of subject entities, which results in the decision to functionally deprecate the FRBR group 3 entities, concept, object, event, place. To provide guidance in the consolidation the Review Group started with an examination of the users and user tasks as defined in the three models. The proposed combined user tasks is as follows:

Find. To search on any relevant criteria in order to bring together information about one or more resources of interest

Identify. To determine the suitability of the resources found and to distinguish between similar resources

Select. To choose (by accepting or by rejecting) specific resources

Obtain. To access the content of the resource

Explore. To use the relationships between one resource and another to place them in a context

The Review Group intends to simultaneously produce a textual description of the consolidated model and to declare it as a namespace. This will require careful indication of versions for any classes and properties whose semantics are changed. Following IFLA's normal procedures for such documents, the draft descriptive text will be made available for world-wide review and the comments received resolved by the FRBR Review Group, prior to recommending its approval by the Cataloguing Section.

FRBRoo

The discussion so far has referred entirely to the FRBR conceptual models in their original “classic” entity-relationship formulation. In parallel with these developments, in 2003 the FRBR Review Group formed a joint working group[18] with the International Council of Museum’s Committee on Documentation (CIDOC) CRM Special Interest Group,[19] with the goal of preparing an object-oriented statement of FRBR as a compatible extension (Riva, Doerr, and Žumer) of the CIDOC CRM (Conceptual Reference Model).[20] FRBRoo version 1.0[21] was approved and published in January 2010. The focus of the four meetings since then has been to include the entities, attributes and relationships defined in FRAD and FRSAD in FRBRoo.[22] The most recent meeting was 30 April-3 May 2012 in Heraklion, Crete, Greece. Version 2.0 of FRBRoo is now close to completion and will be released for comment. An RDF representation of FRBRoo version 1.0.2 exists but not in the Open Metadata Registry, current plans are to import FRBRoo version 2.0 into the OMR as soon as the bulk import feature is available.

As with any translation process, these cross-community alignment discussions have revealed unstated assumptions, imprecise definitions, apparent inconsistencies and a myriad other issues and ideas, all of which are providing valuable input into the FRBR Review Group’s consolidation process.

[18] http://www.ifla.org/en/node/928.

[19] http://www.cidoc-crm.org/who_we_are.html.

[20] http://www.cidoc-crm.org/definition_cidoc.html.

[21] http://www.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V1.0.1.pdf.

[22] Meeting minutes available at: http://www.cidoc-crm.org/special_interest_meetings.html.

References

Doerr, Martin, Pat Riva, and Maja Žumer. “FRBR entities: identity and identification”. *Cataloging and classification quarterly* 50.5/7. (2012). (Cit. on p. 113).

Dunsire, Gordon. *Declaring FRBR entities and relationships in RDF*. 2008. http://www.ifla.org/files/cataloguing/frbrrg/namespace-report.pdf. (Cit. on p. 107).

—–. “Interoperability and semantics in RDF representations of FRBR, FRAD and FRSAD”. *Concepts in Context: Proceedings of the Cologne Conference on Interoperability and Semantics in Knowledge Organization, July 19th–20th, 2010*. Würzburg: Ergon Verlag, 2011. (Cit. on p. 113).

—–. “Representing the FR family in the Semantic Web”. *Cataloging and classification quarterly* 50.5/7. (2012). (Cit. on p. 113).

—–. “Subjects in the FR family”. *CC:DA/SAC joint meeting, 27 June 2011*. 2011. http://gordondunsire.com/pubs/pres/FRSubjects.ppt. (Cit. on p. 113).

IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional Requirements for Bibliographic Records: Final Report*. München: K.G. Saur, 1998. Available with amendments and errata at: http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records. (Cit. on p. 105).

IFLA Working Group on Aggregates. *Final report*. 2011. http://www.ifla.org/files/cataloguing/frbrrg/AggregatesFinalReport.pdf. (Cit. on p. 112).

IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR). *Functional requirements for authority data : a conceptual model*. Ed. Glenn E. Patton. München: K.G. Saur, 2009. (Cit. on p. 105).

IFLA Working Group on the Functional Requirements for Subject Authority Records (FRSAR). *Functional requirements for subject authority data (FRSAD) : a conceptual model*. Ed. Marcia Lei Zeng, Maja Žumer, and Athena Salaba. München: De Gruyter Saur, 2011. Pre-print available online at: http://www.ifla.org/node/5849. (Cit. on p. 105).

Riva, Pat. “Functional Requirements namespaces published”. *SCATNews* 37. (2012). <http://www.ifla.org/files/cataloguing/scatn/scat-news-37.pdf>. (Cit. on p. 111).

—–. “Report from the FRBR Review Group mid-year meeting”. *SCATNews* 37. (2012): 15–16. <http://www.ifla.org/files/cataloguing/scatn/scat-news-37.pdf>. (Cit. on p. 113).

Riva, Pat, Martin Doerr, and Maja Žumer. “FRBRoo: enabling a common view of information from memory institutions”. *IFLA WLIC 2008*. 2008. http://archive.ifla.org/IV/ifla74/papers/156-Riva_Doerr_Zumer-en.pdf. (Cit. on p. 115).

Riva, Pat and Mirna Willer. *Declaring IFLA ISBD and FRBR family of conceptual models in RDF*. 2012. Presented at five years on, jsc/dcmi seminar, april 27, 2012, british library. (Cit. on p. 108).

PAT RIVA, Bibliothèque et Archives nationales du Québec, Canada and IFLA FRBR Review.
messmer@bsb-muenchen.de

Riva, P. ”FRBR Review Group initiatives and the world of linked data”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5482. DOI: 10.4403/jlis.it-5482. Web.

ABSTRACT: The FRBR Review Group has the mandate to review and maintain IFLA's family of conceptual models and to develop guidelines and interpretive documents to assist in the application of the models. In September 2011 the Aggregates Working Group submitted its final report, clarifying the modeling of three distinct types of aggregates within an FRBR framework. Since the approval of the FRSAD model in 2010, the FRBR Review Group has been concentrating its efforts on the development of a consolidated statement of the three models (FRBR, FRAD, FRSAD) in the entity-relationship formalism. The interrelationships between the models are particularly illuminating with regards to the subject entities and relationships and the “naming” entities. In parallel, the FRBR/CIDOC CRM Harmonisation Working Group has been working to extend FRBRoo to include the entities, attributes and relationships defined in FRAD and FRSAD. A third essential aspect of the FRBR Review Group's work is the development of a series of namespaces for the entities, relationships, attributes and user tasks as defined in the three models. This process has raised a number of interesting questions and often clarifies the underlying semantics of the models.

KEYWORDS: Library linked data; Semantic web; RDF

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

ISBD adaptation to SW of bibliographic data in linked data

Elena Escolano Rodrìguez

Introduction

The achievement of any business model depends on the users' belief, the concept is known as "true economy";[1] belief in truth and in the quality of data will be the best investment in the future of ubiquity. Consequently, the emphasis will be on the quality of this information. Indeed it is very important it is controlled and truthful. In this context the source of information is a relevant value, given that it is the source that lends authority to data. If expressed as linked data, the information which has at length been selected, controlled, validated, recorded and structured in cultural institutions as libraries' databases, will certainly take on an important role. In this way, libraries, museums, and archives can retake a prominent position on the web; they have traditionally selected, structured and organized information and have at the same time contributed to convert information into quality data. Nowadays, the need to provide access to both integrated heterogeneous knowledge and distributed homogeneous knowledge in several domains is considerably grow-

[1] http://www.chiefmartec.com/2010/03/business-models-for-linked-data-and-web-30.html.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-6305

ing. The objective is to use the web as a single global database, so the linked date tool will allow for this global web. It is necessary to recognize the ever increasing importance attached to the discovery of objects both digital and not. Therefore, the data that describes the objects should be available there where users are, integrated in a global web which means they should be open and as a result reusable. In the last years the semantic web cloud has increased in a significant way. This is confirmed by the growth occurred from 2007 to 2010 and up to September 2011;[2] the rise becomes evident observing the graphs in the green part, concerning the publications domain. This phenomenon underlines the importance for library community to have their structured and controlled data available on the web with this new tool. Consequently, for libraries this will represent:[3]

- the extensibility and the integration of their data with those of other institutions, with the consequent increase of information that users can retrieve; in this way it would be possible to complete, aggregate and link the library data with other structured information in different ways, in accordance with other standards. The integration can guarantee a greater and better service to the user, not only by virtue of the involvement of museums and archives but thanks to the collaboration with new centres and institutions, or with other products and information sources; this process would allow the widespread diffusion of information recorded by libraries and the integration with other web segments, such as Wikipedia, Geonames, etc.;

[2] http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.html.

[3] Linked Data Incubator Group wiki: http://www.w3.org/2005/Incubator/lld/wiki/Draft_Benefits and http://www.w3.org/2005/Incubator/lld/wiki/Benefits.

- the compliance with requests for public administration transparency; the libraries of public institutions could take charge, upfront, of the planning and coordination of this political action, avoiding duplications of projects within the same sector which should be linked to each other. In this process the libraries could take on a real crucial role;
- the semantic understanding of library language, so far quite unclear for the users, can contribute to optimise the results obtained during the search and consequently the whole library services;
- the possibility of presenting the search results in the user's language. In the future, it will allow, or at least make it easier, to define large cooperation areas, thanks to the automatic conversion in the linguistic form accepted by a specific community. These applications will also concern and improve the multilingual cooperative cataloguing, that is the records could be created and presented in one language without the necessity of creating a new "record".

IFLA contribute

The work carried out by the IFLA ISBD Review Group has also been carried bearing in mind this scope: making available in the cloud the information stored in our database, or as explained in the consolidate ISBD edition at paragraph "A.1.2 Scope": "improving the portability of bibliographic data in the semantic web and consequently the interoperability of the ISBD standard in connection with other content standards (International Federation of Library Associations and Institutions. ISBD Review Group and International Federation of Library Associations and Institutions. Cataloguing

Section. Standing Committee p. 1). Linked data is necessary for participating in the web of data, but for taking part in the semantic web, putting data on the web and link them is not enough: there are other necessary requirements which, according to Berners-Lee ("Linked Data - Design Issues"), are:

1. using URI for identifying or referring to sources. The URI (Uniform Resource Identifier) is the characters set used to indicate univocally the names of the resources on the web and are expressed in a machine-readable form;

2. using HTTP URIs, so that the user can look for and locate resources through them (this is called dereferencing)

3. providing useful information about the resource when we search it with URI, using standards (for example RDF, SPARQL);

4. including links with other URIs for finding out linked information.

The studies on the semantic web are specifically focused on formal ontologies, that is, the logical structure in which the semantic of a particular domain is organized. Aiming at integrating and managing the knowledge of this dispersed information, the research has also contributed to facilitate relationships between ontologies, specifying their context clarify how widespread knowledge is related to several resources. Some information can be automatically captured and information related to the source can explain the context. So it can be reduced the presence of non-intentioned or unwanted meaning in the ontology, obtaining a greater clearness and facilitating the analysis and the search. In order the libraries participate in the semantic web, it was necessary to create the ontology that reflect the logical structure of the library domain, providing useful information to make it understandable. Particularly, in the library field there was much work done on definition of a common well structured and standardized basis, represented by IFLA very much consolidated standards, which ensure quality, exchange capability and sustainability. This regulation encourages the right development of the semantic web, because standards are important components for linked data. IFLA's contribution and participation to this process is justified not only because the utility and the importance it has for libraries but also for additional reasons: it was considered essential to protect the own terminology, at the same time specifying the context and the origin of the metadata (that is a very important issue in linked data). In addition, for ISBD it was a main objective to reposition the IFLA standard and its value as important tool for the delivery and reuse of structured authorized bibliographic data in the Internet environment. IFLA has carried out several actions. First of all, it decided to declare its own models and standards in the Resource Description Framework (RDF). It was followed the recommendation, by the advisor Gordon Dunsire, to the FRBR Review Group in 2008 , and to the ISBD Review Group in 2009 . Its application was decided and authorized during that meeting of the ISBD Review Group, at the IFLA Conference held in Milan in 2009. Work started on the declaration of ISBD set of elements in RDF, in order to present and be submitted to IFLA Cataloguing Section for approval as part of the ISBD consolidated edition of 2011. For such purpose, it was necessary to create a namespace that would properly identify the URIs of RDF declarations by IFLA for its own models and standards, what was recommended in the 2008 report. The recommendation consisted in protecting elements, terms and definitions related to the IFLA models and standards, using a sort of brand, to save them from unlike interpretations from other standards. This action helped also to achieve what has been mentioned before related to the business model, in which basis to obtain results or indirect benefits, it is necessary this quality brand. To carry out this work, study and elaboration a Namespace Group, coordinated since 2009 by Gordon Dunsire, was created within IFLA with the objective of management of the IFLA standards declarations. When establishing the namespace, several issues were taken into account. The namespace had to be clear, short, expanded and applicable to each model and standard. It was decided to adopt the URL form (which begins with http://...), that in the future may be dereferenced, in order to retrieve the RDF or the HTML file when the URI is processed as an ordinary URL. Once the decision was made, the focus was on the namespace structure: it was decided to adopt this quality mark: http://iflastandards.info, considering the potentiality of URL to be intelligible both for computers and humans. Following, the abbreviations of the standard referred would be identified, such as for instance: http://iflastandards.info/ns/isbd/elements. Then it should be considered how to identify the element in the URI. URIs can contain letters and numbers. It could be useful to remember that URI is specifically defined for machine understanding, it is not a label intended for the user even if it can guide him. In fact, the context of an element could be briefly identified with one word but at risk of misleading him to believe this textual information is similar to a label: the label in itself is not sufficient, the programmer, the human being, has to read the full declaration corresponding to the URI with its definition for the correct application. Due to these reasons debated at IFLA General Conference held in Gothenburg in 2010, after which it was decided that URIs would be opaque, without reference to a specific language, because IFLA has to recognize and encourage the multilingualism; therefore, in order to guarantee linguistic neutrality, a numerical solution should be adopted. An opaque URI would also extend its use to linguistic communities different from the English ones ensuring, at the same time, access to these ontologies in other languages without the necessity of creating independent URIs. The declarations contain important information such as metadata name, label, definition, notes used for extending the information or its application, the filiation (whether it is property or sub-property), the state of acceptance, etc. The utility of translation affects definitions, notes and also the labels. Using an opaque URI and specifying the language in which you desire to obtain the information, it is possible to collect all declarations in different languages with the same URI. If an opaque URI had not been used, it would have been necessary to create one for each language to be afterwards linked to the others as ″same as″. The problems related to translation will be further developed bellow. The labels refer more to the comprehension of the programmer than to the machine; it was necessary to disambiguate and adapt them because the relations present in FRBR are coincident for several entities (in RDF classes) so it was necessary to specify the domain of the relationship; and also in some cases it was not clear the relationship orientation (the range in RDF). For example in Italian: ″ha come forma variante″ is a relationship (property or subproperty for RDF) which can be applied both to the entity/class ″Person″ and ″Corporate Body″. Therefore it was necessary to add information in brackets, to identify more specifically the classes which the properties belong and the direction of the relationship. Both FRBR models and ISBD standard include controlled vocabularies. In the former case it consist of the user′s tasks while in the ISBD correspond to the terms used for Area 0: : Content form and media type. Vocabularies were identified by completing the URI with the expression ″terms″ http://iflastandards.info/ns/isbd/terms/ and the notation which indicates the concrete term of the normalized vocabulary. It stands evident that the vocabulary, which is recorded in the language of the cataloguing agency, would be simply converted into another language when the record or the information is captured by other agency, especially in a cooperative environment, in a controlled, normalized and automatic manner. In this way, it is also possible to map or create correspondences with similar vocabularies but not structured in the same way; as for example with the Resource Description & Access (RDA) vocabulary used to describe content and support: Content Type and Media Type. However this is not possible without problems for establishing an equivalence relationship of ″same as″ type, because, in some cases, there is not a total correspondence 1=1. Until now these declarations have been made in Open Metadata Registry, which a space created by the W3C is containing several ontologies about different domains, but in the future it will be possible to transfer these declaration to a specific section in the IFLA website where they can be hosted and managed. Regarding the sustainability and maintenance of the IFLA Namespace is sill and issue on course.

Multilingualism development

The basis for the semantic web is basically in English, which has worrying consequences about cultural and linguistic diversity. Even if English is recognized a IFLA working language, there are also other six official languages that require the development of multilingualism. The first issue of the ISBD/XML Working Group plan, approved in November 2011 , states the intention of promoting the translation of ISBD and the declarations in OMR, in addition of the definition of guidelines for translators. From my participation, on several occasions, in debates concerning the translation of IFLA declarations, I am going to highlight some issues that affect many Latin languages such as Spanish, Italian, Croatian, Slovenian etc. The significant topics discussed are the following:

Style issues

As far as labels are concerned, there is a good practice, drawn from some communities of the semantic web, to use capital letters for classes names in RDF. Moreover, in English, words are joined together, what is called CamelCase, for instance the ISBD subclass: ParallelTitleCompoundEncodingScheme, but this is not possible to apply to Spanish. In some cases the use of capital letters could be accepted, even for prepositions while the conjunction without space is not accepted. Therefore, in Spanish it was accepted the use of capital letters for the first letter of the first word or for every word, but without joining the words. Another issue regards the property labels: they are always verbal phrases. In fact their aim is to serve as predicate in the RDF triple RDF: Subject – Predicate – Object. With respect to Latin alphabets according to the best practices used for the semantic web community, it is recommended to write in lower case. From the beginning it was adopted the convention to avoid, as far as possible, to use the indefinite articles, when possible, with the aim of normalizing and reducing the length of the labels. Likewise, when having to choose between the singular and plural, it was preferred to use the singular, whenever possible. These decisions were also applied during the creation of the ISBD set of elements, since the standards were being revised at that time.

Sources of reference

In the RDF data model, the source of reference, the text of the standard, is essential for programmers and developers; indeed, they could use and consult it as an additional aid to make a better semantic contextualization of the property. From the start, a decision taken by the FRBR Review Group was that labels, definitions and scope notes of the RDF's framework would be kept aligned and matched, as possible, with the text of FRBR; this would have the advantage of allowing natural language processing. In particular, labels would have match with the text accepted in the standard; concerning definitions, their alignment with the text is important, even if few modifications are required to adjust it to the context and making it understandable, that is, they will be as extracts; in the case of scope notes more flexibility in the compiling is also allowed. At the beginning, as there was no experience on which to base our work, the way we decided to follow for translating into Spanish these dispositions (labels, definitions, notes), was inevitably their literal translation from the English version. This decision presented lot of problems for the comprehension of the text in the language of translation, Spanish, and, at the same time, for the respect to the official standard text. In the case of labels, for example, it was necessary to add prepositions to help interpreting the properties, which were not in the English version of the FRBR report. We continued to work in this way for a while. However, after facing many situations, partially already cited, which were useful as forced the group to reconsider certain issues and decisions. As, for example, it was considered that the Spanish developers would have preferred to use the official Spanish translation of the FRBR report as reference source and, therefore, the declarations would have to align with the official Spanish version, instead of the English one. That revoked the initial decision to base the Spanish translation of the RDF declarations on that available in English, always avoiding the semantic ambiguity. Two solutions arised: If available, we would have to use the official translation of the reference source for the declaration in RDF, and if not present, it would be necessary to base the translations on the English declarations, concerning labels, definitions and scope notes. Obviously, if translations of reference sources (standard and models) are not updated represent other serious problem. Even for this case, it was thought various possibilities: full translation of labels, definitions and scope notes (that in the Spanish panorama had already been made for the elements of ISBD); or, more simply, translation of labels (matching with the state of art of the translations in Spanish of the model of the FRBR family). During the meeting of the ISBD/XML Study Group, November 4, 2011 in Edinburgh (UK), Gordon Dunsire stressed the ambiguity of the term ″statement″ in the OMR as meaning ″aggregated elements″, whereas in ISBD the meaning for such term is ″the information from the source″ (eg., ″1.4.5.10 parallel title and parallel statement of responsibility″). It has been recognized as necessary and urgent that the ISBD Review Group revise the current labels and definitions in OMR, and that the ISBD/XML Review Group provides a report on the possible need of change from the work on the ISBD application profile that is being prepared.

Qualifications

As previously mentioned, in the declarations concerning properties, especially in the case of the FRBR models relationships, it was necessary to use parentheses for the disambiguation, as there is homonymy depending on whether it applies the relation to a kind of entity or to another. As the name of the relation is the same even if it is applicable to different entities, it was necessary to do a disambiguation by adding consecutively, in brackets, the main class of the property (that is the domain); the necessity to disambiguate the second term of the relation has required the adoption of further brackets for the orientation of the relation (the range). The activity of translation was also useful as revision for these qualifications; it appeared that the use of parentheses to other languages than English was not clear nor systematic.

Vocabularies

OMR contains also the declarations of the model controlled vocabulary; in the case of ISBD, the vocabulary consist on the terms used in the Area 0 of the description Content form and Media type). Vocabularies are designed for their display to the user. It has been used Simplified Knowledge Organisation System (SKOS) for their representation. With regard to translations, we face the following problem: the grammatical flexion for masculine and feminine adjectives in Spanish and in other languages. Specially in the case of ISBD qualifiers for the content forms of sense and type of content - mainly the latter - that, in Spanish and other languages, have the same gender declination of the names from which they depend; that is, different declination when it is masculine or feminine adjective to be used. Thus we have:

- Imagen (cartográfica) and Objeto (cartográfico);
- Música (notada) and Movimiento (notado).

During the process of translating the vocabularies of ISBD Area 0 we, Spanish, opted for the compound form with slash, that is, "cartográfico/a", that could be used in this way, although if not matched with natural language, or by giving the option to libraries to use the most convenient type. This is the model that languages similar to Spanish have followed in their translation of the Area 0, but elsewhere, for example in the publication of ISBD's examples, we used the simple form. This solution is not applicable, however, for the disposition in SKOS, that provides the preferred label (prefLabel) and does not allow more than one prefLabel per language. The preferred label is the one we expect will be used for the friendly display and that contains the semantics. The case is still under study and debated within the Namespace Task Group. For the moment, the Grupo de Ingeniería Ontológica of the Universidad Politécnica de Madrid proposed a possible solution with the ontology of LIR / Lemon. It will take time to explore and implement it. Therefore, at the moment, and to avoid the compound form that would not coincide with the natural language, we have declared the two forms of masculine and feminine as SKOS ″alternate labels″, excluding a preferred label. The solution seems to be consistent with SKOS, problems will arise during the applications that require the preferred label for display purposes. As a label in the SKOS model cannot be alternative and preferred at the same time, in the future it will be necessary to eliminate one or both. The subject is under study and affects many languages. There are rules that have different vocabularies from those accepted by IFLA, with this tool will be easier to map (or find matching) with those vocabularies, and simultaneously link with other languages. The recording in OMR allows the status of publishing at different levels, and each component of the triple can have its own status. There are no fixed rules for the status of the record, but the general use is that definitions should remain in the same way when the status is ″published″ and labels and scope notes can be changed.

Recent actions

It is not possible to say that all has been said in advance influenced the changes that can be perceived in the new consolidated ISBD edition, but indeed they have certainly a prominent impact on several ones and have also motivated some decisions. IFLA has always had the objective of updating standards to the technological innovation with the aim to support all different kind of libraries improvement, but without forgetting the cumulated experience over the year and the different status and resources libraries could have, so it is fundamental that IFLA developments are useful for any kind of library, ensuring the scalability. The approach from ISBD view comes from the recognition that the human judgment and logic are essential as criteria to select the value data and record it in the description of the resource. This does not mean undervaluing or diminishing the technology that permits the automatic collection of data, but the rules have to be created with the goal of ensuring the quality; and the tools used could change according to the moment, the situations and the cases. On 31st January 2011 the IFLA cataloguing section approved the new consolidate edition published by De Gruyter Saur in July. The changes that could be noticed include those concerning primarily the review of a standard: variations of editorial style and changes in the introduction for better orienting its application, revision or addition of new definitions for removing ambiguity and other examples are included. Once the Review Group clarified what "data element" means, it was possible to modify the standard making it less repetitive, more consistent, easy and logic to apply. The main modifications in the final edition are due to reasons above mentioned, that have compelled to a careful analysis of the ISBD elements; to the search of a better consistence and quality of data, as well as the interoperability among these "data elements" and those coming from other standards. During the ISBD group meeting held in November 2011 emerged the following considerations:

- DC Application Profile is still under development;
- the term "statement" is ambiguous: in OMR the meaning is "aggregated elements", while in ISBD "the information from the source" ;
- collaboration with JSC on the development of a representation of the RDA/Onix Framework in RDF;
- mapping between ISBD Area 0 and the RDA/Onix Framework;

- liaison with DCMI and with appropriate groups on translation issues;
- mappings between the ISBD and UNIMARC namespaces.

As said before, standards are important references for linked data and for semantic web. During the development of the new RDA cataloguing standard, the ISBD review group suggested that it would be useful a meeting with JSC to discuss further the differences and the similarities between them, but it was after its publication in 2010, that it was possible the meeting. ISSN Network was also interested in taking part in the meeting and renovate a tripartite agreement that was achieved in 2000 . This is coincident with another purpose of the ISBD Review Group regarding "the necessity to continue activities on harmonization of the ISBD, ISSN, RDA and other national and international cataloguing rules aligned with the FRBR model and the International Cataloguing Principles" and this tripartite meeting will be the first step toward the harmonization. The meeting addressed the topics identified as potential difficulties to the realization of interoperability among ISBD, ISSN and RDA. During the meeting to homogenize ISBD-ISSN-RDA, (Glasgow 3rd-4th November, 2011) many common issues were dealt concerning RDA rules, among them:

- sources of information;
- elements which have the same name but different definitions;
- criteria for the order of selection of the information;
- different interpretation of the data nature.

In this context, the harmonization has meant functional interoperability, so that records created according to at least one of these rules would be reusable by an agency employing one of the other standards. Both ISBD Review Group and JSC agreed with the creation of correspondences between set elements of the two standards in RDF. As regards the diverse vocabulary which both standards recommend for the content form and media type, it was considered that a strict mapping between ISBD and RDA was not possible. Both standards are based on RDA/ONIX Framework for Resource Categorization document, which contains the general categorization used by other community's standards. Therefore during the meeting in Glasgow it was decided to declare this categorization in RDF and to make the mapping from each standard to RDA/ONIX categorization. The definitions of the ISBD and RDA elements will be reviewed to determine if they are semantically equivalent or if they are subproperty to each other. Indeed it was decided to update the Appendix A of RDA by the ISBD Review Group and to include the existing mappings and guidelines will be developed in an ISBD application profile for RDA. Meanwhile the DCMI/RDA Task Group has become Bibliographic Metadata Task Group[4] in which ISBD Review Groups is represented. Its aim regards the definition of components of current and emerging library, publishing, and related bibliographic metadata standards as RDF vocabularies for use in developing Dublin Core application profiles and semantic mappings. Consequently, its tasks are: Explore "obvious" mappings between known element sets and between value vocabularies, and identify issues, solutions, etc.; Give feedback on the draft ISBD-RDA/ONIX alignment and methodology; Give feedback on the draft IFLA guidelines on translations of namespaces. It is necessary to wait for the report of the meeting held in London 26th April 2012, in which ISBD had much attention.

[4] http://wiki.dublincore.org/index.php/Bibliographic_Metadata_Task_Group.

Conclusions

In conclusion, the importance of the data is growing considerably and its validity and authority is becoming fundamental, this is mainly due to the relevance of its reuse considering that if it was not true it would generate and increase mistakes, creating erroneous relations. We hope that the purpose and means for the dissemination of IFLA standards, through the namespace, and the importance of considering IFLA standard as reference model, will have the expected reception. At the present moment the IFLA standards are recognized as semantic web standards for bibliographic metadata.[5] In addition, many on-going projects are using the IFLA URIs, such as the Universitätsbibliothek of Mannheim[6] which is adopting the ISBD URI and the British Library.[7] Also the Deutsche Nationalbibliothek which is using URIs for the representation of FRBR entities in RDA, is to replace them with the official version of the IFLA URIs,[8] while the Biblioteca Nacional de España that has presented in December 14th 2011 its project of catalogue published in linked data, using the IFLA ontologies: FRBR, FRAD, and ISBD. Obviously, that is not the end of the matter. In the article it has been seen the efforts for adapting the standards to the new environment, the semantic web, and also the efforts to facilitate its comprehension, so that the information which come from different libraries that apply various standards can be linked and interoperable. It is making headway toward the semantic web and the link among libraries, their languages, formats, etc., but aren't we forgetting the opposite part of the cloud not directly related to our competence, which is not structured ac-

[5] http://www.w3.org/2005/Incubator/lld/wiki/Library_Data_Resources.

[6] http://data.bib.uni-mannheim.de/dokumentation_en.html.

[7] http://www.bl.uk/bibliographic/datasamples.html.

[8] http://www.dnb.de/DE/Service/DigitaleDienste/LinkedData/linkeddata_node.html.

cording to our standards. Many people have said that the librarian language is obscure to users, at this respect linked data could make easier its understanding and guarantee the communication with others languages. IFLA has contributed for favoring this process, now it is necessary that the developers adapt systems to this new technology. our bibliographic universe.

References

Berners Lee, Tim. “Linked Data - Design Issues”. (2006). <http://www.w3.org/DesignIssues/LinkedData.html>. (Cit. on p. 122).

International Federation of Library Associations and Institutions. ISBD Review Group and International Federation of Library Associations and Institutions. Cataloguing Section. Standing Committee. *ISBD International Standard Bibliographic Description*. Berlin: De Gruyter Saur, 2011. (Cit. on p. 121).

Elena Escolano Rodrìguez, Biblioteca Nacional de España.
elena.escolano@bne.es

Escolano Rodrìguez, E. ”ISBD adaptation to SW of bibliographic data in linked data”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6305. DOI: 10.4403/jlis.it-6305. Web.

Abstract: Linked data is the current paradigm. All works, projects and applications have it as main tool, because of its potentiality. In linking data, the selected information, authorized, validated, recorded and structured in their databases for long time by cultural institutions as libraries, necessarily is going to play an important role. The work carried out by IFLA ISBD Review Group has had this goal in mind, to make possible that the information in our databases will be in the cloud, that is, ”enhance the portability of bibliographic data in the semantic web environment and the interoperability of the ISBD with other content standards” as said in the Purpose of Consolidated ISBD, 2010. Many voices have spoken about the obscure language for users that represent the library language. Linked data can help in making it understandable. To reach this big objective much more, work than converting information in linked data is necessary. Declarations in RDF, definitions, and translations are essential to make really multilingual understanding, not only English semantic web.

Keywords: Namespace; ISBD/XML; FRBR; Semantic web; Library Linked Data Project

Acknowledgment: Paper based on the Lectio Magistralis in Biblioteconomia, held at the University of Firenze, on March 6th, 2012 within the Master biennale in catalogazione, directed by prof. Mauro Guerrini. I would like to thank Claudio Venturato for the Italian translation of the paper.

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

RDA and the Semantic Web, Linked Data Environment

Barbara Tillett

Connecting to friends, colleagues, customers, and others on the Internet is an everyday experience for most people these days. We use email, Twitter, Facebook, and other social networking systems quickly and easily when there is wifi or an Internet service provider that reaches our geographic location – even as we move around. This change in our communication systems even extends sometimes to replacing phone calls with communications like Skype or Facetime. A former phone communication can now be a multi-media experience where you not only talk but also see each other (or groups of people), share pictures or videos or documents quickly and easily all at the same time. Where are libraries in this world?

Our collections are being digitized and information about our traditional and digital collections is being made accessible through the Web. Our bibliographic descriptions and the information that we provide about the people, families, and corporate bodies used in our descriptions is available for re-use. This has been demonstrated through the popular Virtual International Authority File (VIAF) and various linked data projects, such as the Library of Congress posting of the Library of Congress Subject Headings (LCSH) and other controlled vocabularies on the Web at id.loc.gov. These projects also have shown that the data libraries provide can be very useful in

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-6303

a linked data environment for mashups and pathways to related information that may be of interest to the Web searcher – either through showing them added facets they may wish to consider to refine their search or suggesting new directions or related resources they may also like to see. Library data about our resources is no longer just meant to be stored in catalog drawers as the inventory tool to access an individual library's collections. It can now be made available to anyone, anywhere in the world, at any time.

Libraries share data because they find they collect similar things and can re-use the descriptions of the manifestations collected by other libraries or re-use the authority data about a person, family, or corporate body they also have represented in their collections. But unlike the start of sharing bibliographic and authority data among libraries (mostly by exchanging MARC formatted records or before that through buying catalog cards for the new acquisitions), the data can be shared globally rather than "exchanged" for redundant storage locally. That is, it can be made available in a linked data environment, so libraries do not need to replicate the same data over and over, but instead share it mutually with each other and with others using the Web. This environment helps reduce the library costs and expands the accessibility of information. Library data need not be just in the form of a citation in a bibliography or other linear listing, but the descriptive and authority data can be re-used and packaged in creative new ways that can be context-appropriate to a user's needs.

Library data is dynamic. It can actually start with the descriptive data and identifying information that the creator of a work provides, augmented by information from a rights management organization, a publisher or manufacturer or distributor, further enhanced by a cataloger to provide a classification and/or subject terms to help find that resource, while others may add more content information or expand the relationships, and maybe even later the data can be enhanced by a scholar with some special expertise or knowledge about the resource or the topic it covers. Many other related resources can be connected to the description and pathways emerge in the bibliographic universe. The role of libraries in this chain of description and access is an important one to help organize information and to connect users to the information they are looking for. We are part of the information network and have a tremendous legacy to contribute. Since the late 1990's, the library world has worked towards increased sharing based on agreed international principles (IFLA's International Cataloguing Principles (*ICP*)), internationally accepted conceptual models with their user tasks and recommended mandatory data elements (IFLA's FRBR (*FRBR*) and FRAD (*FRAD*)) as the foundation for how we are re-visualizing our descriptions of our resources. These changes caused us to re-examine our cataloging standards, like ISBD (*ISBD*), and our cataloging codes, like the Anglo-American Cataloguing Rules (AACR) and led to changes to position libraries for the future. Certainly the reconceptualization of the AACR was a huge milestone to move libraries towards seeing their descriptions as more than citations intended only for a library catalog. The move to Resource Description & Access (RDA) is a move towards an internationally shared vision and internationally maintained content code intended for the digital environment that we now find ourselves in.

Resource Description & Access (RDA)

Building on the foundations for the IFLA conceptual models, especially of the Functional Requirements for Bibliographic Records (FRBR) user tasks, national level bibliographic record mandatory elements, and the entity-relationship conceptual model, and the

Functional Requirements for Autority Data (FRAD) expansion of that model, RDA instructs us how to build well-formed metadata that can be re-used for multiple purposes. RDA is also built on the foundations of the Statement of International Cataloguing Principles (ICP), and work continues to align the instructions into closer harmony with those principles through discussions with various communities, like the music library world, to adjust the Anglo-American Cataloguing Rules, 2nd edition (AACR2) rules that were carried forward into RDA into truly FRBR/ICP principle-based instructions, rather than case law of the past. This new cataloging code is a major shift in perspective to make our bibliographic and authority data more useful in today's linked data environment and beyond. Unlike codes of the past, RDA ties the description and access to the user tasks and the specific entities that we are describing. RDA enables us to identify the resource no matter what its format – book, sound recording, score, motion picture, game, map, photograph, data file, whatever. All resources share some basic identifying information, and we can add more identifying characteristics and relationships as needed. Yet RDA has ties with the codes of the past in the objectives to collocate all the works of a creator, all the expressions of a work, all the manifestations of an expression, all the items held. RDA enables us to indicate other related resources and information, to identify the persons, families, corporate bodies, and places, so we can collocate things by or about those entities. It enables us to link to other information about those entities on the web. Such information can then be used for multiple purposes, like creating mashups, as shown with VIAF, Library Thing, and several other services; providing users with useful groupings of information and links to related things of interest. RDA moves us beyond the vocabulary of catalog cards (main entry, headings, see references etc.) to more clearly label the identifying characteristics of and relationships among entities to provide pathways for users and machines. It uses controlled vocabularies and makes them available in a linked data environment (through the Open Metadata Registry), so the Uniform Resource Identifier (URI) for each term can be used and identified with the various language versions of the term for the value being labeled.

Having URIs also can apply to standard phrases, like [publisher not identified], or [place of publication not identified]. One would hope future input systems for cataloging will suggest such standard terms through drop down menus or other devices, so catalogers would not need to key in that standard information. Such features of RDA and the Web-based vocabularies are intended to facilitate the development of future systems to make cataloging easier and to enable multilingual displays for users around the world. RDA also encourages the re-use of descriptive data from publishers, offering us the option to transcribe, that is, take data as we see it on the resource being described or from the data that accompanies the resource coming from the publisher. RDA avoids abbreviating so the user can understand what we're trying to tell them, in line with the ICP.

The new Bibliographic Framework Initiative from the Library of Congress recognizes that our past methods of communicating bibliographic and authority data through exchanging records are in need of an overhaul. With the help of interested parties and experts from around the world, we have started the process of rethinking our approach, exploring the possibilities in this Semantic Web, linked data environment. As we explore the new possibilities and build an internationally shared vision, we must always keep our users foremost in mind, as they are our customers, the reason we collect resources, the reason we organize information. We will continue to work towards collaborative approaches to reduce cataloging costs and to enrich user experiences throughout the world as they seek information in our bibliographic universe.

References

International Federation of Library Associations and Institutions. ISBD Review Group and International Federation of Library Associations and Institutions. Cataloguing Section. Standing Committee. *ISBD International Standard Bibliographic Description*. Berlin: De Gruyter Saur, 2011. (Cit. on p. 141).

International Federation of Library Associations and Institutions. Study Group on the Functional Requirements for Bibliographic Records. *Functional Requirements for Bibliographic Records*. München: Saur, 1998. http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf. (Cit. on p. 141).

Patton, Glenn E., ed. *Functional requirements for authority data: a conceptual model: final report december 2008: approved by the Standing committees of the IFLA cataloguing section and IFLA classification and indexing section*. München: Saur, 2009. (Cit. on p. 141).

Tillett, Barbara and Ana Lupe Cristán, eds. *IFLA Cataloguing Principles: the Statement of International Cataloguing Principles (ICP) and its Glossary in 20 languages*. München: Saur, 2009. (Cit. on p. 141).

BARBARA TILLETT, Library of Congress.
btil@loc.gov

Tillett, B. "RDA and the Semantic Web, Linked Data Environment". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6303. DOI: 10.4403/jlis.it-6303. Web.

ABSTRACT: Connecting to friends, colleagues, customers, and others on the internet is an everyday experience for most people these days. We use email, Twitter, Facebook, and other social networking systems quickly and easily when there is wifi or an internet service provider that reaches our geographic location – even as we move around. This change in our communication systems even extends sometimes to replacing phone calls with communications like Skype or Facetime. A former phone communication can now be a multi-media experience where you not only talk but also see each other (or groups of people), share pictures or videos or documents quickly and easily all at the same time. Where are libraries in this world?

KEYWORDS: Semantic web; RDF; Library Linked Data Project

Submitted: 2012-06-01
Accepted: 2012-08-31
Published: 2013-01-15

Linked and open data: RDA and bibliographic control

Alan Danskin

What is RDA?

Resource Description & Access (RDA) is a new content standard for describing resources and enabling access to them. This is often misunderstood, so I want to begin by clarifying what we mean by a content standard. Let's begin by thinking about an information resource; it could be any kind of resource, a book, a newspaper, a film, a song, a disk. How do we describe it? This is not as simple as it sounds. Cultural Heritage and information resources are complex and multifaceted. How do you describe a CD? Should you describe the disk, should you describe the content of the disk? How much of the content of the disk? What about the people associated with it? Are they a group or individuals? Where should we take the information from? What information is going to be useful? These are all valid questions, but are they all equally important? How do we know what is important? These are just some of the questions that RDA helps to answer. RDA will give you guidance on how to interpret the information on the resource. RDA will tell you what sources of information you should use. It will tell you how to record the title and how to gauge the significance of people or other enti-

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5463

ties associated with the resource. RDA is the latest manifestation of a cataloguing tradition that was developed in the 19th century. From a cataloguer's perspective, RDA has to do the same things that the Anglo-American Cataloguing Rules, 2nd edition (AACR2) have done for the last forty years, but the rapidly changing technological environment in which RDA has developed creates additional demands.

RDA is different from its predecessors, including Anglo-American Cataloguing Rules, which originated in an era when print was the dominant means of communication. RDA is one of the first general cataloguing codes to be created since digital began to supplant print as the dominant technology for communication. The negative reactions that greeted the drafts of RDA and which continue to be repeated can be attributed in part to the fact that RDA instructions and guidelines look very similar to AACR2. Many prospective users are disappointed because they feel that such a big technological change demands an equally big response in the instructions; and for some it calls into question the need for RDA at all. If there is relatively little change in RDA, what is the justification for adopting a new standard?

Justification and criticism

The original justification for RDA came from the conference on the future of AACR2, held in Toronto in 1997. The conference identified structural issues with AACR2 which were too deeply rooted to be corrected through the normal revision process. These included the confusion of content with carrier; arrangement by class of materials and the cultural bias of many instructions. Initial attempts to address these problems through a new edition of Anglo-American Cataloguing Rules (AACR) did not go far enough and RDA devel-

oped from the realisation that a new standard was need to address these issues.
Responses to the initial drafts of RDA were mixed. There were many who felt that AACR2 wasn't broken and didn't need to be fixed, but there were also many respondents who argued that RDA did not go far enough. Frequent criticisms were:

- RDA should be an open standard;
- RDA should be less like a cataloguing code and more like a data dictionary;
- RDA should be more explicitly structured around Functional Requirements for Bibliographic Records (FRBR).

Joint Steering Committee (JSC) took these concerns seriously and has addressed them. These issues have a bearing on RDA's compatibility with linked data.

Open standard

RDA is an open standard in the sense the any interested party can contribute to its development. Editorial control over the standard resides with the Joint Steering Committee for Development of RDA. JSC is made up of representatives of those communities already committed to RDA: the Australian Committee on Cataloguing (ACOC); the American Library Association (ALA); British Library (BL);Canadian Cataloguing Committee (CCC), Chartered Institute of Library and Information Professionals (CILIP) and Library of Congress (LC). In January 2012, the Deutsche Nationalbliothek (DNB) became the most recent member. But you don't have to be a member of JSC or one of its constituencies to contribute to RDA development. Any organization or individual can suggest changes directly to the Chair of JSC. In Europe, EURIG (European RDA Interest Group) offers an informal forum for discussion and collaboration.

RDA is open but not free. RDA development is paid for by charging for access to the standard. This is a common model for standards and it is fair in the sense that development is paid for by users. RDA has to be sustainable in the long term and no change to the business model can be considered until the investment has been recouped. It is recognised that not every user of RDA requires access to the instructions; therefore some content is being made freely available as Linked Open Data. The Committee of Principals agreed that the RDA element set and the RDA vocabularies should be placed in the public domain to enable their re-use. The element set and the RDA vocabularies are being published in the Open Metadata Registry (OMR).[1]

Element set and metadata

RDA was developed to be independent of any specific schema or format.[2] RDA conforms to broad principles of good metadata[3] practice. RDA is based on IFLA's FR family of models, which define the entities or objects of interest specified in RDA. The RDA element set corresponds to the attributes and relationships defined for those entities.

RDA gives definitions for each element and specifies how the content (or value representations) of the element should be recorded.

[1] Open Metadata Registry http://metadataregistry.org/, RDA namespace http://rdvocab.info/.

[2] JSC. Statement of objectives and principles for RDA. 1 July 2009. 5JSC/RDA/Objectives and Principles/Rev/3
http://www.RDA-jsc.org/docs/5rda-objectivesrev3.pdf.

[3] JSC Encoding RDA data. 31 May 2007 5JSC/Editor/3, http://www.RDA-jsc.org/docs/5editor3.pdf.

RDA also incorporates some features of an application profile by specifying whether an element is considered to be core or non-core. An element is considered to be core if it is required to support a basic user task. Some elements are considered to be core only in particular circumstances or for particular types of resource.

RDA also specifies controlled lists of values or vocabularies, which are mostly open allowing new terms to be added to meet changing needs. A few vocabularies, such as the RDA/ONIX Framework for Resource Categorization (ROF), are closed. ROF was developed in conjunction with Editeur to provide a flexible, extensive set of attributes and values to describe both the intellectual content of a resource and the characteristics of its carrier (Dunsire, "Distinguishing Content from carrier: the RDA/ONIX Framework for Resource Categorization").[4]

FRBR and Functional Requirements for Autority Data (FRAD)

RDA is an implementation of two models: Functional Requirements for Bibliographic Records and the Functional Requirements for Authority Data (IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report*). RDA does not specifically implement the Functional Requirements for Subject Authority Data (IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR), *Functional Requirements for Authority Data: A Conceptual Model*), which was developed too late for inclusion. In response to criticisms (based on early drafts) that RDA was not suf-

[4] RDA/ONIX Framework for Resource Categorization, version 1.0 (Released 1 August 2006). 5JSC/Chair/10, http://www.RDA-jsc.org/working2.html#chair-10.

ficiently "FRBRised", JSC comprehensively revised the structure to relate the sequence to the FRBR and FRAD models.[5]
Few current cataloguing systems or schema use the FRBR or FRAD models. Three possible implementation scenarios were identified for migration to RDA:[6]

- scenario 1: fully relational or object oriented database structure;
- scenario 2: Integrated Library System (ILS) structure, with separate bibliographic, authority and holdings records;
- scenario 3: flat file structure in which the bibliographic record is the vehicle for all the metadata.

To support migration from scenarios 2 and 3, RDA includes features which are rightly regarded as out of place in a modern metadata standard. For example, RDA provides detailed instructions on how to build authorized access points by stringing metadata attributes together into a unique key. For example, the authorised access point for the work *Bleak House* would be:

> Dickens, Charles, 1812-1870. Bleak House

The authorised access point for a French translation would be:

> Dickens, Charles, 1812-1870. Bleak House. French

This is the kind of data currently required by some systems. In a relational database system the string would be replaced by an identifier representing an authority record for the work; in a linked data context the different components and the relationships between them would be represented by URIs.

[5] JSC RDA Scope and Structure. 1 July 2009. 5JSC/RDA/Scope/Rev/4, http://www.RDA-jsc.org/docs/5rda-scoperev4.pdf.

[6] JSC RDA, FRBR/FRAD Implementation scenarios. 23 January 2008. 5JSC/Editor/4, http://www.rda-jsc.org/docs/5editor4.pdf.

RDA as linked data

RDA was originally conceived of at a less granular level than Resource Description Framework (RDF) properties, but that is not a barrier to its use in a linked data context. Meaningful linking between resources is inherent in RDA.

RDA Appendix J defines relationship designators to enable consistent and explicit linking between bibliographic resources. Drawing upon Tillett's taxonomy of bibliographic relationships, RDA makes it possible to express derivative, descriptive, whole-part, accompanying, equivalent or sequential relationships (Tillett); refinement of terms enables nuances of the relationship to be expressed. For example, *The bored of the rings* is a derivative work, which is an imitation of, or more explicitly, a parody of, *The lord of the rings*. Approximately 200 bibliographic relationships are specified in RDA.

RDA Appendix I defines relationship designators to connect resources to persons, families, or corporate bodies. For example Charles Dickens is the author of the work *Bleak house*; Andrew Davies is the screenwriter of the 2005 work *Bleak House*.

The relationship designators specified in appendices I and J have been registered on the Open Metadata Registry. Seventy five RDA vocabularies have also been registered in the OMR. The vocabularies constitute a rich source of metadata, with applications beyond libraries. The vocabularies range from Applied material to Video Format and encompass terms as diverse as spoken word: (Content type), quarterly (Frequency); female (Gender); serial (Mode of issuance). A total of 810 terms have been registered in vocabularies.

RDA vocabularies have great potential for reuse. The relationship designators are particularly valuable for explaining why x is related to y, but it is important to be aware that they are not yet stable. The majority of RDA terms in OMR currently have the status "New Proposed". JSC, with the assistance of Metadata Management Associates, is in the process of changing the status to "published".

At present, only a relatively small number of vocabularies have been published because JSC is confirming definitions for all of the terms. This process has inevitably raised some issues regarding duplication of terms between vocabularies and the forms of terms. JSC intends to resolve these issues before publishing the terms. Publication of the terms in the OMR signals to the community that the URIs for the concept represented by the term can be reused with confidence. The RDA elements are also registered on OMR with the status of "New-Proposed". JSC is reviewing the RDA element set based on feedback from testing and extensive discussions with Metadata Management Associates (MMA). JSC has to be confident that the element set is stable before the elements are published. Several difficult issues have had to be resolved in order to attain the required level of confidence.

Constrained or unconstrained elements

An aspect of the linked data vision is that metadata can break down barriers, including those silos erected within the cultural heritage sector to meet the specific needs of museums, archives and libraries. Placing constraints on linked metadata elements is a barrier to reuse. For example, RDA Publisher's Name is an RDF property with domain manifestation. This is consistent with the FRBR model but it makes the element unattractive to users or communities who do not perceive a need to distinguish between Work, Expression Manifestation and Item. It has taken some time for JSC to understand these perspectives and from JSC's perspective an element set without FRBR cannot be RDA. It was therefore agreed that an unconstrained (or unbound) element set should be created in addition to the constrained elements. As illustrated below, the constrained RDA elements have been modeled as sub-properties of the unconstrained elements. In this example the constrained element is currently distinguished by a parenthetical qualifier in the label. Publisher's name (Manifestation) has domain Manifestation, as is reflected in the URI. The unconstrained element Publisher's Name is the "parent" of the constrained element, "Publisher's name (Manifestation), but its own domain is unbounded. The registration of constrained and unconstrained elements in the same namespace is likely to be confusing to prospective users and will also complicate dissemination of information about the element sets. Different options for resolving these problems are being discussed by the stakeholders and need to be resolved before the elements can be published.

The RDA namespace also includes FRBR entities for RDA element set. This element set was registered because FRBR and FRAD entities were required by RDA but had not been registered by IFLA.

Interoperability and mapping

In addition to FRBR and RDA, the International Standard Bibliographic Description element set was published on the OMR in 2011; basic MARC 21 elements have also been published and the Dublin Core Element set has been available since 2008. The increasing availability of element sets and vocabularies in RDF creates new possibilities for interoperability and mapping. At the JSC meeting in Glasgow, in 2011, Gordon Dunsire reported on work to map between the RDA and International Standard Bibliographic Description (ISBD) element sets and vocabularies (Dunsire, "Mapping ISBD and RDA element sets: briefing/discussion paper"; "Mapping ISBD Area 0 vocabularies to RDA carrier and content"). Two different approaches were followed.

The ISBD Area 0 vocabularies and the RDA vocabularies for content type, media type and carrier type were mapped in a hub and spoke model to the RDA/ONIX Framework for Resource Categorization.

The ISBD and RDA element sets were mapped using the element definitions and scope notes and the more explicit semantics in property and class declarations. The exercise implied that RDA and ISBD elements are sub-properties of other properties, which have neither ISBD nor RDA as domains or ranges. To put it another way, unconstrained or unbounded elements could support mapping between the constrained ISBD and RDA element sets.

At the DCMI UK Regional Meeting hosted by the British Library in April, 2012 Dunsire explored these ideas further, using MARC 21 and Dublin Core. A significant implication of this work is that it is possible to build mappings between element sets which have different levels of specificity. Therefore it is possible to link MARC 21 ″Target audience“; Dublin Core Terms ″audience“; FRBR ″has intended audience“; and RDA ″intended audience“ by means of an unconstrained property ″intended audience“, which can itself be link to ISBD ″has note on use or audience“ through a further unconstrained property, ″has note on use or audience. Rich RDA metadata linked as sub-properties of less granular elements can be dumbed-up into simple Dublin Core for applications that don't want RDA ("Turtle Dreaming").

Similar approaches could be followed for other metadata schema with wider application than libraries. There is much common ground between RDA and Friend of a Friend (FOAF); both define properties of the person. The registration of RDA properties in RDF will enable a more rigorous comparison of their semantics and how and whether they relate will become clearer. For example, RDA does not refine the components of a personal name, but FOAF does have properties for given name and family name.

Internationalization

One of the objectives of RDA was to internationalize the instructions. Much work has been done to remove AACR2's cultural bias and RDA is already being translated into several languages, including German, French and Spanish. In the OMR URIs representing elements or controlled terms can have labels and definitions specific to each language community. This holds out exciting possibilities in future for metadata which can be presented in the language appropriate to a specific audience or even to a specific user.

Bibliographic framework

Realising the benefits of RDA has always been dependent on the development of schemata and systems to unlock its potential. Linked data concepts, developing more or less concurrently with RDA, offer exciting possibilities to make library metadata open and actionable. But linked data also has to demonstrate its ability to meet the demanding requirements of resource discovery and data management. Nor is RDA only about the metadata we create tomorrow, unlocking legacy metadata, which represent humanity's literary and intellectual heritage is at least important.

Library of Congress announced the Bibliographic Framework Transition Initiative in early 2011. The initiative is intended to map a path from the current bibliographic framework, built upon the exchange of MARC records, towards a new framework in which library metadata can be shared and reused without being transported and replicated. In October, we learned that the framework would be looking towards solutions based on RDF and linked data. In May, LC announced the appointment of Zepheira to lead a modelling initiative to translate MARC 21 into a linked data model.[7]

At last, the pieces are coming together which will enable the links to be forged.

References

Dunsire, Gordon. "Distinguishing Content from carrier: the RDA/ONIX Framework for Resource Categorization". *D-Lib Magazine* 13.1/2. (January/February 2007). <http://www.dlib.org/dlib/january07/dunsire/01dunsire.html>. (Cit. on p. 151).

—. "Mapping ISBD and RDA element sets: briefing/discussion paper". (2011). <http://www.rda-jsc.org/docs/6JSC-Chair-4.pdf>. (Cit. on p. 155).

—. "Mapping ISBD Area 0 vocabularies to RDA carrier and content". (2011). <http://www.rda-jsc.org/docs/6JSC-Chair-5.pdf>. (Cit. on p. 155).

—. "Turtle Dreaming". (2012). <http://dcevents.dublincore.org/index.php/BibData/fyo>. Presented to the seminar "Five years on" British Library, London 27th April, 2012. (Cit. on p. 156).

IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional Requirements for Bibliographic Records: Final Report*. München: K.G. Saur, 1998. (Cit. on p. 151).

—. *Requisiti funzionali per record bibliografici. Rapporto conclusivo*. Roma: ICCU, 2000. Trans. of *Functional Requirements for Bibliographic Records: Final Report*. Munich: K.G. Saur,

IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR). *Functional Requirements for Authority Data: A Conceptual Model*. Ed. Glenn E. Patton. München: K.G. Saur, 2009. (Cit. on p. 151).

—. *Requisiti funzionali per i dati di autorità. Un modello concettuale*. Roma: ICCU, 2010. Trans. of *Functional Requirements for Authority Data: A Conceptual Model*.

Tillett, Barbara B. "Bibliographic Relationships". *Relationships in the Organization of Knowledge*. Ed. Carol A. Bean and Rebecca Green. Dordrecht: Kluwer Academic Publishers, 2001. 19–35. (Cit. on p. 153).

[7] Bibliographic Framework Transition Initiative
http://www.loc.gov/marc/transition/.

ALAN DANSKIN, British Library.
alan.danskin@bl.uk

Danskin, A. ”Linked and open data: RDA and bibliographic control”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5463. DOI: 10.4403/jlis.it-5463. Web.

ABSTRACT: Resource Description & Access (RDA) is a new cataloguing standard which will replace the Anglo-American Cataloguing Rules, 2nd edition, which has been widely used in libraries since 1981. RDA, like AACR2, is a content standard providing guidance and instruction on how to identify and record attributes or properties of resources which are significant for discovery. RDA is also an implementation of the FRBR and FRAD models. The RDA element set and vocabularies are being published on the Open Metadata Registry as linked open data. RDA provides a rich vocabulary for the description of resources and for expressing relationships between them. This paper describes what RDA offers and considers the challenges and potential of linked open data in the broader framework of bibliographic control.

KEYWORDS: Library linked data; RDA; Bibliographic control

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

Library of Congress Classification as linked data

Kevin Ford

What is linked data

The Library of Congress has published a select number of classes from the Library of Congress Classification (LCC) system as linked data as a new offering of its Linked Data Service,[1] commonly known as id.loc.gov. The offering, while still considered a beta project, provides URIs for resources that represent a simplified version of the underlying data found in the source MARC Classification records. The beta service also furnishes URIs for classification number resources that either derive directly from the underlying data or are the result of a synthesis between a schedule resource and a table resource. Although the data are presented in MADS/RDF[2] and SKOS[3] where appropriate, LCC as linked data is accompanied by a small LCC ontology to more accurately describe the types of classification resources and the relationships between them, especially where MADS/RDF and SKOS Class and Property definitions were seen as insufficient. This paper explores the publication of LCC as linked data and the accompanying ontology by contextualizing them with

[1] http://id.loc.gov.
[2] http://www.loc.gov/mads/rdf.
[3] http://www.w3.org/2004/02/skos.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5465

respect to prior efforts representing LCC as linked data, representing Dewey as linked data, and the appropriateness of SKOS for library classification data, especially given the historical need for a distinct MARC format for Classification.

The Library of Congress classification system has existed since the late nineteenth century "to organize and arrange the book collections of the Library of Congress" (*Library of Congress Classification*). The system is organized into twenty-one classes, most of which are further divided into subclasses. Each class represents a field of knowledge, such as Art, Law, or History. Each subclass is further divided into more specific topics that basically adhere to a hierarchical representation of the field of knowledge. Like most classification systems, LCC is subject-based. The resulting "number", therefore, represents a distinct topic within the field of knowledge. For decades LCC has been printed, bound, and distributed (at cost, basically) and still is today. One may acquire, for a price, the entire 41-volume set or one may choose individual classes or schedules. LCC is also accessible via ClassificationWeb,[4] which is a sophisticated web application designed to assist catalogers with the assignment and creation of LCC classification numbers. It is offered as a subscription service for which LC charges a fee. Also for cost (basically), the Library of Congress Classification is available in MARC21 format and is made available as a bulk download, with periodic updates, from the Library's Cataloging and Distribution Service. Notably, the raw data, though available, requires purchase and is not presented in accordance with linked data methods and principles.

The Library of Congress Classification as linked data does have a history, albeit a short and little known one. Karen Coyle laboriously scraped the first four levels (more or less) of all LC Classification classes from PDF documents hosted on the LC website to a plain

[4] https://classificationweb.net.

text file (that is, something far more accessible for machines) and uploaded the resulting text file to archive.org.[5] This work dates to, and therefore the data predates, September 2007.[6] The PDF documents, which are still available (though perhaps updated since), present a detailed outline of LCC. Ed Summers then took the text file, generated a basic SKOS RDF representation from it, and developed a very simple website where he published the SKOS data.[7] This work was little publicized, but it is still active and accessible. Summers's code is on GitHub.[8]

Coyle's text file simply lists the classes (A, B, C, and so on) and the first three levels, if appropriate, of each subclass (AC, AE, AG, and so on). The concept's label at any given level is matched with the class number. Because only the first few levels of LCC are outlined, most classification numbers represent a range of more specific topics. Missing – nearly universally – from the detailed outline are language-specific divisions within topics, temporal divisions within topics, and form divisions within topics, in addition to simply greater granularity and specificity, such as the distinction between "General works" and "Special topics." From Coyle's text file, Summers generated a skos:Concept Resource for each classification number and associated label. He took each classification number and appended it to a base HTTP URI (in a namespace he controls) to create an unique identifier for the resource and he made the lexical label for the topic (and class number) the skos:prefLabel. He generated skos:broader and skos:narrower relationships between classification topics when the classification number represented an encompassing range or a more specific range respectively. Summers created something akin

[5] http://ia600304.us.archive.org/0/items/LcClassificationA-z/lc_class.txt.

[6] http://ia600304.us.archive.org/0/items/LcClassificationA-z/LcClassificationA-z_meta.xml.

[7] http://inkdroid.org/lcco.

[8] https://github.com/edsu/lcco.

to an LCSH-like pre-coordinated heading with the labels of narrower topics (i.e. those that fit contextually with broader topics): the skos:prefLabel of narrower topics contains the labels of its broader relations, the labels of which are separated by two hyphens. The data collected by Coyle, which may have been all that was reasonably possible to collect, were limited to a class number, label, and hierarchy. The first three levels of the Dewey Decimal Classification system – the Dewey Summaries – have been available as linked data since 2009.[9] OCLC published the full Dewey Decimal Classification as linked data in Summer 2012. As with Summers's design, each topic is a skos:Concept with broader or narrower relations to any given topic's hierarchical relatives. Published as it was by OCLC, the available data are richer, including information about provenance and licensing (no fewer than four statements for each Concept), creation and modification times, among a few others. Unlike Summers's design, OCLC reserved the skos:prefLabel exclusively for the lexical label of the given Concept – broader relations are not strung together with the topic's label to create the skos:prefLabel. OCLC's URI design patterns warrant special mention. Pains have been taken to embed some semantics into the URI pattern, reserving, essentially, one namespace each for "non-information resources (abstract or concrete real-world objects), generic resources, and their representations" (OCLC). Although some of the URI examples do not appear to function presently, the focus on URI composition and the need to represent a variety of different resource types bears on the representation of all aspects of publishing classification systems such as DDC and LCC as linked data.[10] A diverse number of resource types are also very relevant to LCC. In addition to the embedded semantics in the Dewey URIs, this issue received greater elucidation

[9] http://dewey.info.

[10] The actual service at http://dewey.info features diverse URI patterns, all of which appear to function, for all types of information resources.

by Panzer and Zeng in two related publications (Panzer and Zeng; Zeng, Panzer, and Salaba).
The authors explored how to model classification schemes (notably DDC) in SKOS. Among other findings, the authors discuss how classification systems include "assignable" and "non-assignable" concepts. In DDC, an example of a non-assignable concept is a centered entry, or a classification number range or span for which there are likely a number of more specific topics and, therefore, specific numbers. In LCC, this is referred to as a range. There is also the issue, as Panzer and Zeng note (2009), of synthesized concepts (a classification number and topic that are a result of combining two concepts in the classification system) and non-synthesized concepts. One risks some semantic incoherency when attempting to model all these types of things, and to establish appropriate relationships between them, purely in SKOS. Panzer and Zeng considered the need to create, minimally, an extension to the core SKOS vocabulary, but it was clear that an altogether separate attempt might be necessary, in a namespace entirely distinct from a SKOS one, to correctly capture the semantics and relationships. These same issues also materialized during the process of trying to represent LCC in SKOS.

SKOS – the Simple Knowledge Organization System – is designed "to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading lists and taxonomies within the framework of the Semantic Web".[11] SKOS has proven to be extremely versatile and effective at representing thesauri, subject heading lists, and taxonomies (though, in part as a result of being intentionally simple, there can be some loss of granularity with respect to library data). In fact, data represented using the MARC Format for Authority data, such as subject heading lists like LCSH, map effortlessly to SKOS. This is seen readily and simply

[11] http://www.w3.org/2004/02/skos.

when decomposing a MARC Authority record into MADS/RDF and SKOS. For MARC Authority, a valid (i.e. not deprecated) authority record is the Concept. The 1XX - the main heading - becomes the authoritative or preferred label. MADS/RDF provides a means to capture the type of concept, be it a Topic, Geographic, GenreForm, or Temporal notion, and a few others. MADS/RDF also provides support for better representation of pre-coordinated headings. MARC Authority 4XX fields are variant or alternate labels. 5XX fields represent various relationships between terms, of which broader and narrower relationships are the most popular. MADS/RDF added a few additional relationships, such as those needed to accurately record connections between earlier or later established concepts, and a new resource type to clearly denote deprecated resources. A number of note fields defined in MARC Authority also have one-to-one mappings to MADS/RDF and SKOS. But MADS/RDF and SKOS classes and properties have been far less amenable to classification data, or at least to library-specific classification systems such as DDC and LCC.[12] This is essentially the difficulty Panzer and Zeng encountered during their research and it is the same encountered when attempting to publish LCC as linked data. At least when it comes to library classification systems such as DDC and LCC, this is unsurprising.

The influential consideration here lies with the MARC21 format for Classification.[13] More specifically, its very existence. Formally but provisionally published in June 1990, the MARC21 Format for

[12] This probably has to do a lot to do with the relative complexity of classification systems, especially with respect to how classification numbers are constructed, when compared to thesauri or "subject heading lists;" the aggregate expertise of the SKOS designers and members of the working group with respect to classification systems; and, partly as a natural extension of the previous point, a certain amount of partiality and attention given to, and in favor of, thesauri and "subject heading lists" during the development of SKOS.

[13] http://www.loc.gov/marc/classification.

Classification was specifically developed to facilitate the exchange and printing of classification data, most notably LCC and DDC (Guenther). Importantly, the new MARC format was, however, the result of an attempt to modify the MARC format for Authority data (this work started in 1987/1988). After identifying most of the changes that would be required of the MARC Authority format, a draft of the proposed changes was presented to the committee overseeing changes to the MARC formats (MARBI). Following this review, and the early development period generally, it was clear that "there was less overlap with the authority format than originally anticipated, and ... [the MARC Authority] codes and conventions were too constraining" (Guenther). The proposal for classification data was rewritten to be a separate format, which would become the MARC Format for Classification by 1991.

The MARC Format for Classification – and its development process – took into consideration the very same semantic difficulties encountered by Panzer and Zeng, and the present author, when faced with "skosifying" complex library classification data, and a difficulty that is compounded by the unsuitable nature of the RDF data element semantics. The MARC Format for Classification can represent class schedules and tables, neither of which is necessarily assignable as is. The format can represent ranges and hierarchy. Naturally, it has full support for notes and index terms. But SKOS semantics are not rich enough this type of information. That said, SKOS can reasonably represent (assignable) classification topics and even class number ranges. It is with this information in mind, and the background work by Panzer and Zeng, that it was decided to present LCC as linked data as much as possible in MADS/RDF and SKOS but to define a small vocabulary in OWL to faithfully represent LCC-specific data and data elements where MADS/RDF and SKOS fall short.[14]

[14] http://id.loc.gov/ontologies/lcc.

Although there are a few ontological constraints on the data, constraints do not presently extend to how the data are used. For example, while it could be possible to infer "assignable" versus "non-assignable" resources from the intersection of select Classes in the ontology, this type of modeling has not been undertaken. As such, it is an experimental offering that attempts to make no semantic restrictions on its use but which strives to represent the derived and underlying data accurately. The ontology is also specific to LCC; it makes no attempt to model data elements specific to other classification systems, such as DDC. Also, though it would be unwise to rule out OCLC developing an ontology for DDC, the explicit declaration of classes in the small LCC ontology transfers the semantics embedded in dewey.info URIs to the data itself. ("Smart" URIs and clear data semantics are not mutually exclusive and could, in fact, be complementary.) A select number of Library of Congress Classification classes are available from LC's linked data Service,[15] commonly known as id.loc.gov.[16] This offering - at the time of this publication - is very much a beta offering. During this stage, the data and its representation are subject to change, especially as more is learned about how the data is used and better ways for it to be represented are determined or developed. Nevertheless, it is an attempt not only to publish an RDF representation of the underlying data used to construct classification numbers but also to publish the classification numbers themselves. To this end, an effort has been made to apply the tables to schedules, thereby synthesizing a classification number, as appropriate.

In order not to become too mired in MADS/RDF [17] and SKOS [18] semantics and restrictions, everything is a MADS/RDF Authority

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

[16] http://id.loc.gov/ontologies/lcc.html.

[17] http://www.loc.gov/mads/rdf.

[18] http://www.w3.org/2004/02/skos.

and SKOS Concept, with the exception of Index Terms, which can be interpreted as variants. They are therefore instantiated as MADS/RDF Variants and SKOS/XL Alternate Labels. The authoritative label - the preferred label and the tightly controlled term - is reserved for the main caption or term. This is therefore similar to how OCLC created Dewey resources and a departure from how Summers presented the data. The full lexically represented hierarchy that one finds in the source MARC records is recorded simply as an rdfs:label so that it is still available for parsing and potentially for display purposes. The classes and properties in the LCC ontology, therefore, are the real carriers of distinction between Library of Congress Classification resources published at id.loc.gov.[19] The LCC ontology provides a way to describe the "underlying data," which is a reference to the data one would find in a MARC classification record. Data in the MARC classification record include information about classification-specific resource types such as tables and schedules, and data describe details about how to apply table numbers to base numbers to generate and assignable classification number. As such, the LCC ontology defines Classes and Properties sufficient enough to accurately represent LCC data in RDF and sufficient enough to synthesize class numbers from schedules when and however appropriate. The ontology is a significant simplification of the MARC Classification codes, data element definitions, and conventions. One such simplification touches on the identification of different types of ranges defined in MARC Classification. Because there appears to be no meaningful distinction between a MARC Summary Range and MARC Defined Range with respect to their representation in RDF, specifically for LCC, these types are simply an LCC Range. On the

[19] I have endeavored to capitalize the word "Class" (and Property) when referring to an OWL or RDF Class (or Property). Whenever referencing an entity associated directly with LCC - such as classification number, LCC class, class schedule, or class number - I have presented the word in all lowercase letters.

other hand, it was deemed necessary to define an additional Table type - a Guide Table - where the MARC Classification format made no clear distinction between the two. A Guide Table is hierarchically the broadest table concept and carries the Table Rule, which is the instruction needed to synthesize a classification number between an LCC Schedule and an LCC Table. The small LCC ontology includes Classes for a Schedule, Range, Table, Guide Table, and Table Rule, all of which are types of resources that are somewhat unique to classification schemes. Additionally, classification-specific properties have been defined that relate these classes to each other, such as one that relates a Table to its Guide Table or another that relates a Guide Table to one or more Schedules, to which the Guide Table may apply. At all other times, MADS/RDF, which is fully mapped to SKOS, is employed (all data are, of course, also outputted as SKOS). Naturally, these Table, Guide Table, and Schedule resources are "underlying data" and are generally considered to be "non-assignable," that is they are resources that should not be used to describe another resource, such as a bibliographic one. Because these resources often have a one-to-one relationship with an underlying MARC Classification record, the LCCN of the underlying record has been used as part of the URI scheme. An LCCN that begins with "CF" represents a schedule; one that begins "CT" represents a Guide Table or Table. However, when classification resources are described with the ClassNumber OWL Class, the resource could be described as assignable. The URIs for these resources end in a classification number or range.

A ClassNumber resource may be an LCC Range or a MADS/RDF Topic. The former - an LCC Range - generally represents a group of concepts hierarchically related to the broader concept represented by the range. Of course, ranges are not assignable when traditionally assigning classification numbers to physical bibliographic resources. MADS/RDF Topic was used when the resource represented a single,

`lcc:GuideTable/lcc:Table`	http://id.loc.gov/authorities/classification/ct96152584
`lcc:Schedule`	http://id.loc.gov/authorities/classification/cf94051344
`lcc:ClassNumber`	http://id.loc.gov/authorities/classification/ND1360-ND1360.6

Table 1: Table showing example URIs based on different LCC types. Note how LCCN is last token of URI in the first two examples versus the classification number range in the last example.

distinct concept.[20]

MADS/RDF and SKOS broader and narrower relationships were asserted between all concepts whether they represented non-assignable underlying data or assignable classification numbers and ranges. However, broader and narrower relationships are expressed between concepts based on whether they represent underlying data (schedules, tables, and guide tables) or classification numbers. Schedules link to tables, guide tables, or other schedules for example; classification numbers link to other classification numbers. For example, an LCC Schedule or LCC Table, both of which are considered non-assignable resources and represent underlying data, may record broader or narrower relationships to other LCC Schedules or LCC Tables respectively, but will not carry such a relationship to an LCC Class Number. That said, there are defined relationships in the LCC ontology created expressly to accurately capture the relationship between underlying data resources, such as an LCC Table, and an LCC Class Number. For example, lcc:isynthesizedFromTable and lcc:synthesizedFromSchedule records from which LCC Schedule or LCC Table the LCC Class Number derives.

The LCC ontology has helped considerably in maintaining a separation of concerns and avoiding the pitfalls of representing this information purely, or at least mainly, in SKOS. Additionally,

[20] http://id.loc.gov/authorities/classification/B4877.S4.html.

because the non-assignable or underlying data has also received representation in RDF, it is possible for others to experiment with this information. In fact, it is known beyond any doubt that the representation of LCC Table resources as tables, LCC Guide Tables as guide tables, LCC Schedules as schedules, and the inclusion of Table Rules in RDF is sufficient to derive and synthesize classification numbers from these resources. The creation of classification numbers and resources, as seen at id.loc.gov, is the result of programming manipulation of LCC Schedule and LCC Table numbers (when tables were required and as part of the process of applying the table rules) and smart querying of the LCC Table data in RDF loaded into a triplestore. Ultimately, focus to date has been almost entirely on the accurate generation of classification numbers from LCC Schedules and, when required, LCC Tables. The MARC Classification records contain numerous ways to link one classification schedule or range to another, often in a separate class altogether. No attempt has been made to extract this information and establish the relationship between the two concepts in the data. Where MADS/RDF or SKOS relationship properties are insufficient, it is anticipated that new LCC properties will be created. Additionally, the data, as presently available, represents a snapshot of any given class - no updates or changes to those classes have been taken into consideration. The types of changes classification numbers undergo, how those changes are recorded in the data, and how changes may or may not affect the RDF representation of LCC remain open, and as yet unexplored, issues. Given that the use of MADS/RDF provides a means to indicate the type of concept - here everything is a MADS/RDF Topic - one wonders whether it would be possible to identify the type of concept especially at the narrower hierarchical levels where the concept might be distinctly temporal in nature (18th century) or a form (General works or Cantos) or a specific language (Russian).

Although a little time has been spent linking LCC resources with LCSH and LC Names resources, more work can be done here too. Naturally, linking LCC to Dewey would be a high priority endeavor. There has also been a long-standing desire to use the Library of Congress Classification as an entry point to the bibliographic catalog versus merely a means to locate a book on a shelf (Chan).

Considerable work remains, but it is hoped that this beta offering will energize developers and stimulate additional innovation. In particular, we look forward to learning of new use cases, especially ones that will explore new uses of the data. For our part, we will continue to make entire classes available as time and resources permit. And we will continue to augment the data and accompanying ontology to ensure that the data being offered is as rich as possible and necessary to accurately represent the data and promote new development.

References

Beall, Julianne and Joan S. Mitchell. “History of the Representation of the DDC in the MARC Classification Format”. *Cataloguing & Classification* 48.1. (2010): 48–63.

Chan, Lois Mai. “Library of Congress classification as an online retrieval tool. Potentials and limitations”. *Information Technology and Libraries* 5.3. (1986): 181–192. (Cit. on p. 173).

Guenther, Rebecca S. “The Development and Implementation of the USMARC Format for Classification Data”. *Information Technology and Libraries* 11.2. (1992): 120–131. (Cit. on p. 167).

Library of Congress. *Library of Congress Classification*. http://www.loc.gov/catdir/cpso/lcc.html. (Cit. on p. 162).

OCLC. *Using the API*. 2010. http://oclc.org/developer/documentation/dewey-web-services/using-api. (Cit. on p. 164).

Panzer, Michael and Marcia Lei Zeng. “Modeling Classification Systems in SKOS: Some Challenges and Best-Practice Recommendations”. *Semantic interoperability of linked data: Proceedings of the International Conference on Dublin Core and Metadata Applications, Seoul, South Korea*. Ed. S. Oh, S. Sugimoto, and S.A. Sutton. 2009. Seoul: Dublin Core Metadata Initiative and National Library of Korea, 2009. 3-14, http://dcpapers.dublincore.org/index.php/pubs/article/view/974/944. (Cit. on p. 165).

Zeng, Marcia Lei, Michael Panzer, and Athena Salaba. “Expressing Classification Schemes with OWL 2 Web Ontology Language”. *Paradigms and conceptual systems in Knowledge Organisation: Proceedings of the Eleventh International ISKO Conference, University of Rome, Italy*. Ed. C. Gnoli and F. Mazzocchi. 2010. 356–362. (Cit. on p. 165).

Kevin Ford, Library of Congress.
kefo@loc.gov

Ford, K. ”Library of Congress Classification as linked data”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5465. DOI: 10.4403/jlis.it-5465. Web.

Abstract: In 2009 and in 2011, the Library of Congress made two of its largest authority files –Subject Headings and Names - available as linked data via LC’s linked data service, id.loc.gov. Both are offered in MADS/RDF and SKOS. It is LC’s objective, in 2012, to publish another of its largest authority files as linked data: LC Classification. However, whereas the source records for Subject Headings and Names are encoded in the MARC Authority format, from which there is a relatively straightforward mapping to MADS/RDF and SKOS, LC Classification records rely on the MARC Classification format. Mapping from LC Classification to MADS/RDF or SKOS has been a little more challenging. For example, records that represent classification ranges, which are not Concepts intended to be assigned, are not easily accommodated in SKOS. This presents additional problems when needing to accurately represent the relationships in RDF for LC Classification. With comparison to the publication of LCSH and Names at id.loc.gov, this paper will examine issues encountered – and how those challenges were addressed – during the conversion of LC Classification to MADS/RDF and SKOS for release as linked data at id.loc.gov.

Keywords: Library linked data; Library of Congress Classification; Ontology; SKOS; MARC21; Authority control

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

Dewey linked data: making connections with old friends and new acquaintances

Joan S. Mitchell, Michael Panzer

We adress the history, uses cases, and future plans associated with the availability of the Dewey Decimal Classification (DDC) system as linked data. Parts of the DDC have been available as linked data since 2009. Our initial offering included the DDC Summaries (the top three levels of the DDC) in eleven languages exposed as linked data in dewey.info, an experimental web service. In 2010, we extended the content of dewey.info[1] by adding assignable numbers and captions from the Abridged Edition 14 data files in English, Italian, and Vietnamese. In mid-2012, we -extended the content of dewey.info yet once again by adding assignable numbers and captions from the schedules and geographic table in the latest full edition database, DDC 23. We will discuss the behind-the-scenes development and data transformation efforts that have supported these offerings, and then turn our attention to some uses of Dewey linked data plus future plans for Dewey linked data services.

[1] http//dewey.info.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5467

History

The history of Dewey linked data is an evolving story of opportunity and experimentation, with an eye toward usability and use of the data. In 2009, the DDC 22 Summaries, an authorized derivative work based on the top three levels of DDC 22, had already been translated into ten languages (more languages than the full edition of the DDC on which the data were based). We decided to experiment with making the DDC Summaries available as linked data in an experimental web service, dewey.info. Our initial design goals included:

- provide an actionable URI for every class;
- encode the classification semantics in RDF/SKOS;
- provide representations for machines and for humans;
- make the data usable under a widely understood license used in the Semantic Web community.

Publishing Dewey as linked data required development decisions on several different fronts. First of all, we had to develop a URI pattern that would support the identification of several different kinds of entities and relationships. The URIs had to act as dereferenceable identifiers that could deliver representations of the referenced resources in a RESTful manner. Each class had to be identified with a URI and the data had to be presented in a reusable way. In developing the URI pattern, we had to provide for the full complexity of the DDC at any time: identification of the scheme, parts of the scheme, edition, language, and time slice. Figure 1 shows the status of DDC 22 at the time of initial development of URIs for the DDC.

DDC 22 was initially published in 2003; the various DDC 22 translations were published in 2005 (German), 2007 (French), 2009 (Italian),

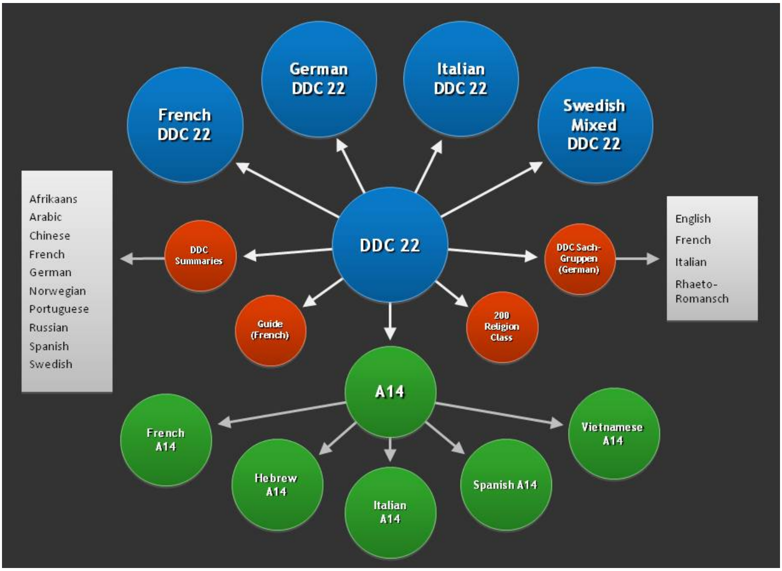


Figure 1: Versions of the DDC based on DDC 22.

and 2011 (Swedish-English mixed version). Abridged Edition 14 (a logical abridgment of DDC 22) was published in 2004; translations followed in 2005 (French), 2006 (Italian and Vietnamese), and 2008 (Hebrew and Spanish). The DDC Summaries based on DDC 22 were published in English and ten other languages at the time of the introduction of dewey.info. Besides the DDC Summaries, figure 1 includes two other authorized derivative works based on DDC 22: 200 Religion Class (2004), an updated subset of DDC 22; Guide de la classification décimale de Dewey, a French-language customized abridgment of DDC 22, and DDC Sachgruppen, a German translation of selected DDC 22 top-level classes (including some below the three-digit level) developed for the primary use case of organizing the national bibliographies of Germany, Austria, and Switzerland (the four languages in the box on the right-hand side of figure 1 are translations of DDC Sachgruppen; all five language versions are used in the national bibliography of Switzerland).

Dewey.info includes representations for machines and humans; the latter is particularly important in order to illustrate the DDC data offerings to a wider community beyond traditional users of value vocabularies from the library community. The data in dewey.info are presented in human (XHTML+RDFa) and machine (RDF) versions (the machine version of dewey.info has three different RDF serializations: RDF/XML, Turtle, and JSON). The Dewey URIs have the following general pattern: http://dewey.info/{object-collection}/{object}/{snapshot-collection}/{snapshot}/about}. Specific documents have a variable resource name component and allow specification of content language and type (format):
http://dewey.info/{object-collection}/{object}/{snapshot-collection}/{snapshot}/{resource-name}.{language}.{content-type}.

An object is a member of the DDC domain and part of an object collection. The object collection specifies the type of the object. The object collection is a mandatory component and can have one of the values "scheme," "table," "class," "manual," "index," "summary,"

and "id." A specific object from that collection follows if required. For example:
http://dewey.info/class/576.83/
http://dewey.info/scheme/
http://dewey.info/table/2/

A snapshot is used to refer to versions of objects at specific points in time. Snapshots can be part of a snapshot collection, e.g., "e22," referring to every concept version that is part of Edition 22 of the DDC. In the following examples, the first URI is an example of a snapshot, the second is an example of a snapshot collection, and the third is an example of a snapshot-collection/snapshot/ combination. snapshot-collection/snapshot/.
http://dewey.info/class/641/2009/
http://dewey.info/class/641/e22/
http://dewey.info/class/641/e23/2012-08/

Language and format are also accommodated in the URI:
http://dewey.info/class/641/about.it
http://dewey.info/class/641/about.rdf
http://dewey.info/class/641/about.it.html

While SKOS is often the RDF vocabulary of choice for representing controlled vocabularies on the Web, its initial development was largely informed based on thesaurus-like knowledge structures. Panzer ("DDC, SKOS, and linked data on the web") and Panzer and Zeng ("Modeling Classification Systems in SKOS: Some Challenges and Best-practice Recommendations") have noted some of the challenges in representing classification data in SKOS. Since the initial DDC linked data offering did not include complicated note types and relationships between classes other than those expressed by the notational hierarchy, the shortcomings in SKOS noted elsewhere with respect to the representation of classification data did not pose a major roadblock in the exposure of the DDC 22 Summaries in dewey.info.

The query http://dewey.info/class/641/about.it.rdf delivers the following machine-actionable representation in RDF/SKOS, which focuses on presenting concept metadata together with number and caption information plus basic semantic relationships.

Note that the two main entities retrieved are http://dewey.info/\class/641/ and http://dewey.info/class/641/2007/02/about.it, connected through a dct:hasVersion relationship:

Listing 1: Example of concept metadata representation in RDF/SKOS.

```
<rdf:RDF xmlns:rdf=’’http://www.w3.org/1999/02/22-rdf-syntax-
    ns#’’ xmlns:xhv=’’http://www.w3.org/1999/xhtml/vocab#’’
    xmlns:cc=’’http://creativecommons.org/ns#’’ xmlns:dct=’’
    http://purl.org/dc/terms/’’ xmlns:skos=’’http://www.w3.
    org/2004/02/skos/core#’’>
<rdf:Description rdf:about=’’http://dewey.info/class
    /641/2007/02/about.it’’>
<rdf:type rdf:resource=’’http://www.w3.org/2004/02/skos/core#
    Concept’’/>
<xhv:license rdf:resource=’’http://creativecommons.org/
    licenses/by-nc-nd/3.0/’’/>
<cc:attributionName>
  OCLC Online ComputerLibrary Center, Inc.
</cc:attributionName>
<cc:attributionURL rdf:resource=’’http://www.oclc.org/dewey
    /’’/>
<dct:isVersionOf rdf:resource=’’http://dewey.info/class
    /641/’’/>
<dct:language rdf:datatype=’’http://purl.org/dc/terms/RFC4646
    ’’>
it
</dct:language>
```

```
<skos:notation rdf:datatype=’’http://dewey.info/schema-terms/
    Notation’’>
641
</skos:notation>
<skos:inScheme rdf:resource=’’http://dewey.info/scheme
    /2007/02/about.it’’/>
<skos:prefLabel xml:lang=’’it’’>
Cibi e bevande
</skos:prefLabel>
<skos:narrower rdf:resource=’’http://dewey.info/class
    /641.01/2007/02/about.it’’/>
<skos:narrower rdf:resource=’’http://dewey.info/class
    /641.2/2007/02/about.it’’/>
<skos:narrower rdf:resource=’’http://dewey.info/class
    /641.3/2007/02/about.it’’/>
<skos:narrower rdf:resource=’’http://dewey.info/class
    /641.4/2007/02/about.it’’/>
<skos:narrower rdf:resource=’’http://dewey.info/class
    /641.5/2007/02/about.it’’/>
<skos:narrower rdf:resource=’’http://dewey.info/class
    /641.6/2007/02/about.it’’/>
<skos:narrower rdf:resource=’’http://dewey.info/class
    /641.7/2007/02/about.it’’/>
<skos:narrower rdf:resource=’’http://dewey.info/class
    /641.8/2007/02/about.it’’/>
<cc:morePermissions rdf:resource=’’http://www.oclc.org/dewey/
    about/licensing/’’/>
<dct:created rdf:datatype=’’http://www.w3.org/2001/XMLSchema#
    dateTime’’>
2000-01-01T00:00:00.0+01:00
</dct:created>
<dct:modified rdf:datatype=’’http://www.w3.org/2001/XMLSchema#
    dateTime’’>
2006-01-28T22:04:16.000+0100
```

```
</dct:modified>
</rdf:Description>
<rdf:Description rdf:about=’’http://dewey.info/class/641/’’>
<rdf:type rdf:resource=’’http://www.w3.org/2004/02/skos/core#
    Concept’’/>
<rdf:type rdf:resource=’’http://www.w3.org/2002/07/owl#Thing
    ’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2009/08/about.en’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2009/08/about.fr’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2009/08/about.es’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2009/08/about.sv’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2009/08/about.pt’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2009/08/about.ru’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2009/08/about.zh’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2009/08/about.ar’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2009/11/about.af’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2010/03/about.no’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2009/03/about.de’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2009/07/about.en’’/>
<dct:hasVersion rdf:resource=’’http://dewey.info/class
    /641/2007/02/about.vi’’/>
```

```
   <dct:hasVersion rdf:resource=''http://dewey.info/class
       /641/2007/02/about.it''/>
   <dct:hasVersion rdf:resource=''http://dewey.info/class
       /641/2011/07/about.gd''/>
   <dct:hasVersion rdf:resource=''http://dewey.info/class/641/e23
       /''/>
   <skos:notation rdf:datatype=''schema-terms/Notation''>641</
       skos:notation>
   <skos:notation rdf:datatype=''http://dewey.info/schema-terms/
       Notation''>641</skos:notation>
   <skos:inScheme rdf:resource=''http://dewey.info/scheme/ddc/''/
       >
   <skos:inScheme rdf:resource=''http://dewey.info/scheme/
       schedule/600/''/>
   </rdf:Description>
   </rdf:RDF>
```

Finally we needed an appropriate license model. We make data on dewey.info available under a Creative Commons BY-NC-ND license.[2]Licensing information is embedded in RDF and RDFa following the Creative Commons Rights Expression Language (ccREL) specification.[3] In the RDF/SKOS extract above, the following licensing information is embedded in the RDF:

Listing 2: CC license embedded in RDF/SKOS

```
   <xhv:license rdf:resource=''http://creativecommons.org/
       licenses/by-nc-nd/3.0/''/>
     <cc:attributionName>OCLC Online Computer Library Center, Inc.
         </cc:attributionName>
     <cc:attributionURL rdf:resource=''http://www.oclc.org/dewey
         /''/>
```

[2] http://creativecommons.org/licenses/by-nc-nd/3.0.

[3] http://wiki.creativecommons.org/CcREL.

```
<cc:morePermissions rdf:resource=’’http://www.oclc.org/dewey
    /about/licensing/’’/>
```

A year after the initial offering, we extended the data available in dewey.info with the addition of assignable numbers and captions from Abridged Edition 14 in three languages (English, Italian, and Vietnamese). This extension added about 3500 additional records for each language to the data already available in dewey.info. While the DDC Summaries represented a broader set of languages than available in the full and abridged translations, the new abridged-edition offerings were a subset of the languages in which the edition had been translated. Why were English, Italian, and Vietnamese chosen? The simple answer was that each was available in the same proprietary format, ESS XML, for which we already had an RDF/SKOS transformation.

Parallel to the linked data work, the Dewey editorial team was making a major data transformation of another type—moving from the proprietary ”ESS” format to one based on the MARC 21 Classification and Authority formats. In 2009, the DDC Summaries were transformed from ESS XML to RDF/SKOS; we used the same transformation to make the Abridged Edition 14 data available in dewey.info. In 2010, OCLC moved to a new underlying representation for the DDC, adopting one based on the MARC 21 formats for classification data (to represent class records) and authority data (to represent Relative Index and mapped terminologies associated with class records). At the same time, OCLC adopted MARCXML as the distribution and ingest format for DDC data across versions, and moved to a new data distribution and ingest model (previously, data transfers were handled at the individual file level over an ftp site). We made a decision to delay the distribution of additional DDC data in dewey.info until we could productionize the data transformation and distribution process operating on the new format and within the distribution environment. This meant taking the data encoded in MARCXML from the distribution server, applying the RDF/SKOS transformation stylesheet, and associating the result with a ″subscription,″ automatically creating an Atom feed of data sets that a user agent (in this case, dewey.info) could pick up from the distribution server over a RESTful interface. A model of the process is shown in figure 2.

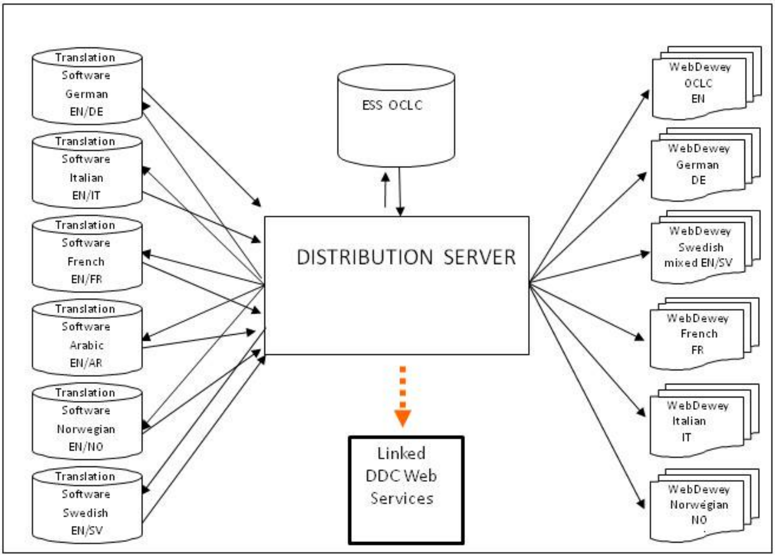

Translation Software German EN/DE
Translation Software Italian EN/IT
Translation Software French EN/FR
Translation Software Arabic EN/AR
Translation Software Norwegian EN/NO
Translation Software Swedish EN/SV
ESS OCLC
DISTRIBUTION SERVER
Linked DDC Web Services
WebDewey OCLC EN
WebDewey German DE
WebDewey Swedish mixed EN/SV
WebDewey French FR
WebDewey Italian IT
WebDewey Norwegian NO

Figure 2: Dewey distribution environment.

We installed the pieces on the distribution server that would make this possible in May 2012. In in mid-June 2012, we added assignable numbers and captions from the DDC 23 schedules will be available to dewey.info ; this addition of over 38,000 numbers increased the available Dewey linked data nearly tenfold. In August 2012, we further extended Dewey linked data by adding the assignable notation and captions from Table 2 (the Dewey geographic table).

Next steps

Our next planned offering is the linking of ”new acquaintance,” GeoNames, to Table 2 data. Because we want to manage all editorially curated data (including mappings) with the OCLC ESS system, this will require short-term and long-term changes to geographic data within the system. In order to allow the provision of geographic data on the class level, the Dewey editorial team developed MARC PROPOSAL NO. 2011-10,[4] which was approved by MARBI in June 2011. The proposal defines new fields that allow for the storage and display of geographic codes in MARC classification records, thereby enabling the reuse of parts of the Relative Index links to GeoNames (generated by the matching algorithm) on the class level in applications downstream, e.g., in linked data representations of the DDC.

Use cases

In addition to linking plans, we report on use cases that facilitate machine-assisted categorization and support discovery in the Semantic Web environment. It is important to have use cases for Dewey linked data, and to solicit new use cases that might inform decisions about our data offering. Institutions such as Bibliothèque nationale de France, the British Library, and Deutsche Nationalbibliothek have made use of Dewey linked data in bibliographic records and authority files .FAO has linked AGROVOC to our data at a general level. We are also exploring links between the DDC and other

[4] http://www.loc.gov/marc/marbi/2011/2011-10.html.

value vocabularies such as VIAF, FAST, ISO 639-3 language codes, and MSC (Mathematics Subject Classification). Today, we would like to focus on three uses cases, a caption service, the ”old friend” of DDC synthesized number components associated with categorized content, and the ”new acquaintance” of DDC-GeoNames links.

Caption service

Querying Dewey linked data

The first use case is a simple one: querying Dewey linked data by a Dewey number to have the associated caption delivered as an explanation of the number. For example, the query http//dewey.info/class/945.5/about will return information about class 945.5, including the captions ”Regione della Toscana” and ”Tuscany (Toscana) region.” There are also two ways in which this data is made accessible to machines and can therefore be used in an automated way as part of a library catalog or other discovery tool. The HTML page for class 945.5 contains structured data in RDFa markup, which means that user agents will be able to distill caption information as regular RDF triples.

Another very powerful and flexible way is directly accessing the triple store using the SPARQL endpoint.

Listing 3: Query that returns all distinct captions associated with class number 945.5

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT DISTINCT ?caption WHERE {
  {GRAPH ?g
     {?concept skos:notation ’’945.5’’^^<http://dewey.info/
          schema-terms/Notation>;
               skos:prefLabel ?caption
```

```
      }
    }
  }
```

Note that the endpoint supports HTTP bindings of the SPARQL protocol, meaning that the endpoint serves as a general web service interface (in case the linked data presentation is not preferred).

DDC-DDC number components links

The second use case is an enhancement of data in dewey.info taken from the DDC itself: links to Dewey synthesized number components. The concept is simple: What if we linked every synthesized number to its component parts? For example, 641.59455 represents the cooking of Tuscany (641.59 Cooking characteristic of specific continents, countries, localities + T2—455 Tuscany [Toscana] region). The underlying Dewey data includes the MARC 21 765 Synthesized Number Components field: 765 0# $b641.59 $z2$s 455 $u641.59455 By establishing a link between 641.59455 and T2—455 (represented as ""$z 2$s 455" in the 765 field and as "2–455" in the URI string), it is possible to isolate the geographic facet and use it to foster alternative approaches to discovery. The potential enhancements to such discovery is discussed in the next section.

DDC-GeoNames links

Linking Dewey data with GeoNames offers the opportunity to extend the boundaries of categorization and discovery. Since GeoNames has emerged as not only the dominant source for geographic coordinates in the linked data space, but also as a leading provider of identifiers (URIs) for geographic entities, a GeoNames term can act as a general equivalent or a boundary object for data from different domains that have never been directly mapped to each other. The linking of two concepts in different schemes or from different domains to the same GeoNames entity helps to establish a common ″aboutness″ of these two terms.

Figure 3 illustrates how a common link to a GeoNames term from a geographic class in dewey.info and from a New York Times subject heading for the same geographic area establishes a strong (albeit implicit and untyped) relationship between these two terms because both entities are ″about″ the same city. Also, by extension it can be assumed that all articles and other resources indexed with the NYT heading should be discoverable by the DDC class, therefore adding to the amount of categorized content that can be retrieved by using this DDC number in a discovery interaction. Links to datasets like GeoNames extend the boundaries of DDC classes on a conceptual level as well. Whereas a traditional mapping between KOS usually connects entities of the same type (e.g., concepts), linking in the sense of the Semantic Web can connect different kinds of named/i-

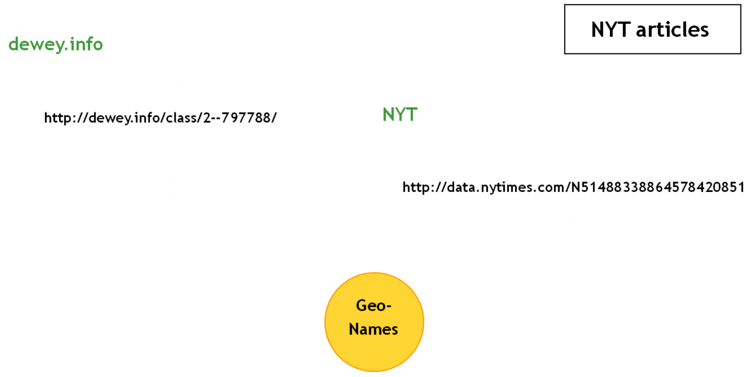

NYT articles
dewey.info
http://dewey.info/class/2--797788/
NYT
http://data.nytimes.com/N51488338864578420851
Geo-
Names

Figure 3: Links to GeoNarmes.

dentified entities. While a mapping between concepts often operates with variations of semantic relationships traditionally employed by thesauri (e.g., broader/narrower, related, whole/part), linking of different types of entities requires a new set of relationships tailored to the domain model of the linked dataset or value vocabulary. In the case of GeoNames, in order to store the links in MARC, we have to use a traditional mapping relationship. However, in a linked data version, the SKOS mapping relationships (corresponding to traditional thesaurus relationships) cannot be used to link Dewey classes and GeoNames terms, because GeoNames URIs identify a gn:Feature, which is defined as ″a geographical object″ and, being a subclass of http://schema.org/Place, as an entity with a ″physical extension.″ In other words, GeoNames (like many other ontologies) does not contain descriptions of or identifiers for concepts of places; it contains descriptions of and identifiers for the places themselves. In such cases, a relationship like foaf:focus should be used, which ″relates a conceptualisation of something to the thing itself.″ A GeoNames URI identifies a locality, not a concept of a locality. This operation effectively connects a Dewey concept with a different set of relationships, which can be used to present information seekers compelling tools to identify and select geographic features for resource discovery. In essence, it opens up a new perspective or viewpoint on the arrangement of classes in Dewey.

Figure 4 on the facing page shows in parallel two different kinds of neighborhoods applicable to T2—6626 Niger. The established Dewey ″neighborhood″ shows the class in the context of the DDC notational hierarchy. Linking this class to its corresponding GeoNames feature, however, allows for reusing GeoNames' gn:neighbour relationship and applying it directly to this Dewey class. The right-hand side shows the concept T2—6626 surrounded by features that neighbor the country in its foaf:focus in the physical world.

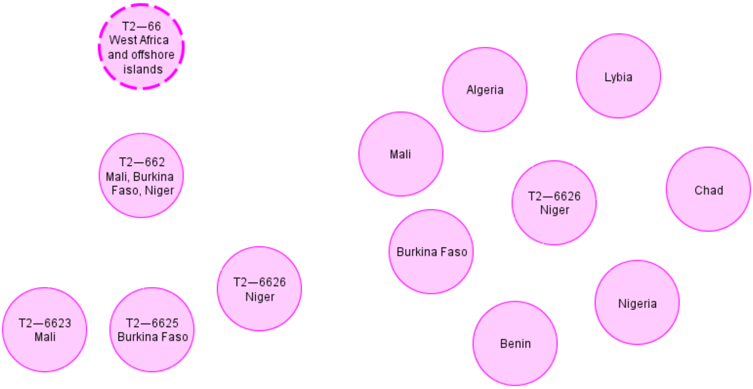


Figure 4: Two views of T2—6626 Niger.

Taking this one step further, linking all geographic Dewey concepts to GeoNames allows for an on-the-fly switching of the viewpoint as needed, effectively allowing for transforming the concepts temporarily into features, and, by using inherited properties like geographic coordinates, placing them on a map (figure 5 on the next page).

Furthermore, DDC classes can utilize more than just relationships inherited from geographic features. The links allow also for a more expressive typing of related DDC entities and open the door to geospatial reasoning over the underlying DDC data. For example, usually it is not clear whether a Dewey number represents a country (or another type of entity). But in the above example, the ″inherited″ types allow for basic viewpoint-transgressing queries such as: ″Display all Dewey numbers that represent countries that are adjacent to

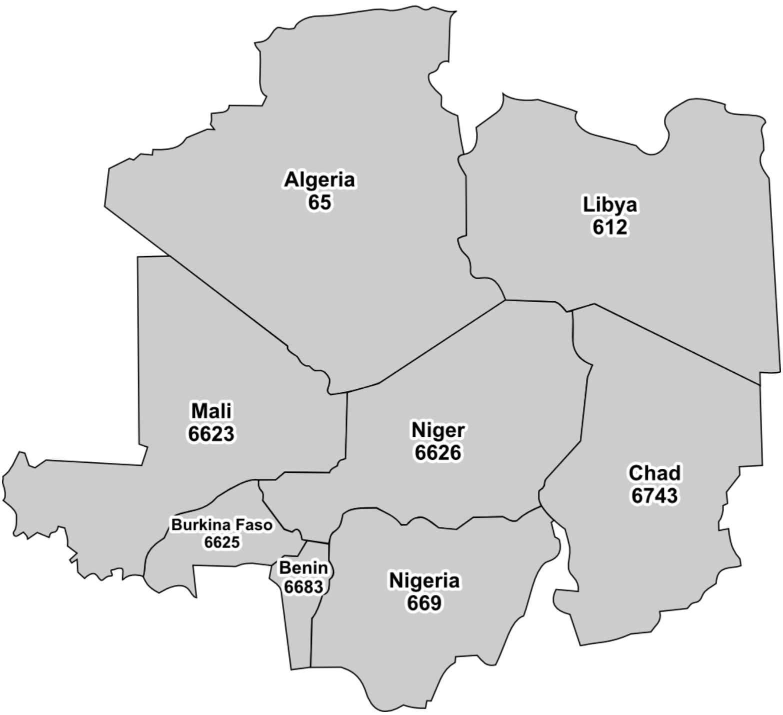


Figure 5: Blending of Dewey viewpoint and geographic viewpoints.

T2—6626."

Figure 6 shows another example of transgressing viewpoints. Table 2 is mainly arranged by continents, which means that countries that span different continents are separated notationally, i.e., they don't occupy a contiguous span of Dewey numbers. This may even be true for cities in these countries, e.g., Istanbul in Turkey occupies subdivisions of both T2—4 and T2—5. While Dewey provides all necessary relationships in order to relate the European and Asian parts of Turkey, they are divided notationally, making it not a simple task for a discovery system to offer the user a compelling way of selecting subentities for retrieval. Using the inherited gn:neighbour relationship, however, makes it easy to display classes about the European part of Turkey e.g., T2—49618, shown with its Relative Index terms in yellow) and the Asian part (e.g., T2—5632, shown with its Relative Index terms in green) together in a geobrowser like Google Earth using the geographic viewpoint.

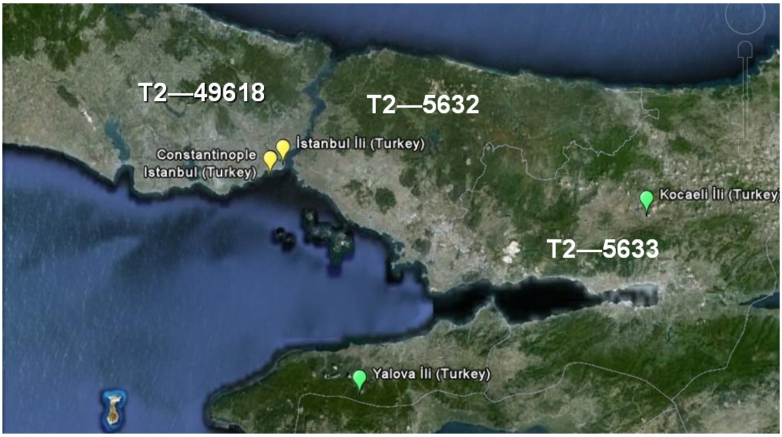


Figure 6: Overlaying Dewey classes and Relative Index terms on a map using properties of linked entries.

Conclusion

The contents of dewey.info and links to Dewey data have evolved over time as we have taken advantage of various opportunities for experimentation. With each addition, we have considered possible use cases for the additional data. The following statement appears in the last paragraph of the final report of the W3C Linked Library Data Incubator Group (2011) :

> Linked data follows an open-world assumption: the assumption that data cannot generally be assumed to be complete and that, in principle, more data may become available for any given entity.

The schema-less RDF data model allows for a substantial degree of freedom (compared to the relational database paradigm) in leveraging existing data by enrichment and addition of new connections almost ad hoc. Our efforts to publish the DDC as a linked data value vocabulary have taken place in a rich and evolving Dewey ecosystem. Figure 7 shows the current state of translations and versions published, planned, or under way based on DDC 23 data; where known, expected publication dates are shown in parentheses. Figure 8 shows the current mappings and crosswalks between the DDC and other knowledge organization systems.

We expect to continue extending linked DDC data within the rich environment described in figure 7 on the next page and figure 8 on the facing page to meet use cases in categorization and discovery.

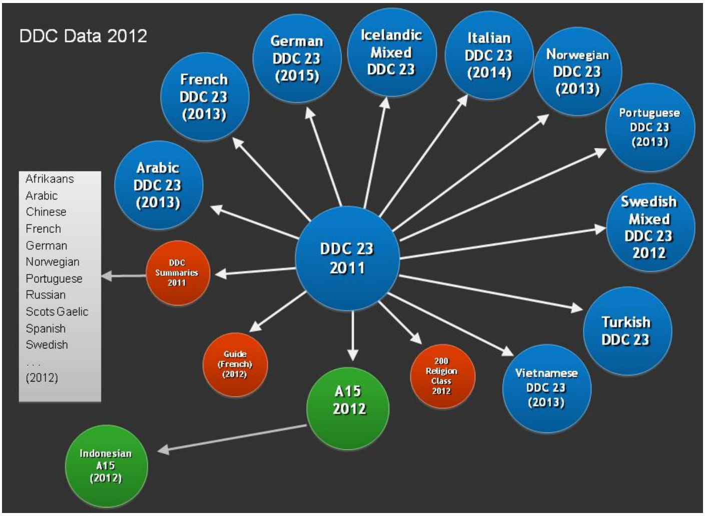


Figure 7: Editions and versions based on DDC 23.

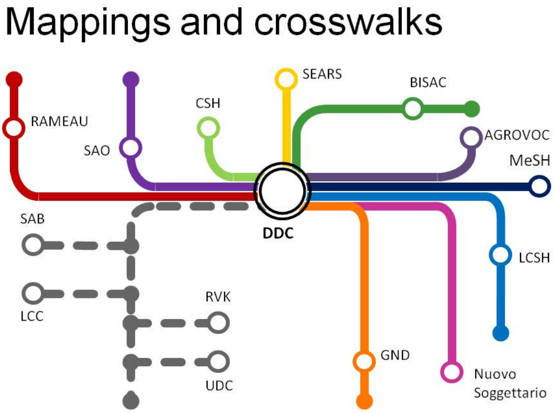


Figure 8: Mappings and crosswalks to the DDC.

References

Panzer, Michael. “DDC, SKOS, and linked data on the web”. Proc. of Everything Need Not Be Miscellaneous: Controlled Vocabularies and Classification in a Web World, Montréal, Canada, August 5 2008. 2008. http://www.oclc.org/news/events/presentations/2008/ISKO/20080805-deweyskos-panzer.ppt. (Cit. on p. 181).

Panzer, Michael and Marcia Lei Zeng. “Modeling Classification Systems in SKOS: Some Challenges and Best-practice Recommendations”. *Semantic interoperability of linked data: Proceedings of the International Conference on Dublin Core and Metadata Applications*. Seoul, October 12-16 2009. Ed. S. Oh, S. Sugimoto, and Sutton S.A. Seoul: Dublin Core Metadata Initiative and National Library of Korea, 2009. 3–14. http://dcpapers.dublincore.org/ojs/pubs/article/view/9748. (Cit. on p. 181).

Joan S. Mitchell, OCLC.
mitchelj@oclc.org
http://staff.oclc.org/d̃ewey/joan.htm
Michael Panzer, OCLC.
panzerm@oclc.org
http://staff.oclc.org/d̃ewey/michael.htm

Mitchell, J.S., M. Panzer. ”Dewey linked data: making connections with old friends and new acquaintances”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5467. DOI: 10.4403/jlis.it-5467. Web.

Abstract: This paper explores the history, uses cases, and future plans associated with availability of the Dewey Decimal Classification (DDC) system as linked data. Parts of DDC system have been available as linked data since 2009. Initial efforts included the DDC Summaries in eleven languages exposed as linked data in dewey.info. In 2010, the content of dewey.info was further extended by the addition of assignable numbers and captions from the Abridged Edition 14 data files in English, Italian, and Vietnamese. During 2012, we will add assignable numbers and captions from the latest full edition database, DDC 23. In addition to the ”old friends” of different Dewey language versions, institutions such as the British Library and Deutsche Nationalbibliothek have made use of Dewey linked data in bibliographic records and authority files, and AGROVOC has linked to our data at a general level. We expect to extend our linked data network shortly to ”new acquaintances” such as GeoNames, ISO 639-3 language codes, and Mathematics Subject Classification. In particular, the paper examines the linking process to GeoNames as an example of cross-domain vocabulary alignment. In addition to linking plans, the paper reports on use cases that facilitate machine-assisted categorization and support discovery in the semantic web environment.

KEYWORDS: DDC; Dewey linked data; Dewey Decimal Classification

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

Linked Heritage: a collaborative terminology management platform for a network of multilingual thesauri and controlled vocabularies

Marie-Veronique Leroi

Context

The semantic web and linked data are now well known principles of the Web. Applying the rules of the linked data has been defined as a priority for the development of Europeana, the European Digital Library. Although the normalisation of metadata is a work that has been investigated for years, the importance of terminologies to understand and exploit these metadata in a structured way has risen to the foreground thanks to the linked data and semantic web.

The semantic web is defined by Tim Berners-Lee as "the Web of data with meaning in the sense that a computer program can learn enough about what the data means to process it". The Web is not about documents anymore but it is about data. Therefore Linked data can be explained as "The semantic web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data". The Web of Data makes

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5471

sense only if the data are linked.

Many European projects are contributing specific content to Europeana. Cultural heritage is rich thanks to its diversity. Galleries, Libraries , Archives and Museums, also known as the GLAM sector, have a different perception of cultural objects. Besides the diversity of the content itself, there is also a huge diversity of languages and expert terminologies. There is therefore a need for the harmonisation of terminologies at European level to enable a better understanding of the content available at European level.

The Michael project[1] was one of the first European project to take into account the issue of multilingualism in a European context. Indeed this portal offers a multilingual access to digital cultural heritage in 12 languages. The Michael Culture association that has been created after the end of the project in order to sustain the activity of the European portal has been involved in the European projects contributing to Europeana.

Athena has been one of the major projects contributing content to Europeana since almost 1,8 million objects descriptions were provided via this initiative. We present in this paper how the work on terminologies initiated within the Athena Project is now reused and implemented within the Linked Heritage project.

Athena

The Athena Project that started in November 2008 and ended in April 2011, aimed at providing content from European museums to Europeana. The Michael Culture Association has been involved in the Athena Project as workpackage leader for the one dedicated to terminology and multilingualism (WP4). The main objectives of this

[1] Michael: http://www.michael-culture.org

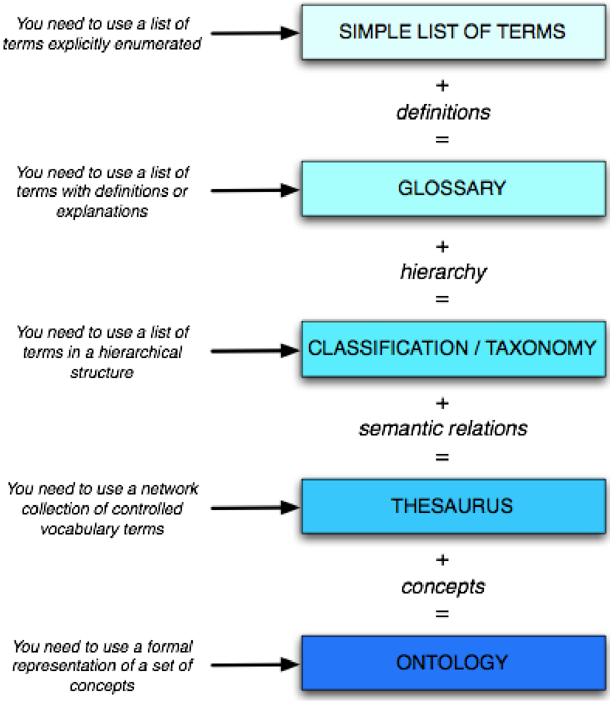

You need to use a list of terms explicitly enumerated
SIMPLE LIST OF TERMS
+
definitions
=
You need to use a list of terms with definitions or explanations
GLOSSARY
+
hierarchy
=
You need to use a list of terms in a hierarchical structure
CLASSIFICATION / TAXONOMY
+
semantic relations
=
You need to use a network collection of controlled vocabulary terms
THESAURUS
+
concepts
=
You need to use a formal representation of a set of concepts
ONTOLOGY

Figure 1: Kinds of terminology resources

workpackage were to provide the European museums with recommendations for terminology management. Therefore we proceeded with a definition and presentation of the different kinds of terminology resources. Indeed as many kinds of terminology resources can be hidden behind the «controlled vocabulary»phrase, we made a synthetic view in order to guide the institutions for answering our survey.

The survey launched among the Athena Partners was intended to

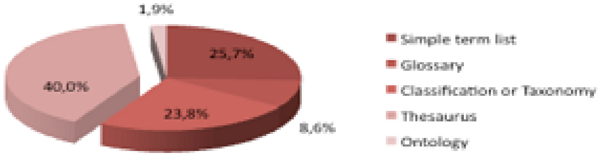


Figure 2: Kinds of terminology

achieve a complete state of the art and give us a precise overview of the terminology use and methods in the European museums. All the results of the state of the art could be found on the Athena and Linked Heritage wiki and on the deliverable of the project.[2]
The results of the survey shown us that most of the cultural institutions, 40% (on 105 total answers) use a thesaurus-type terminology. Indeed most of the cultural institutions use a structured controlled vocabulary. Thesaurus offers both hierarchical and associative relations between the descriptors and it is then a very powerful and simple tool for indexing and cataloguing.
Since it has been acknowledged as a W3C recommendation in August 2009 and since it is the format expected by Europeana, SKOS (Simplified Knowledge Organisation System) is used more and more to bring all the terminology resources into a interoperable format. Very few of the institutions who answered the survey had already SKOSififed terminology, e.g. terminology converted in theSimplified Knowledge Organisation System (SKOS) format. Most of the museums manage their vocabularies within their collections management system. These tools are generally proprietary and often allow for an XML export and in some cases for a SKOS export. Besides the iden-

[2] D4.1: identification of terminology resources in museums: http://www.athenaeurope.org/getFile.php?id=398

tification of terminology resources in use in the European museums, the second task of the workpackage was to define some guidelines and a tutorial for SKOS.

Recommendations

The first step is about the conception of your terminology. So to say, at this stage an institution manage its terminology "internally" in order to make a thesaurus in a "human" perspective. We gave the different key-steps for an institution that is about to create on its own a new terminology or adapt a terminology already in use in order to optimize your digital resources descriptions on Europeana. These operations have to be done in priority since they determine the two other steps. In this step, the institution has to think of the domain(s) the terminology will cover, who will be the users expected to use it, what will be the languages the terminology will be available in. The terms and their organisation within the thesaurus structure is defined at this level.

Then the second step consists in making the terminology interoperable. Now it is about rising the terminology out of the museum. Indeed when a terminology is SKOSified, it takes into account the machine perspective. With this second step, we are in the perspective of the semantic web. In this second step, the institution has to evaluate if SKOS is the relevant format for the kind of terminologies it manages. SKOS, for example, won't be the appropriate format in the case of authors/persons' names. After this evaluation, the main object of this step is to proceed with the conversion of the terminology into SKOS. Some tools are available for validating the SKOS output.

Finally we address the cultural institutions our last recommendations as they concern the networking of their terminology with

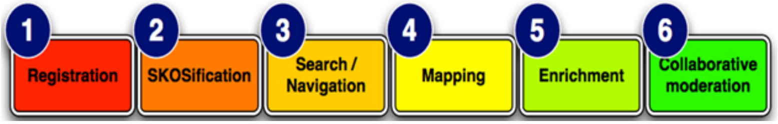


Figure 3: Benchmark - workflow

others. At this third and last stage, for an institution, it is about being visible in Europe in a network perspective by integrating their terminology in a network of SKOSified terminology. Define some metadata to describe the administrative details of the terminology is a necessary task within this step. On the basis of the metadata provided by other terminologies, terminology resources that can be mapped with the one of the institution will be identified. With this last step, we are in the perspective of the linked data.

From Athena to Linked Heritage

The WP3 of Linked Heritage relies on the legacy of the WP4 of the Athena Project and has been organised to tackle in the best way the following two aspects: content management and technical developments for terminology management. Within the Athena WP4, by gathering experiences from museums and other heritage domains, we identified the most logical process and functional needs related to the management, semantic interoperability and enrichment of terminologies. In doing this, the project identified some use cases and set up a benchmark. The work on the use cases made it clear that first of all a workflow specification was needed for a collaborative production and moderation of cultural heritage terminologies. With the help of the expert working group on terminologies and the uses cases defined earlier on, the following workflow was considered:

This defined workflow helped to evaluate existing tools, interfaces and methods that might be suitable for this work. Unfortunately none of the evaluated tools could actually manage the entire process as shown above. Some of the tools came close, offering search and navigation, as well as semantic mapping and enrichment in a collaborative environment. But there was no complete software environment available offering the possibility for cultural heritage institutions to upload, register and SKOSify the terminologies first, before proceeding to the next steps. This was exactly the point of failure in the workflow of existing tools, because most cultural heritage institutions use own in-house reference terminologies and haven't got the available resources for managing them in a standard interoperable format such as SKOS. This technological lack has been fully considered in the framework of the Linked Heritage WP3 as a complete software will be developed in order to tackle all the steps and process identified within Athena.

From theory to practice: Terminology Management Platform (TMP)

The Terminology Management Platform (TMP) will be this complete software for terminology management and is meant to follow the same structural workflow presented above. Linked Heritage goes beyond the work done in Athena with the purpose to develop a prototype of a tool able to deal with the different steps of the defined workflow and thus to lessen the economical efforts the institution has to make when wanting to share the terminology in an exchangeable format to the community and Europeana.

The work done in WP4 of the Athena Project resulted in the proposed solution to design and the implement an integrated software environment for terminology management, enabling any institution to manage its terminology according to Europeana ingestion rules. In Linked Heritage this proposed solution will become reality with the development of a prototype of a Terminology Management Platform (TMP) for the cultural heritage sector to collaboratively create a network of interlinked multilingual terminologies in a Europeana compliant format (SKOS). In the development of this platform, the expertises of four technical partners are brought together in a combined effort to create an integrated environment for terminology management. This approach of combining expertises and integrating existing tools into a single web environment, allows us to select and combine best practice technological features in a time and cost efficient way. Considering the outcomes from the Athena Benchmark and the first tasks achieved within Linked Heritage WP3, here are the identified features for the Terminology Managment Platform (TMP):

- to be a web service: For collaborative work online;
- to have a user-friendly GUI: Adapted for a non-expert use in European museums, lirabries and archives;
- to combine open-source components: Such a service must stay independent of proprietary codes and formats;
- to be logically structured with an intuitive Workflow: The user must find which actions to do according to his/her needs;
- to be flexible enough to be adapted to new standards: What if SKOS is updated in a new version or evolving towards an ontology description?

As said above, four technical partners are involved in the development of the TMP. Each of them has its own experience of development and usability and this diversity is very enriching for the whole

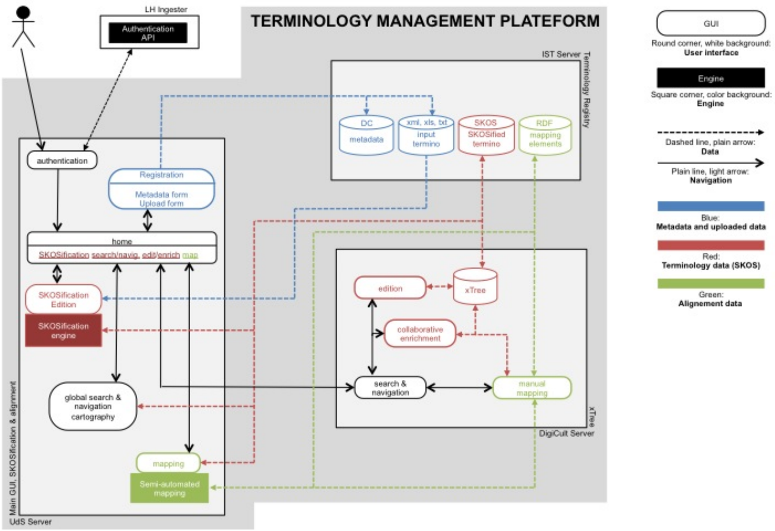


Figure 4: TMP Architecture

toolset. The schema in figure 4 represents the technical architecture of the TMP.

Digicult, the technical partner from Germany is already developing and maintaining an editing tool, xTree. The Institute of Science and Technology (IST) from Portugal has already been involved in projects contributing to Europeana and especially in the development of a metadata registry. This knowledge and experience is now applied for the Linked Heritage TMP since Istituto di Scienza e Tecnologia (IST) is in charge of developing the terminology registry of the TMP. The National Technical University of Athens was already involved in the Athena Project and is the creator of the Mint tool that is used for the ingestion of metadata. In order to guide the institutions who will use the TMP, we decided to use the same authentication process than the one used in the Mint tool for ingestion so institutions won't need to register themselves in the TMP if they are already using the Ingester. Finally, the University of Savoie from France who has a strong experience in knowledge engineering is in charge of developing a tool for SKOSification and for visual navigation. But most of all, University of Savoie is working on bringing together the different technologies and modules from the rest of the technical partners.

Conclusion

Linked Heritage take a great benefit from the Athena Project as a legacy since it reuses the Mint tool for ingestion and the LIDO format. In the field of terminologies it even goes one step beyond putting into practice all the recommendations and workflow defined in Athena. The Terminology Management Platform will be a great deal for the institutions who will be then autonomous for the management and mapping of their terminologies. This will also enable a better awareness on the semantic web and linked data issues.

MARIE-VERONIQUE LEROI, Ministère de la culture et de la communication (France).
marie-veronique.leroi@culture.gouv.fr

Leroi, M.V. ”Linked Heritage: a collaborative terminology management platform for a network of multilingual thesauri and controlled vocabularies”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5471. DOI: 10.4403/jlis.it-5471. Web.

ABSTRACT: Terminology and multilingualism have been one of the main focuses of the Athena Project. Linked Heritage as a legacy of this project also deals with terminology and bring theory to practice applying the recommendations given in the Athena Project. Linked Heritage as a direct follow-up of these recommendations on terminology and multilingualism is currently working on the development of a Terminology Management Platform (TMP). This platform will allow any cultural institution to register, SKOSify and manage its terminology in a collaborative way. This Terminology Management Platform will provide a network of multilingual and cross-domain terminologies.

KEYWORDS: Athena Project; Library linked data; Michael Project; Terminology Management Platform; TMP

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

The Nuovo soggettario as a service for the linked data world

Giovanni Bergamin, Anna Lucarelli

Introduction

The Nuovo Soggettario (hereinafter, NS) edited by the National Central Library of Florence (BNCF), is the main Italian subject indexing tool for various kinds of resources. It has been developed in collaboration with the Italian National Bibliography (BNI) which holds a leading role in the bulding and development of subject indexing tools in compliance with the International Federation of Library Association (IFLA) recommendations (The NS employment by Italian National Bibliography is also described in Jahns, *Guidelines for subject access in National bibliographies*) and other International standards. This tool is used by general and specialized Italian libraries (indexers, researchers, users), in particular those participating in the Servizio Bibliotecario Nazionale (SBN), and is also employable in archives, multimedia libraries and documentation centres. The NS entered into the tradition of the analytico-synthetic languages; the system consists of a semantic and syntactical apparatus and, in compliance with the uniform and specific heading principles, it is conceived as a system to be applied in both pre-coordinated (the terms are combined in subject strings) and post-coordinated indexing environments (the terms are extracted from a controlled

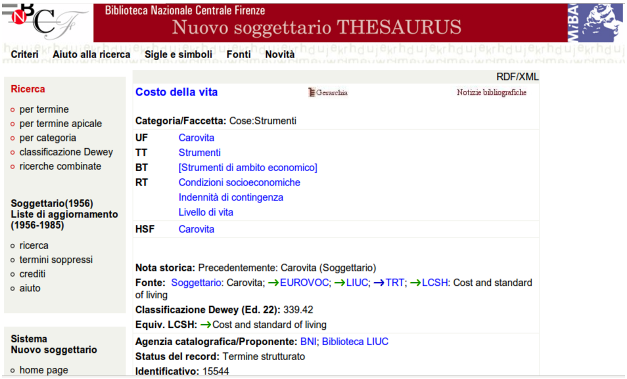
JLIS.it. Vol.4, n.1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5474

vocabulary and used as key words). The main component of the NS is a universal thesaurus built in compliance with the International standards, available online from the 2007.[1] It is a tool continuously being developed and currently accessible on the BNCF website. At the moment the Thesaurus consists of 46,000 terms derived from the 1956 Soggettario and its updates (which are being controlled and standardized), from new terms introduced for the semantic relationship network and from new terms proposed by the BNI indexers and other partners (Lucarelli et al., "The Nuovo soggettario Thesaurus: structural features and web application projects"). The terms are organized inside a structure based on four main categories and on semantic relationships determined by standards (*ISO2788:1986 – Documentation, guidelines for the establishment and development of monolingual thesauri. Documentation, principes directeurs pour l'établissement et le développement de thesaurus monolingue*; *ISO25964/1:2011 – Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval*). They are equipped by a rich apparatus of notes, connections with formerly preferred terms (historical variants), an indication of the correspondent numbers of the Dewey Decimal Classification, as well as by Sources which are in constant updating and employed for the control of morphologies and meanings.[2] The Thesaurus is integrated with the BNCF opac and with the opac of the other libraries that adopt it. The users can navigate from the controlled vocabulary to the bibliographic records. Regarding Linked data, the Thesaurus is linked with other thesauri, with some encyclopedias (such as Wikipedia and the prestigious Italian Treccani encyclopedia [3]), and with other cultural instituition's digital resources. The NS thesaurus promotes the Italian language and multilingual information retrieval by its data management software, however is

[1] http://thes.bncf.firenze.sbn.it/ricerca.php.

[2] http://thes.bncf.firenze.sbn.it/fonti.php.

[3] http://www.treccani.it.



Costo della vita

Categoria/Faccetta: Cose:Strumenti

UF	Carovita
TT	Strumenti
BT	[Strumenti di ambito economico]
RT	Condizioni socioeconomiche Indennità di contingenza Livello di vita
HSF	Carovita

Nota storica: Precedentemente: Carovita (Soggettario)
Fonte: Soggettario: Carovita; →EUROVOC; →LIUC; →TRT; →LCSH: Cost and standard of living
Classificazione Dewey (Ed. 22): 339.42
Equiv. LCSH: →Cost and standard of living
Agenzia catalografica/Proponente: BNI; Biblioteca LIUC
Status del record: Termine strutturato
Identificativo: 15544

also in compliance with standards (*Guidelines for Multilingual Thesauri*). A large number of terms has a cross-language equivalence relationship with Library of Congress Subject Headings (LCSH) preferred terms, displayed and linked by "Equiv. LCSH" note. i.e. «Costo della vita»:

In the last period, the NS is developed in two ways:

1. Interoperability: since 2010, metadata are available in Resource Description Framework (RDF)/SKOS format and will be employable in the Linked data world, not only in closely librarians contexts;

2. Automatic indexing: thesaurus is testing in automatic indexing of digital resources; in particular our goal is to reduce the cataloguing expenses.

These developments are outlined with the programs of other countries in the indexing domain, such as demonstrated by IFLA papers (Gömpel and Svensson, "Managing legal deposit for online publications in Germany").

SKOS standard for thesauri

Simplified Knowledge Organisation System (SKOS) is defined as a common data model,[4] developed by W3C Semantic Web Deployment Working Group (SWDWG),[5] for sharing and linking knowledge organization systems (such as thesauri, taxonomies, classification schemes and subject heading systems) within the semantic web. It is an application of the RDF. The most important thesauri, developed by National Libraries, are progressively adopting this standard for their controlled vocabularies. SKOS data are concepts which are independent of the terms used to label them, tagged as RDF triples and encoded using any concrete RDF syntax. The concepts, which are expressed by preferred terms in the thesaurus and used as descriptors in indexing system, are identified with URIs and are labeled with skos:prefLabel, expressed in one or more natural languages. The standard assigns alternative lexical labels to conceptual resources which have not a URI: skos:altLabel to represent a relationship between terms in a thesaurus that both represent the same concept; skos:hiddenLabel to represent misspelled variants of other lexical labels, abbreviations and acronyms. The standard expects the possibility to define and qualify the concept with some other information expressed by some labels which came from skos:note superclass (skos:definition; skos:scopeNote; skos:example: gives examples for the use of the terms; skos:historynote: it may be applied to a preferred or non-preferred term or to a concept. It should be used when a new preferred term is added to the thesaurus or change is made to an existing term that affects the concept's scope in different periods of application; skos:editorialnote: gives some administration information; skos:changenote: documents the different choices and modifications). The hierachical ad associative

[4] http://www.w3.org/TR/skos-reference.

[5] http://www.w3.org/2004/02/skos.

thesaural relationship, established between concepts, are labelled with skos:broader, skos:narrower, skos:related.

NS in SKOS format

Our thesaurus has been converted in SKOS format at the beginning of 2010. It was presented as a prototype at the IV Summit di Architettura dell'informazione (Motta and Rodighiero, "Il thesaurus del Nuovo soggettario interpreta SKOS") and then improved within the Digital resources automatic indexing project, developed in the BNCF since 2011 (Viti, "Interoperabilità fra thesauri generali e thesauri specialistici in ambito economico-finanziario. Il caso del Nuovo soggettario"). Our work has followed many stage and now is growing gradually in comparison with current developments. One of the most important problems starting with the prototypal stage was about the impossibility that SKOS – even if it defines an expressive array of sibling terms and collections of concepts – recognizes node labels as conceptual units which belong to hierarchical relationships; the standard calls them exclusively skos:Collection. The application doesn't establish links between the members of arrays and the general concept which expressed the same array. Instead each member of the array (skos:member) is directly linked with the concept which comes before the node label and not with the array identified by skos:Collection. Through the URI's skos:Concept we could verify if a skos:Member belongs to a skos:Collection and rebuild the whole hierarchical relationships. For example, a direct link can not be established between the skos:Concept Bambini, skos:Collection [Bambini secondo l'attività] and skos:member Bambini artisti. During our conversion we have found other problems; in particular, there where some difficulties for translation of two types of semantic relationships:

1. historical variants relationship (expressed with HSF, Historical see for) links some preferred terms with some preferred terms in the past which are no longer accepted;

2. the multi-word terms splitting relationship (expressed with USE+/UF+) create reciprocal link between multi-word terms and single word terms derived from factoring.

In the first case, we have refined the historical variants tagged skos:altLabel class as sogi:obsoleteTerm. Practically, the the historical variants begin a non preferred term. About the splitting of the complex concepts, at the moment, we have decided not to implement the SKOSXL extension (which identifies also the terms by an URI, not only the concepts), because about this we have not found some examples of applications. At the moment, the splitting relationship is expressed by a note in a specific field. The apparatus of note (definition, scope note, history note, sources, DDC...) is suitably expressed by SKOS. The syntactical note, that in the thesaurus guides the subject strings constructions, is labelled with skos:example. The assignment of an URI to the concepts promote the interoperabilty between different KOS, that is the possibility of mapping the semantic entities of different conceptual schemes. To realize this aim, the standard establishes three different equivalence levels: skos:closeMatch; skos:exactMatch; skos:broaderMatch e skos:narrowerMatch; skos:relatedMatch.[6] About this, we are testing the creation of equivalences to support the linked data between NS terminology and its equivalents in another vocabularies. We have chosen an empiric approach, based on an international reconnaissance of others SKOS applications. During the creation or maintenance of the NS equivalences can be activated by:

[6] http://www.w3.org/TR/skos-reference.

1. entering in a specific field (Source) the name of the vocabulary you want to cite: if the cited vocabulary is available SKOS, SKOS relationship of NS will be enriched with skos:closeMatch. If the the cited vocabulary is not available in SKOS this citation will be used for the creation of a deep link to the vocabulary (i.e. a direct link to the corresponding term);

2. entering the equivalence in a specific field (Equiv. LCSH) which refers to the Library of Congress Subject Headings equivalences: also in this case we use closeMatch relationship which is conceptual wide-ranging than exactMatch which was used in the initial stage.[7]

AGROVOC	1070
DBPEDIA	800
LCSH	750
ThESS	450
RAMEAU	240
EUROVOC	80

We are testing the settlement of equivalence semantic levels, between NS and ThESS (the thesaurus of Mario Rostoni Library of the LIUC University), by skos:broaderMatch, skos:narrowerMatch, skos:relatedMatch tags.

[7] About this, we have analysed matching procedures between RAMEAU and LCSH, in which the link is an exactMatch or a closeMatch without equivalence level's identification. At the moment, the links between RAMEAU and LCSH are established with a closeMatch (one sense relationship: RAMEAU -> LCSH) while those between LCSH and RAMEAU are established with an exactMatch LCSH<>RAMEAU.

The NS for automatic indexing of digital resources

As already mentioned, in BNCF has been running since 2011 a prototype test for the use of NS for semiautomatic subject indexing of digital resources acquired through legal deposit.[8] The BNCF initiative is in line with other European national libraries initiative (for instance, the Deutsche Nationalbibliotek project in this field is a relevant one (Junger, "Can indexing be automated? - the example of the Deutsche Nationalbibliothek") and takes into account two objectives:

1. the need for change in cataloguing practices due to rising amount of publications in digital format;
2. the sustainability of subject indexing.

Here "automatic indexing" refers to procedures using algorithms and techniques – coming also as result of the latest technological research – that can be used for automatic (or semi-automatic) extraction from a text of "relevant" keywords / key phrases. These procedures may be based on keywords / key phrases extraction and assignment with or without support of a controlled vocabulary. According to recent tests in progress at the international level, automatic indexing seems to produce better results – in term of precision and recall – if assisted by controlled lists (such as thesauri). In our prototype, the process of extraction of keywords / key phrases is managed by the software application Keyword indexer (Biblioteca Nazionale Centrale di Firenze, "Procedure automatizzate di estrazione di parole e frasi chiave: specifiche tecnico-funzionali"). This application requires, as preliminary step, the

[8] The prototype was developed in collaboration with two Italian companies: Casalini libri http://www.casalini.it and @Cult http://www.atcult.it.

creation of a knowledge base (also called learning model) based on sample documents (with associated metadata) and a vocabulary in SKOS format. In particular, as a first test, we created a thematic learning model on the economic and financial sectors, using the following structural components:

1. set of digital full-text documents: a sample of Italian doctoral thesis belonging to the economic and financial sector according to the classification system determined by the MIUR (Ministry of Education, University and Research): the classification symbols are SECS-P/01-13 and SECS-S/01-06;

2. set of metadata associated with the selected set of documents;

3. Nuovo Soggettario (NS) in SKOS format;

This model has been then applied to indexing the 2010-2011 issues of the digital journal LIUC Papers.[9]Keyword Indexer software, using TF/IDF (Term Frequency/Inverse Document Frequency) algorithm, was used to determine the ranking of terms . Obviously final results were affected by every variation of the above parameters.[10] Obviously final results were affected by every variation of the above parameters. For the time being, considering the last configuration of our test (choice of metadata closest to the semantic content

[9] Italian monthly journal focused on social science and in particular on Economics and Management http://www.biblio.liuc.it/pagineita.asp?codice=82. It is edited by Mario Rostoni Library of Carlo Cattaneo University in Castellanza (LIUC) which cooperate with the NS project.

[10] «The TF/IDF weight (term frequency–inverse document frequency) is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The TF/IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others».http://en.wikipedia.org/wiki/Tf*idf

of the document such as title + abstract, title + MIUR - Ministero dell'Istruzione, dell'Università e della Ricerca - classification symbol), findings are to be considered provisional: in this case automatic indexing is not closest enough to intellectual indexing. For these reasons we plan to continue our tests taking into account:

1. a multidisciplinary learning model for the general needs of a National library;
2. refinement of procedures for preparation of metadata to be used for building the learning model: we are considering both intellectual indexing and/or new automatic procedures for extracting topic keywords which could be used as metadata.

In any case is worth considering that all the tests are based on reuse of open source software components freely available on the net.

NS and the Semantic web

For interoperability with other applications, NS is available through the Zthes protocol.[11]Zthes is essentially an evolution of Z39.50-based information retrieval protocol, where the targets are not library catalogs but controlled vocabularies in compliance with ISO 2788 and ISO 5964. Through Zthes, applications can exchange data using the well-known and established mechanism of application interfaces known as Application Programming Interface (API)s. In particular Zthes uses SRU syntax (Search-Retrieval via URL) where requests for access to a controlled vocabulary are included as a parameters within a URL and response messages are tagged using XML syntax: in other words, Zthes uses http protocol - designed for interaction between the user (browser) and machine (web

[11] http://zthes.z3950.org.

Server) - for communication between machine and machine. API based on Zthes are easy to implement but they must however deal with the limitations of all the API based on HTTP protocol and XML syntax for the exchanged messages. In particular, an important limitation is the fact that in general API are not reusable - either at the protocol level or at message encoding level - in different contexts (a custom API is need for different kind of application). Interoperability through the infrastructure of the semantic web (RDF language and the SPARQL protocol in particular) certainly overcomes APIs limitations and this fact is the main reason for making available the NS using SKOS/RDF. As many have noted the success of the semantic web depends on widespread use including the ability of penetration into everyday applications that we use to access information. Among them search engines play an important role and, on the other hand, one can certainly argue that if the search engines are not interested in semantic web, there are little chance to establish semantic web as a widespread infrastructure for the information access. Search engines have long been interested in the semantic of documents (interested in indexing coded data within documents). The recent agreement – known as schema.org[12]– between the most important search engines (Google, Yahoo, Bing and Yandex) with the purpose to commonly define a standard to describe elements within HTML (HTML5) pages based on RDF, can (or should) be an interesting opportunity for libraries (Ronallo, "HTML5 Microdata and Schema.org"). Use of schema.org metadata set - in fact a simple extension of the HTML tags - will allow search engines to "understand" the structure and the nature of a given document. To remain in the library world, as we know search engines already can index the bibliographic records but treating them like any HTML page losing the ability to identify the semantic structure

[12] http://schema.org.

(the elements that characterize the bibliographic record). With a HTML/RDF coding based on schema.org our catalogs, thanks to metadata they contain, will be interpreted as "semantic objects" by the major search engines. This will increase the value of the information produced by the libraries increasing also the likelihood of bringing together '"supply and demand". Of course schema.org is not proposing a new model for bibliographic record, but within the library world schema.org can be used as as strategy to promote on the web here and now the information we produce (value and limitations included). Schema.org has recently decided also to maintain a list of suggested extensions.[13] This list will include both basic and widely used vocabularies (e.g. Wikipedia), and vocabularies produced thanks to a "significant professional contribution" (LCSH is cited as an example). Since for schema.org there is no limitations for extensions, this list will be used by search engines as a priority indication for inclusion of content in the new "semantic indexing service". NS, available as SKOS/RDF, is ready to become also an extension for schema.org for people accessing search engines using Italian language.

References

Lucarelli, Anna, et al. "The Nuovo soggettario Thesaurus: structural features and web application projects". *Subject access. Preparing for the future*. Ed. Patrice Landry et al. Berlin/Munich: De Gruyter Saur, 2011. 155–168. (Cit. on p. 214).

ISO2788:1986 – Documentation, guidelines for the establishment and development of monolingual thesauri. Documentation, principes directeurs pour l'établissement et le développement de thesaurus monolingue. Geneva: International Organization for Standardization, 1986. (Cit. on p. 214).

Classification, IFLA and Indexing Section. Working Group on Guidelines for Multilingual Thesauri. *Guidelines for Multilingual Thesauri*. 2009. (Cit. on p. 215).

[13] http://www.w3.org/wiki/WebSchemas/ExternalEnumerations.

Gömpel, Renate and Lars G. Svensson. “Managing legal deposit for online publications in Germany”. IFLA, 2011. (Cit. on p. 215).

Motta, Marta and Dario Rodighiero. “Il thesaurus del Nuovo soggettario interpreta SKOS”. IFLA, 2010. (Cit. on p. 217).

Viti, Elisabetta. “Interoperabilità fra thesauri generali e thesauri specialistici in ambito economico-finanziario. Il caso del Nuovo soggettario”. Diss. Università degli studi di Udine, 2012.

Junger, U. “Can indexing be automated? - the example of the Deutsche Nationalbibliothek”. 2012. (Cit. on p. 220).

Biblioteca Nazionale Centrale di Firenze, @Cult. “Procedure automatizzate di estrazione di parole e frasi chiave: specifiche tecnico-funzionali”. 2011. 9–11. [Documentazione interna alla BNCF, non pubblicata]. (Cit. on p. 220).

Ronallo, Jason. “HTML5 Microdata and Schema.org”. *Code4lib jurnal* 16. (2012). <http://journal.code4lib.org/articles/6400>. (Cit. on p. 223).

Jahns, Y., ed. *Guidelines for subject access in National bibliographies*. Berlin/Münich: De Gruyter Saur, 2012. 77–79. (Cit. on p. 213).

Biblioteca Nazionale Centrale di Firenze. *Nuovo soggettario. Guida al sistema italiano di indicizzazione per soggetto. Prototipo del Thesaurus*. Milano: Editrice Bibliografica, 2006.

ISO25964/1:2011 – Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. Geneva: International Organization for Standardization, 2011. (Cit. on p. 214).

GIOVANNI BERGAMIN, National Central Library of Florence.
giovanni.bergamin@bncf.firenze.sbn.it

ANNA LUCARELLI, National Central Library of Florence.
anna.lucarelli@bncf.firenze.sbn.it

Bergamin, G., A. Lucarelli. ”The Nuovo soggettario as a service for the linked data world”. *JLIS.it*. Vol.4, n.1 (Gennaio/January 2013): Art: #5474. DOI: 10.4403/jlis.it-5474. Web.

ABSTRACT: Nuovo soggettario (NS), edited by the National Central Library of Florence, is the Italian subject indexing tool for various types of resources. It has been developed in compliance with the IFLA recommendations, and other international standards in the field of subject indexing. This tool has been created for general and specialized Italian libraries, and for museums, multimedia libraries, archives and documentation centres. The main component of the NS is a general thesaurus available on the web since 2007 (http://thes.bncf.firenze.sbn.it/ricerca.php). The thesaurus comprises nowadays approximately 46.000 terms and is updated. It supports the new subject indexing practices and manages terminology deriving from collaboration between the BNCF and other libraries. The project is evolving in many directions and supporting interoperability. The main goal of the availability – since November 2010 – of the NS dataset in SKOS/RDF format, is to promote the use of this tool also beyond the traditional library environment. In this context three working areas have been taken into account: 1) improve accessibility and usability of the NS in the linked data environment: SPARQL endpoint, mapping to other datasets (including LCSH, RAMEAU, AGROVOC, EUROVOC, DBpedia); address the costs of bibliographic control starting from a project of automatic indexing (quality controlled) using NS in SKOS /RDF format and open source software tools; 3) cooperate with other institutions that are publishing linked open data.

KEYWORDS: Library linked data; Nuovo soggettario; SKOS

ACKNOWLEDGMENT: Thanks to Maria Grazia Pepe and to Elisabetta Viti

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

Annotation schema for legal doctrine: a case study on DoGi database

Tommaso Agnoloni, Elisabetta Marinai,
Ginevra Peruginelli, Maria Teresa Sagri,
Daniela Tiscornia

The context

Legal information has specific features due to its nature, different purposes and uses made of it, while preserving a strong requirement for integration of its different sources: legislation, case law and doctrine. Not only have these three sources different functions and purposes, but they require peculiar solutions for ensuring their access. In spite of this diversity there is a strong need, both at national and international level, of new forms of integration between these different sources to ensure effective sharing and interconnection of legal data. In traditional legal information systems data-sets are stored, organized and classified according to the classical distinction among types of legal sources. Apart from the European system (Eur-Lex) and few commercial initiatives the provision of integrated access to data originating from a wide range of distributed, multilingual and heterogeneous data sources is still scarce. Therefore,

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5480

users are obliged to access several providers of legal information and perform their searches step by step, by identifying specialized data repositories, assessing their own search strategy, browsing several databases, and establishing conceptual and legal connections by themselves. Of course there is quite a number of providers of legal information users can turn to in their research activity. As a matter of fact a user should first examine the legal sources hosting the desired information and then locate and check the availability of the resource. In particular, legal doctrine plays a key role in integrating various types of information. It comprises a vast amount of scientific literature produced by a variety of legal sources and meets essential requirements such as interpretation and understanding of rules and judgements as clarified and commented by legal scholars. As such, legal literature does reflect the legal debate within an ever growing and constantly changing society. By merging conceptual synthesis and critical analysis, legal literature can be seen as the final step in the creative process of the legal discourse, after the regulative step, the interpretative reconstruction and the comparison between the factual and the abstract description of a case.

The linked data approach in testing the DoGi-Legal Doctrine database

In the area of law the application of linked open data technology provides several advantages over the current management of legal data, while offering an opportunity for the development of new information systems to legal users.[1] In fact, the linked open data

[1] «A new approach for the notion of legal information is necessary, [. . .] by adapting the term to the new technological possibilities whilst ensuring the authority of the information as a legal source, law could become more of a logical networked system of rules and concepts, in line with that it is assumed to be» (Kirchberger).

environment provides the design principles and the technical infrastructure to publish structured data on the web so as to allow their interconnection and, finally, integration into a larger and distributed data space. The possibility to aggregate different data sets based on the systematic relationships between classes of sources, to meet and combine them regardless of their format and origin, is particularly interesting for the reconstruction of the complex structure of legal discourse. In the legal domain, where knowledge emerges from the connection among legislative, jurisprudential and bibliographic data,[2] the linked data methodology allows the creation of new repositories of knowledge. Furthermore, the development of a collaborative connection between different types of legal information contributes to ensure open access to digital legal knowledge and to develop a *seamless web* for the legal domain. Legal data sets, produced and published independently, are typically structured resources that can be enriched by extra-legal information content such as bibliographic citations related to authors' biographies. Therefore, linked data technologies can provide a standardized mechanism for publication of structured legal doctrine, with the opportunity to gather and explore external legal resources providing an interpretational context. The choice to use legal doctrine for testing a linked data approach in a legal environment is not accidental. For this study the DoGi-Legal Doctrine database[3] has been retained as a valuable resource due to its importance in documenting the evolution of the Italian legal science. The database is, at present, the most comprehensive source of online information on Italian doctrine. The Institute of Legal Information Theory and Techniques of the National

[2] «The needs of information retrieval of the jurist who faces a legal query consists in the selection of all the relevant legislation to this case, the largest possible number of opinions from experts and the largest possible number of significant case law on domain» (Rescigno).

[3] http://nir.ittig.cnr.it/dogi.

Research Council (ITTIG) produces, manages and distributes the database since 1970, the year of its foundation. The corpus includes approximately 250 print and electronic legal journals published by Italian publishers. Here articles; case notes, reviews / comments of legislation and case law, reports on conferences and critical reviews are analytically indexed. This large information resource including approximately 400,000 records consists of bibliographic references of each contribution; abstracts listing the topics discussed by the author, one or more classification codes that identify the legal topics, references of the main sources such as legislation, regulations and case law cited by the author of the contribution with a link to the full text of the legal source[4] The value of the database compared to other available information resources provided by commercial publishers, but also produced on the basis of voluntary institutional initiatives,[5] consists in the fact that DoGi resource provides access to literature not only through bibliographic references, but also through references to legislation and case law cited in the text. The richness and quality of the available information offer new opportunities to create relationships between entities and resources. Giving access to legislation and case law cited in the article, as well as providing the user with a very analytical legal classification, lawyers, as well as ordinary citizens have the opportunity to obtain an overview of legal issues and have access to complete documentation to resolve legal cases of day life. This allows users to know, understand and use data for specific purposes, both professional and academic, or just for simple interest. In this context, linked data technologies can contribute to the development, application and sharing of DoGi records on the web. Following the increasing adoption of advanced web technologies by governments, academia and professionals, the

[4] For futher information see: http://nir.ittig.cnr.it/dogiswish/brochureSwish.htm.

[5] Associazione ESSPER: http://www.biblio.liuc.it/essper/default.asp.

intent is to model the DoGi resources reusing existing descriptive standard schemas such as BiBo, FOAF, DC-Terms and SKOS eventually extended to adapt to the DoGi conceptual structure. The two main objectives of this study are the following:

1. creation of a data model for the DoGi resource as a first work in progress for the legal community in order to lay the ground work for new forms of integration of bibliographic data, legislation and case law available on the web;

2. conversion of the DoGi classification scheme into SKOS and its alignment with other thesauri (such as the BNCF ”Nuovo Soggettario”, Eurovoc). This involves: *a)* enrichment of the collections through the use of controlled vocabularies; *b)* retrieval and integrated search of heterogeneous collections through the use of vocabularies; *c)* analysis of specific topics by navigating through controlled vocabularies; *d)* cross-collection and cross language retrieval, *e)* link between domains, disciplines, different communities.

Definition of the DoGi Data model

As already mentioned, due to its richness in connections to heterogeneous possibly distributed sources, the DoGi database is especially suited to a conversion to linked data formats. One of the peculiarities of the DoGi database is in fact that it can be seen halfway between a bibliographic archive (accessed by bibliographic metadata e.g. author, title, journal..) and a legal sources archive (accessed by legal sources through legal metadata e.g. enacting authority, year, number ...) with the doctrine article acting as a connection hub among legal subjects and the legal documents commented, discussed or supporting the authors argumentation throughout the paper.

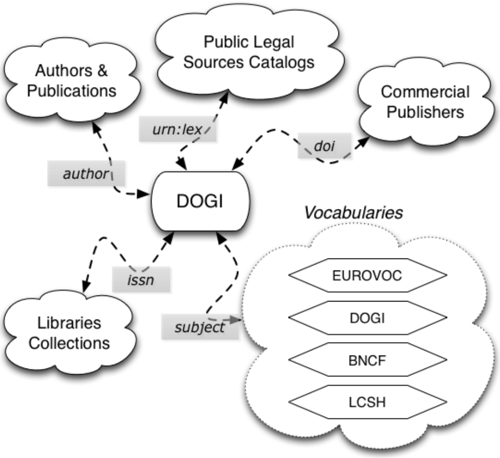


Figure 1: Potential connection of the DoGi dataset in the linked data cloud.

In figure 1 sketches the potential enrichment of the existing DoGi dataset, both in content and access possibilities, made possible by the integration with data published by third parties. In a symmetric way DoGi data would contribute to enrich related data collections. The work carried on consisted in the first place in identifying the standard vocabularies and domain specializations able to model the entities in the DoGi database. Based on the identified models and on the definition of a unified integrated data model, we have been able to define the data mappings necessary to open up to RDF format conforming to the model an excerpt of the DoGi dataset (the bibliographic records, authors and referred legal sources) and of the DoGi classification scheme used for its content categorization in legal subjects. Data were originally made available as XML dumps from the database. Before that, first class resources in the dataset should be assigned globally unique identifiers to be referenced in the linked data space by defining namespace prefixes to prepend to database entities IDs; In the conversion from the database we used the following namespaces as prefixes to corresponding resources identifiers:

```
http://www.ittig.cnr.it/dogi/periodical/
http://www.ittig.cnr.it/dogi/issue/
http://www.ittig.cnr.it/dogi/record/
http://www.ittig.cnr.it/dogi/legalsource/
http://www.ittig.cnr.it/dogi/person/
http://www.ittig.cnr.it/dogi/agent/
http://www.ittig.cnr.it/dogi/descriptor/
http://www.ittig.cnr.it/dogi/term/
```

Modeling DoGi Bibliographic records

As regards bibliographic informations, standard vocabularies can be effectively reused for DoGi data to a large extent. Dublin Core Terms metadata vocabulary, BibO the bibliographic ontology, Foaf friend of a friend ontology have been used as sketched in figure 2 on the next page to represent bibliographic records, bibliographic resources and authors respectively. A mapping of the DoGi records fields (left column) to Dublin Core metadata and bibliographic ontology properties (right column) as used for the conversion of data, is reported in the following tables. The mapping adopted the three distinct hierarchical levels: journal, journal issue and individual article level.

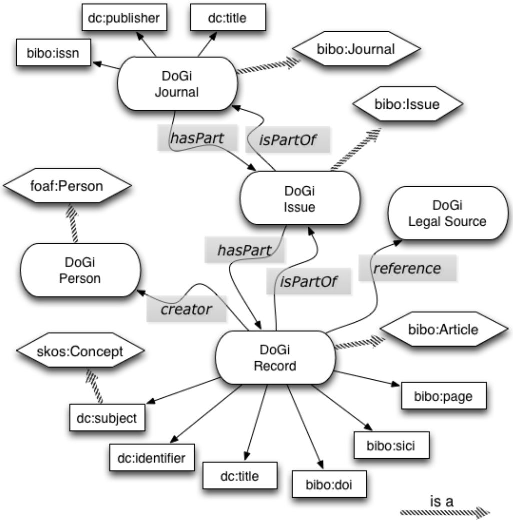


Figure 2: Modeling DoGi metadata reusing DCTerms, BibO, Foaf.

record ID	dc:identifier
	bibo:doi
	bibo:sici
Title	dc:title
Author	dc:creator
Editor	dc:contributor
Language of the original contribution	dc:language
Classification	dc:subject
Outline	dcterms:tableOfContents
Summary	dcterms:abstract
Source of the article *(e.g. Decision Court of ..., International conference on..., etc.)*	dc:source
Appeared in	dcterms:isPartOf
Pages	bibo:pageStart, bibo:pageEnd
Referenced Legal sources (legislation, case law)	dcterms:references

Table 1: DoGi Record / Articolo.

issue ID	dc:identifier bibo:doi
Year of publication	dcterms:issued
Issue number	bibo:issue
Volume number	bibo:volume
Editor	dc:contributor
In Journal	dcterms:isPartOf
Contains article	dcterms:hasPart

Table 2: DoGi Issue

Journal ID	dc:identifier
ISSN code	bibo:issn
Journal title	dc:title
Description	bibo:shortDescription
Publisher	dc:publisher
Director	bibo:director
Contains issue	dcterms:hasPart

Table 3: DoGi Journal

The possible kind of doctrinal contributions catalogued in DoGi have been modelled as specializations of generic bibliographic resources in the dogi namespace (v. figure 3 on page 238).

Listing 1: Different types of doctrinal contribution examined in the DoGi archive.

```
  <!-- JOURNAL -->
<rdf:Description rdf:about=’’http://www.ittig.cnr.it/dogi/
    periodical/1486’’>
  <rdf:type rdf:resource=’’bibo:Journal’’ />
  <bibo:issn>urn:issn:1591-5611</bibo:issn>
  <dc:title>Diritto penale e processo</dc:title>
  <bibo:shortDescription>Mensile di giurisprudenza,
      legislazione e dottrina</bibo:shortDescription>
  <dc:publisher rdf:resource=’’http://www.ittig.cnr.it/dogi/
      organization/ipsoa’’>
</rdf:Description>

  <!-- ISSUE -->
<rdf:Description rdf:about=’’http://www.ittig.cnr.it/dogi/
    issue/1486-13-1’’>
  <rdf:type rdf:resource=’’bibo:Issue’’ />
  <dcterms:identifier>1486-13-1</dcterms:identifier>
  <dcterms:isPartOf rdf:resource=’’http://www.ittig.cnr.it/
      dogi/periodical/1486’’/>
  <dcterms:hasPart rdf:about=’’http://www.ittig.cnr.it/dogi/
      record/2012-G0162’’/>
  <dcterms:issued>2012</dcterms:issued>
  <bibo:volume>13</bibo:volume>
  <bibo:issue>1</bibo:issue>
</rdf:Description>

 <!-- DoGi RECORD -->
<rdf:Description rdf:about=’’http://www.ittig.cnr.it/dogi/
    record/2012-G0162’’>
  <rdf:type rdf:resource=’’bibo:Article’’ />
  <rdf:type rdf:resource=’’dogi:IndependentContribution’’ />
  <dc:identifier rdf:resource=’’http://www.ittig.cnr.it/dogi/
      record/2012-G0162’’/>
  <bibo:doi>10.3280/DUDI2012-001001</bibo:doi>
  <bibo:sici>1591-5611(201201)1:1%69:MEDP%2.0.TX;2-4</bibo:
      sici>
  <dc:creator rdf:resource=’’http://www.ittig.cnr.it/dogi/
      persons/BartoliRoberto’’/>
  <dc:title>Mobbing e diritto penale</dc:title>
  <dc:type>Contributo dottrinale indipendente</dc:type>
  <dc:type rdf:resource=’’http://purl.org/dc/dcmitype/Text’’/>
  <dcterms:language>it<dcterms:language>
  <dcterms:isPartOf rdf:resource=’’http://www.ittig.cnr.it/
      dogi/issue/1486-13-1’’/>
  <bibo:pageStart>85</bibo:pageStart>
  <bibo:pageEnd>94</bibo:pageEnd>
  <dcterms:description>Bibliografia: a pie’ di pagina o nel
      corpo del testo</dcterms:description>
   <dcterms:abstract>Il fenomeno del mobbing risulta molto più
        complesso di quanto si possa credere, verificandosi
       all’interno di dinamiche relazionali senza dubbio
       tradizionali [...] fatti offensivi.</dcterms:abstract>
  <dcterms:tableOfContents>Mobbing e reati relazionali, tra
```

```
        passato e modernità. I modelli classici di tutela penale
         a contrasto del mobbing.</dcterms:tableOfContents>
    <!--. classification -->
    <dc:subject rdf:resource=’’http://www.ittig.cnr.it/dogi#
        s0920’’/>
    <!-- TUTELA DELLA DIGNITA DEL LAVORATORE -->
    <dc:subject rdf:resource=’’http://www.ittig.cnr.it/dogi#
        d4764’’/>
    <!-- Tutela della personalita morale del lavoratore (mobbing
        ) -->
    <!--.Legal sources -->
    <dcterms:references rdf:resource=’’http://www.ittig.cnr.it/
        dogi/legalsource/22367902’’/>
    <dcterms:references rdf:resource=’’http://www.ittig.cnr.it/
        dogi/legalsource/35366803’’/>
     <...../>
  </rdf:Description>

   <!-- DoGi AUTHOR / ORGANIZATION -->
  <rdf:Description rdf:about=’’http://www.ittig.cnr.it/dogi/
      persons/BartoliRoberto’’>
    <rdf:type rdf:resource=’’foaf:Person’’/>
    <foaf:givenname>Roberto<foaf:givenname>
    <foaf:family\_name>Bartoli</foaf:family\_name>
    <foaf:mbox>xxx@yy.com<foaf:mbox>
    <foaf:phone/>
    <foaf:topicInterest/>
    <foaf:depiction/>
    <foaf:homepage/>
    <foaf:weblog/>
    <foaf:publications/>
    <dogi:hasCreated rdf:resource=’’http://www.ittig.cnr.it/dogi
        /record/2012-G0162’’/>
  </rdf:Description>
```

```
<rdf:Description rdf:about=’’http://www.ittig.cnr.it/dogi/
     organization/ipsoa’’>
  <rdf:type rdf:resource=’’foaf:Organization’’/>
  <foaf:name>IPSOA<foaf:givenname>
  <address:localityName>Rome</address:localityName>
  <foaf:mbox>zzz@kk.com<foaf:mbox>
  <foaf:homepage/>
  <foaf:phone/>
  <foaf:depiction/>
  <dogi:hasCreated rdf:resource=’’http://www.ittig.cnr.it/dogi
       /periodical/1486’’>
</rdf:Description>
```

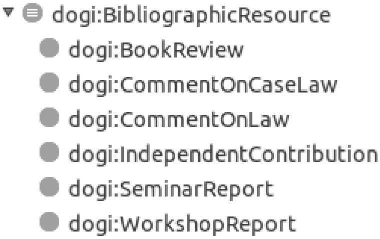
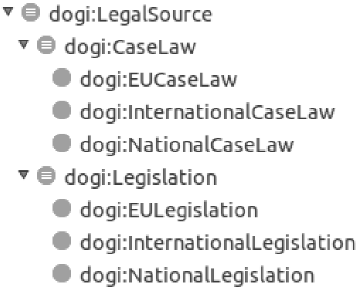


Figure 3: Dogi type of contributions classification.

A multilingual extension of such a representation is straightforward and natively supported by the RDF data model.

Modeling DoGi legal sources

As regards legal sources referred in DoGi records, these are extremely heterogeneous covering both legislation and case law and national, international, European sources. We reproduced this classification introducing the classes reported in figure 4 on the next page in the DoGi annotation schema.



- dogi:LegalSource
  - dogi:CaseLaw
    - dogi:EUCaseLaw
    - dogi:InternationalCaseLaw
    - dogi:NationalCaseLaw
  - dogi:Legislation
    - dogi:EULegislation
    - dogi:InternationalLegislation
    - dogi:NationalLegislation

Figure 4: Dogi legal sources classification.

Legislation

Concerning legislative sources, despite several related initiatives in the legal documentation field, no shared reusable vocabulary for the attribution of legislative metadata covering different legislative sources typology (national, European, International) and for different countries' national legislations, is currently available. However, for the purpose of legislative documents identification the urn:lex schema is increasingly being adopted. urn:lex is a proposed Internet standard for legal document identifiers.[6] The urn:lex namespace aims to facilitate the process of creating URIs for legal sources independent of a document's online availability, location, and access mode. ”Sources of law” include any legal document within the domain of legislation (including bills), case law and administrative acts or regulations. This identifier will be used as a way to represent references (and more generally, any type of relation) among the various sources of law. Being a transparent identifier, urn:lex

[6] http://tools.ietf.org/html/draft-spinosa-urn-lex-06.

can be seen as a serialization of the basic metadata able to describe a legislative act (the country where the legal source has effect, the enacting authority, the type of document, its number, an effective date..). We adopted an extension of this set of basic metadata for the description of referred legislative sources in DoGi documents along with the urn:lex specification for their unique identification (Tab. 4).

Listing 2: descrizione delle risorse legislative riferite nei record DoG

```
<!--cited legislation: art.55 d.lgs. 9 aprile 2008, n.81 -->
<rdf:Description rdf:about=
 ''http://www.ittig.cnr.it/dogi/legalsource/22367902''>
 <rdf:type rdf:resource=
 ''http://www.ittig.cnr.it/dogi/NationalLegislation''/>
 <dcterms:identifier rdf:resource=
 ''urn:nir:stato:decreto.legislativo:2008-04-09:81#art55''>
 <dcterms:isPartOf rdf:resource=
 ''urn:nir:stato:decreto.legislativo:2008-04-09:81''>
 <dc:type>text</dc:type>
 <dc:title>Sanzioni per il datore di lavoro
  e il dirigente
 </dc:title>
 <dcterms:coverage>it</dcterms:coverage>
 <dcterms:language>IT</dcterms:language>
 <dogi:enactingAuthority>Stato</dogi:enactingAuthority>
 <dogi:legislativeDocumentType>Decreto
  Legislativo
 </dogi:legislativeDocumentType>
 <dcterms:issued>2008-04-09</dcterms:issued>
 <dogi:publicationDate>2008-04-30</dogi:publicationDate>
 <dcterms:publisher>Gazzetta Ufficiale
 </dcterms:publisher>
 <dogi:textualCitation>art. 55 d.lgs. 9 aprile 2008,
   n. 81</dogi:textualCitation>
 <dcterms:subject rdf:resource=
   ''http://www.ittig.cnr.it/dogi#s0920''/>
   <dogi:text_url>
   http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:
   decreto.legislativo:2008-04-09;81~art55
   </dogi:text_url>
   <dogi:isReferencedBy rdf:resource=
   ''http://www.ittig.cnr.it/dogi/record/2012-G0162''/>
  </rdf:Description>
```

Legislative document ID	dc:identifier (urn:lex Schema)
Document title	dc:title
Territorial jurisdiction	dc:coverage
Document type	dogi:legislativeDocumentType
Enacting Authority	dogi:enactingAuthority
Date of enactment	dcterms:issued
Publication Date (Efficacy)	dogi:publicationDate
Publisher	dcterms:publisher
Textual Citation	dogi:textualCitation
Classification	dcterms:subject
Document URL	dogi:text_url
Reference to partition container (e.g. Article part of Section)	dcterms:isPartOf
References to other resources	dcterms:references
Resources referring the legislative source (e.g. a DoGi record)	dogi:isReferencedBy

Table 4: dogi:Legislation

Notice that, in a linked open data environment, the national publication offices should be themselves responsible for the open publication of national legislation catalogs in open formats along with the associated RDF metadata as open data sets in the linked data cloud. In such a foreseeable scenario the urn:lex identifier would be by itself sufficient to establish a machine readable connection to the official legislative metadata catalogs, automatically augmenting the DoGI dataset with the official descriptive metadata of the legal sources referred in DoGi doctrinal articles. In this respect the most advanced initiative of open legislation publishing is so far the UK legislation portal[7] part of the British open government data strategy[8] where all the best practices of XML publication of legislative documents along with associated RDF metadata have been put in practice.

In legislation.gov.uk a more sophisticated metadata model – incorporating FRBR,[9] the CEN MetaLex vocabulary, [10] Dublin Core Terms,[11] and the Crown Legislation Markup Language – enabling advanced version control and output of descriptive metadata have been adopted providing also all the metadata able to implement a point in time legislative system.
A rapid change of the scenario in this direction can be foreseen in the near future both at the national and European level under the impulse of the open government data movement and of revised discipline in favour of the right of access to public sector information. A significant example of this trend is the experimentation currently carried on in the EU LATC Project of transition to linked open data of the whole Eur-Lex service of European Union Law along with its connection to national implementing legislation.[12]

Case law

Similar considerations hold for the identification of a set of basic metadata able to identify and describe case law documents. In this case we based on the recommendation establishing Council of Europe conclusions inviting the introduction of the European Case Law Identifier (ECLI) along with minimum set of uniform metadata

[7] www.legislation.gov.uk.

[8] http://www.data.gov.uk.

[9] http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records.

[10] http://www.metalex.eu.

[11] http://www.dublincore.org/documents/dcmi-terms.

[12] http://eur-lex.publicdata.eu.

for case law. The following is an RDF serialization of the metadata related to a Jurisprudential source from the DoGi archive according to the annotation defined in Tab. 5.[13]

```
<!--cited decision: Cass. sez. VI pen. 13 gennaio 2011, n. 685
     -->
<rdf:Description rdf:about= ’’http://www.ittig.cnr.it/dogi/
    legalsource/35366803’’>
  <rdf:type rdf:resource= ’’http://www.ittig.cnr.it/dogi/
      NationalCaseLaw’’/>
  <dcterms:identifier rdf:resource= ’’urn:lex:it:corte.
      cassazione:sentenza:2011-01-13;685’’/>
  <dcterms:identifier rdf:resource=’’ECLI:IT:CCA
      :2011-01-13:685’’/>
  <dcterms:creator>Corte di Cassazione sezione VI penale</
      dcterms:creator>
  <dogi:court>Corte di Cassazione</dogi:court>
  <dogi:courtSection>VI</dogi:courtSection>
  <dogi:judicialDivision>penale</dogi:judicialDivision>
  <dc:type>judicial decision</dc:type>
  <dc:title/>
  <dcterms:coverage>it</dcterms:coverage>
  <dcterms:language>IT</dcterms:language>
  <dcterms:date>2011-01-13</dcterms:date>
  <dcterms:issued/>
  <dcterms:publisher/>
  <dcterms:contributor>Serpico</dcterms:contributor>
  <dcterms:contributor>Milo</dcterms:contributor>
  <dogi:textualCitation>
    Cass. sez. VI pen. 13 gennaio 2011, n. 685
  </dogi:textualCitation>
  <dc:subject rdf:resource=’’http://www.ittig.cnr.it/dogi#
```

[13] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2011:127:0001:0007:IT:PDF.

```
        d4764’’/>
    <dcterms:abstract>
      Il mobbing èsolo vagamente assimilabile alla previsione di
           cui all’art. 572 c.p....
    </dcterms:abstract>
    <dogi:text_url/>
    <dogi:isReferencedBy rdf:resource= ’’http://www.ittig.cnr.it
        /dogi/record/2012-G0162’’/>
  </rdf:Description>
```

Case law ID	dcterms:identifier (ECLI)
	dcterms:identifier (urn:lex)
The full name of the court	dcterms:creator
Name of the court	dogi:court
Section (e.g. number)	dogi:courtSection
Judicial division (e.g. criminal, civil)	dogi:judicialDivision
The country in which the court or tribunal is seated	dc:coverage
The date of the decision	dcterms: date
Language of the instance document	dcterms:language
Document type (e.g. Judicial Decision)	dcterms:type
Title (Preferably the name of the parties or an alias according to national practice)	dcterms:title
Date of deposit	dcterms:issued
The (commercial or public) organization responsible for the publication	dcterms:publisher
Names of judges, Advocate general or other staff involved	dcterms:contributor
Classification (field of law)	dcterms:subject
Abstract or summary of the case	dcterms:abstract
Textual Citation	dogi:textualCitation
Document URL	dogi:text_url
References to other resources	dcterms:references
Resources referring the judicial decision (e.g. a DoGi record)	dogi:isReferencedBy

Table 5: dogi:CaseLaw

Similar remarks made for legislation hold for publication of European and national case law in public portals: the European e-Justice portal and its national sections should be in the near future responsible for the publication, conforming to the aforementioned standards, of judicial decisions in Europe.

Opening DoGi classification scheme

Even more important in order to share with external data providers the conceptual organization of the legal doctrine materials stored in DoGi is the opening of its classification scheme. DoGi classifies its resources according to a consolidated classification scheme covering all areas of national (Italian) law and international law. It is divided into 24 micro thesauri covering general areas of law, each structured in a three level hierarchy of systematic descriptors. Additionally a flat list of analytical terms related to systematic descriptors is provided for further specification of the articles classification. SKOS (the Simple Knowledge Organization Systems) is a W3C Recommendation[14] that provides a means for representing knowledge organization systems (including controlled vocabularies, thesauri, taxonomies and folksonomies) in a distributed and linkable way. From the XML dump of the conceptual scheme stored on the DoGi database, containing for each descriptor its alphanumeric identifier and its lexicalization, we provided a conversion to SKOS preserving the original structure and content. The descriptors have been assigned a globally unique identifier by prefixing their original id with the DoGi prefixes

`http://www.ittig.cnr.it/dogi/descriptor/` for systematic descriptors

`http://www.ittig.cnr.it/dogi/term/` for analytical terms

obtaining for example the following

Listing 3: SKOS/RDF representation for a DoGi description.

```
<skos:Concept rdf:about=''http://www.ittig.cnr.it/dogi/
    descriptor#s1018''>
```

[14] http://www.w3.org/TR/skos-reference

```
   <rdf:type rdf:resource=''http://www.w3.org/2002/07/owl#
       NamedIndividual''/>
   <skos:prefLabel xml:lang=''it''>Responsabilita penale}</skos:
       prefLabel>
   <skos:broader rdf:resource=''http://www.ittig.cnr.it/dogi/
       descriptor#c0014''/>
   <skos:related rdf:resource=''http://www.ittig.cnr.it/dogi/
       term#d4977''/>
 <!-- mapping to BNCF Nuovo Soggettario -->
   <skos:exactMatch rdf:resource=''http://purl.org/bncf/tid
       /12679''/>
 </skos:Concept>
```

The SKOS representation allows DoGi descriptors to be referred by external sources on the web as well as to provide translation in different languages and semantic mapping to external vocabularies. As a first step the correspondences provided by an ongoing joint effort with BNCF (National Library of Florence) of mapping legal descriptors of the ”Nuovo Soggettario” with DoGi descriptors, have been integrated in the SKOS version of the concept scheme using skos:semanticRelation (exactMatch, broaderMatch, narrowerMatch) properties. This lays the foundations for example to use the DoGi vocabulary published on the Linked data space as a bridge for multilingual access, programmatic access from third parties archives to DoGi data by legal subject and fetching resources from external archives linked with mapped descriptors.
Part of the experimentation consisted in loading the SKOS version of the DoGi vocabulary in the open source platform for collaborative multilingual thesaurus editing LegalVocbench. LegalVocbench is based on the web platform Vocbench,[15] originally developed by FAO for the management of the agricultural thesaurus AGROVOC and

[15] http://aims.fao.org/tools/vocbench-2.

recently specialized by ITTIG for the management of legal vocabularies. This will eventually allow further enrichment of the DoGI vocabulary in its content (e.g. multilingual labelling, adding legal definition), structure (e.g. adding semantic relations among descriptors to further enhance of ″related material″ retrieval) and alignment to other legal or general purpose value vocabularies available on the web.

Conclusions

The interoperability of DoGi records on the web is a first step to reach different communities (institutional data providers such as governments, libraries, commercial publishers), but of course this evolution involves an effort of standardization to be shared among various communities. This work has been carried out starting from legal literature with the intent to make legal documentation stakeholders aware of this need, identifying metadata models and schemas, standards and protocols for interoperable information systems and supporting the creation of new forms of *alliances*. This evolution can contribute to a strong development of legal information systems on the web, encourage the emergence of digital content and thus supporting advanced shareable searching of legal data. Furthermore, the adoption of the Library Linked Data approach in the field of legal doctrine offers a unique opportunity to provide stability and persistence of these information resources regardless of format and software tools used for their production, releasing institutions from dependence on a limited number of technical providers.

References

Kirchberger, Christine. “The 'I' in Legal Information Retrieval”. *Vem reglerar informationssamh́allet?* Stockholm: Jure AB, 2010. http://ssrn.com/abstract=1957558. (Cit. on p. 228).

Rescigno, Ugo. “Il giurista come scienziato”. *Rivista di Diritto Pubblico* 3. (2003): 833–864. (Cit. on p. 229).

TOMMASO AGNOLONI, Istituto di Teoria e Tecniche dell'Informazione Giuridica del CNR (ITTIG), tommaso.agnoloni@ittig.cnr.it
ELISABETTA MARINAI, Istituto di Teoria e Tecniche dell'Informazione Giuridica del CNR (ITTIG), elisabetta.marinai@ittig.cnr.it
GINEVRA PERUGINELLI, Istituto di Teoria e Tecniche dell'Informazione Giuridica del CNR (ITTIG), ginevra.peruginelli@ittig.cnr.it
MARIA TERESA SAGRI, Istituto di Teoria e Tecniche dell'Informazione Giuridica del CNR (ITTIG), m.t.sagri@ittig.cnr.it
DANIELA TISCORNIA, Istituto di Teoria e Tecniche dell'Informazione Giuridica del CNR (ITTIG), daniela.tiscornia@ittig.cnr.it

Agnoloni, T., E. Marinai, G. Peruginelli, et al. ”Annotation schema for legal doctrine: a case study on DoGi database”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art. #5480, p. 1–249. DOI: 10.4403/jlis.it-5480. Web.

ABSTRACT: Interoperability today is the key term for the enhancement of databases published on the web: the data, when isolated, have little value, on the contrary, their value increases significantly when different datasets, produced and published independently by different providers, can be reused and freely mashed by third parties. The use of data for new purposes not foreseen by organizations and individuals who publish “raw data” is one of the advantages of linked open data model. To achieve these benefits content and relationships between the entities described in the dataset should be explicitly represented in standard web formats (XML, RDF, URI). The DoGi-Legal Literature database, one of the most valuable sources for online access to legal doctrine, created and managed by the Institute of Legal Information Theory and Techniques of the CNR is following this direction. This paper will define the schema of the data representing the database in RDF format. This will make the DoGi database interoperable between different data and service providers (libraries, publishers, information services for accessing national and European legal information), allowing the creation of new advanced services to be made available on the web of data. In particular, the contribution will focus on the goal to promote semantic interoperability between the DoGi classification schema and other semantic indexing tools in legal domain.

KEYWORDS: Library linked data; DoGi-Legal Literature database
ACKNOWLEDGMENT: Maria Teresa Sagri, Ginevra Peruginelli, Daniela Tiscornia and Elisabetta Marinai are authors of paragraphs at pages 1–5; Tommaso Agnoloni is author of paragraphs on pages 5–21; Ginevra Peruginelli is author of conclusions on page 21.

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

Towards a Web of Data: Applications and Experiences

Semantic technologies and linked data, with a case study at the Consiglio Nazionale delle Ricerche (CNR)

Aldo Gangemi

Web evolution and resource dereferencing

Web is evolving from a web of documents (often called Web 1.0) to a web of entities (called, with subtle differences in meaning, semantic web, web of data, Web 3.0). This evolution is passing also through the availability to users to edit its contents and generate complex social networks through simple interaction paradigms (known as social web or Web 2.0). This is happening primarily thanks to a deeper exploitation of the Web architecture designed since the nineties,[1] which enables the dereferencing and linking of web resources (identified by means of a Web address), through simple communication protocols (e.g. HTTP). For example, when one writes the address (URI) http://www.cnr.it (the web address of the portal of Consiglio Nazionale delle Ricerche, CNR) in a browser, the HTTP client of the browser dereferences that address by communicating with a server at CNR, which returns a HTML page, visualized on its turn

[1] http://www.w3.org/TR/webarch.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5457

by the browser. Other web pages can be present in the visualized page, so creating a network of hypertextual links, which enables the browsing experience. This is basically the web of documents. Sometimes dereferencing is indirect, as in the case when an address represents a call to a database, e.g. when looking for one's tax data in the Agenzia delle Entrate (the Italian fiscal authority) web site: this is still the web of documents, but the documents are generated out of a query to a database, whose answer is rendered in HTML by using XML stylesheets. The case of Web 2.0 is a more sophisticated indirect dereferencing, which also enables direct changes to a database performed by users: applications such as voice protocols, email, tagging, automatic log analysis, opinion mining etc. converge in rich, customizable and dynamic HTML pages, as in the case of Facebook.

Two difficult problems: identity and semantic interoperability

Web 1.0 and 2.0 have two limitations, which actually exist in information systems since centuries ago: identity and semantic interoperability. The identity issue arises e.g. in the following example. Aldo Gangemi has different homepages (one on his institute site, ISTC-CNR, one from the wiki of his lab, STLab, one on the semanticweb.org site etc. He is also registered on many other portals providing services to the citizen, to members of associations, conference committees, commercial services etc. Moreover, he has several accounts of social web applications (e.g. Facebook, Gmail, Flickr, iTunes etc.). Even more, Aldo Gangemi is a datum within public or personal databases, like Google Scholar, DBLP etc.; that datum has identifiers that are owned specifically by those databases, gathering then a sort of ”positional” identity within one of their tables). Finally, Aldo Gangemi is cited in other web pages: articles, bibliographic references, event reports. Now the issue is: how can we know that (the physical or social person) Aldo Gangemi is the entity denoted by his homepages, registrations, accounts, database IDs, citations? Intuitively, the issue is not limited to persons, but it impacts on everything that has an identity: places, organizations, products, services, events, laws, ideas, concepts, fictional things etc. The semantic interoperability issue, besides purely system-oriented problems (e.g. different computational platforms), arises from the identity issue: if we cannot resolve the identity of something across the different sources and systems that refer to it, it gets really difficult to aggregate (i.e. assemble) and integrate (i.e. appropriately connect) the information about it. This is quite limiting when considering that the relations between something and something else can be similar within different systems: the relation between Aldo Gangemi and the email messages addressed to him, or between him and his recipients, are similar in any emailing system, but those systems assign different identities to the same persons, if any. In addition, each system works on a proprietary infrastructure: different languages, formats, protocols etc. All this makes data integration between different systems partial in the best cases.

Some traditional solutions

In the last years, a sort of cartel has emerged between commercial service providers such as Facebook, Google etc., in order to make social network data interoperable: this however concerns only data exchange that are commercially interesting for those systems, and third party applications that count on them. Database management systems use complex procedures to integrate their data when it is required: schema integration, identity resolution, data warehousing etc. Each process is typically made ad hoc on a pair of databases. Partial solutions for data integration also come from data mining or natural language processing techniques. For example, there are effective statistical approaches for named entity recognition and resolution, as well as for discovering similarity and indirect relations in data. Document annotation is an approach that comes back at least to the beginning of 20th century: a document, or part of it (paragraphs, terms) are annotated with a category or tag taken from some knowledge organization system: thesauri, classification schemes, nomenclatures, controlled vocabularies, which have developed in most scientific, library, and commercial disciplines. Exemplar cases of similar large efforts include SnoMed, ICD, MeSH (medicine), Getty thesaurus (cultural heritage), Agrovoc (agriculture) etc. Recently, annotation procedures are assisted either by computational support for manual annotation, or by automatic annotation algorithms (e.g. text classification), with variable precision.

The web of data

In 2006, Tim Berners-Lee introduced linked data, a simple and elegant method[2] to realize some practical data identity integration and interoperability on the Web. Linked data are aimed at realizing a web of data (or Entities, or Things, depending on the interest to data management, to entity linking, or to sensors and things in the physical world). Linked data is one of the technologies for the semantic web (discussed in the next section), and consists of four principles and many good practices. The principles include:

1. use web addresses (URI) as names for entities/things;

[2] http://www.w3.org/DesignIssues/LinkedData.html.

2. use HTTP URIs so that people can look up and dereference those names;

3. when someone looks up a URI, provide useful information, using the standards (RDF, RDFS, SPARQL, OWL, RIF);

4. include links to other URIs, in order to be able to discover more things and data.

Among good practices, it's useful to mention those that have best supported the Linking Open Data (LOD) bootstrap, whose state of play is visualized periodically as a cloud[3]:

- use open licenses to obtain highly reusable data;
- use non-proprietary formats (e.g. CSV instead of Excel);
- use W3C open standards (typically RDF,[4] SPARQL,[5] OWL[6]) to identify things, so that people can point at your stuff, new links can be created, better queries and more extended reasoning can be performed.

These practices also fit recommendations from the Open Data movement, and are currently adopted in many different fields, including Public Administration data[7] and are used in the integration and enrichment of data, for example for the expert finding task.[8]
The LOD Cloud contains linked data from many different domains, in particular biomedicine, cultural, multimedia, bibliographic, geographic etc. An example of the potential of linked data is shown in

[3] http://linkeddata.org.
[4] http://www.w3.org/RDF/.
[5] http://www.w3.org/TR/rdf-sparql-query/.
[6] http://www.w3.org/2004/OWL/.
[7] http://data.gov; http://data.gov.uk; http://dati.gov.it.
[8] http://data.cnr.it.

figure 1, a graph built automatically by an application (RelFinder[9]), which incrementally visualizes the relations between any two entities, provided that they have an identity on the web of data. In the figure 1, graph building starts from the entities:

```
<http://dbpedia.org/wiki/Neo-positivism>
<http://dbpedia.org/wiki/Francis_Bacon>
```

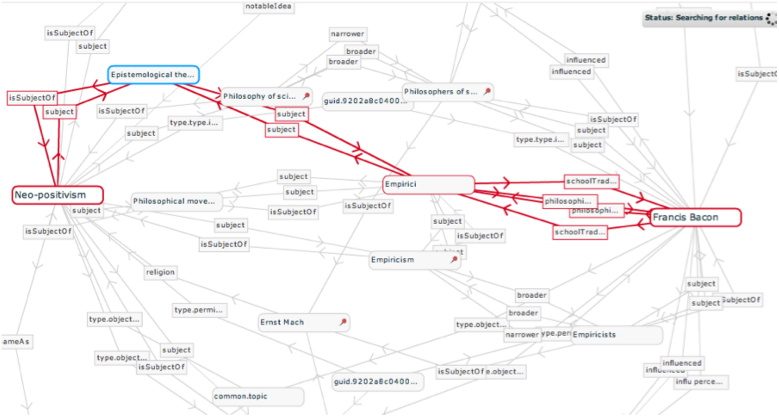


Figure 1: The emerging relations between two entities across the Linking Open Data graph.

Semantic web standards

W3C open standards, primarily RDF,[10] SPARQL[11] and OWL,[12] enable elegant and homogeneous representation of, as well as querying

[9] http://www.visualdataweb.org/relfinder.php.
[10] http://www.w3.org/RDF/.
[11] http://www.w3.org/TR/rdf-sparql-query/.
[12] http://www.w3.org/2004/OWL/.

and reasoning on, the data from most traditional data structures and data models.
RDF is based on a recursive data structure, called triple, made of a Subject, a Predicate, and an Object, analogously to the most abstract grammatical structure of Western languages, SVO (Subject-Verb-Object).

Listing 1: Sample RDF triples.

```
<http://www.cnr.it/ontology/cnr/individuo/
    unitaDiPersonaleInterno/MATRICOLA1582>
  <http://www.w3.org/2000/01/rdf-schema#label>
  ''Aldo Gangemi''

<http://www.cnr.it/ontology/cnr/individuo/
    unitaDiPersonaleInterno/MATRICOLA1582>
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
  <http://www.cnr.it/ontology/cnr/personale.owl#
      UnitaDiPersonaleInterno>

<http://www.cnr.it/ontology/cnr/individuo/
    unitaDiPersonaleInterno/MATRICOLA1582>
  <http://purl.org/dc/terms/subject>
  <http://dbpedia.org/resource/Category:Semantic_Web>
```

RDF triples can be queried via the SPARQL language.

Listing 2: Query on triples in 1

```
SELECT ?l
WHERE {
?x <http://purl.org/dc/terms/subject>
<http://dbpedia.org/resource/Category:Semantic_Web>.
?x <http://www.w3.org/2000/01/rdf-schema#label> ?l}
```

The query in listing 2 on the preceding page gets the answer:

```
l
''Aldo Gangemi''
```

Each triple contains Subjects and Objects that have a type, which is on its turn a Class, e.g.

```
<http://www.cnr.it/ontology/cnr/personale.owl#
    UnitaDiPersonaleInterno>
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
    <http://www.w3.org/2002/07/owl#Class>
```

Each triple contains a Predicate (or Property), which together with classes forms the vocabulary (also called schema or ontology) used by a dataset. In cases where logical validation and reasoning is required, a vocabulary is defined in the OWL (Ontology Web Language) standard,[13] a language that allows the use of automated reasoners to derive logical inferences out of data structures. For example, an automated reasoner infers the inverses of existing triples, the symmetric triples, the triples holding transitively (when appropriate rules have been defined for the vocabulary) etc.
With the expressive power of OWL and SPARQL on the web of data, one can make complex questions to heterogeneous knowledge sources, e.g. in the Romal Law domain, the following natural language questions can be formalized as queries, but terms need to be mapped to appropriate entity types in RDF and OWL. In this case, underlined terms are supposed to be mapped as classes, bold-faced **terms** as properties, and *terms* in *italics* as specific entities or values:

-which *Roman Law* sources **contain** maxims **concerning** *stipulation*, cite *Ulpian*, and **include** commentaries **published in** *the last 10 years*?

[13] http://www.w3.org/2004/OWL/.

-which cases **appeared in** *Common Law* systems **contain** interpretations **relative to** contracts **analogous to** *stipulatio*?

In order to improve vocabulary quality and inference capabilities, additional axioms need to be defined (e.g. what type of entities can be analogous to what, what can be cited in what etc.). Therefore, vocabulary design requires a certain accuracy and quality control, which can be obtained by means of approaches oriented to user requirements, and with the reuse of standard vocabularies and ontology design patterns,[14] known to describe the domain of interest, and/or solving the modeling problems emerging from user requirements.

Semantic applications

Availability of large open data can provide a good motivation to develop next generation applications, which build on both existing and novel solutions, focused on the semantic paradigm: using meaning of data as a widespread organizational schema.
The life cycle of a semantic application is typically the following:

1. reengineering existing data, by producing datasets in RDF triples (data) and OWL (vocabularies);

2. linking between entities in multiple datasets, and production of new triples ;

3. extraction of new entities and triples by means of data mining and natural language processing techniques, and production of new triples;

[14] http://www.ontologydesignpatterns.org.

4. reasoning on the logical structures obtained from previous steps, and possible production of new triples (materialization);

5. publishing of datasets on appropriate platforms, for SPARQL querying;

6. presentation of enriched data to be used by web users: textual, graphic, rich snippets, explorative etc.

The life cycle reflects a multiple interpretation of the term semantics. In steps 2 and 3, we refer primarily to the linguistic semantics that is implicit in the analyzed texts; related technologies are those of text and data analysis, and aims at recognizing entities, names, terms, relations, facts, topics etc. Only once we have extracted them, we can produce new formal triples. In steps 1, 4, 5, we refer to logical (or formal) semantics of data and schemas; related technology is basically what we have mentioned in previous sections as ″semantic web″ (which is a mix of web science and knowledge representation). In step 6., we refer to the semantics of user interaction.

Technologies oriented to linguistic semantics allow e.g. to recognize entities in texts, and to resolve their identity with respect to known datasets. Once identity has been resolved, it is possible to enrich the dataset with known relations between that entity and other entities. For example, given the following text from the proceedings of European Union Parliament:

> The sensitivities of Northern Ireland are too important for any ill-informed bandwagoning on the International Fund for Ireland. Raytheon has been welcomed to Derry by no less than Nobel Peace Prize winners, John Hume – one of our own colleagues, and David Trimble. Raytheon will be funded by the Industrial Development Board in Northern Ireland. Not one euro nor one Irish pound from the International Fund for Ireland is going to Raytheon.

it's possible to use a ″named entity recognizer″ like http://www.alchemyapi.com/api/demo.html in order to recognize several names (e.g. Northern Ireland, International Fund for Ireland, Derry, John Hume etc.), whose identity can be automatically resolved by an ″entity resolver″ like http://wit.istc.cnr.it/stlab-tools/wikifier as e.g. the entities:

```
<http://dbpedia.org/resource/Northern_Ireland>
<http://dbpedia.org/resource/John_Hume>
<http://dbpedia.org/resource/Derry>
<http://dbpedia.org/resource/David_Trimble>
```

Once identified, we can query LOD to find out other linked entities, e.g.

```
<http://dbpedia.org/wiki/Mark_Durkan>
<http://dbpedia.org/wiki/David_Cameron>
```

A complex graph emerging from LOD triples when the four entities above are searched together for their links can be then retrieved (and e.g. visualized in the RelFinder tool[15]). For example, the following triples are found:

```
<http://dbpedia.org/wiki/Mark_Durkan> <http://dbpedia.org/
    ontology/placeOfBirth> <http://dbpedia.org/resource/Derry>

<http://dbpedia.org/resource/Derry> <http://dbpedia.org/
    ontology/country>
<http://dbpedia.org/resource/Northern_Ireland>
<http://dbpedia.org/wiki/Mark_Durkan> <http://dbpedia.org/
    ontology/predecessor> <http://dbpedia.org/resource/
    John_Hume>
```

[15] http://www.visualdataweb.org/relfinder.php.

```
<http://dbpedia.org/resource/Northern_Ireland> <http://dbpedia.
    org/ontology/leaderName>
<http://dbpedia.org/resource/Derry>
```

Figure 2 summarizes this kind of simple process: linguistic knowledge can be used to enrich (and give access to) formal knowledge, while the latter, besides generating the implicit knowledge that is implicit in triples (deductive inferences), can also be used as background knowledge by the algorithms that extract new linguistic knowledge (as applied in the entity resolution task).

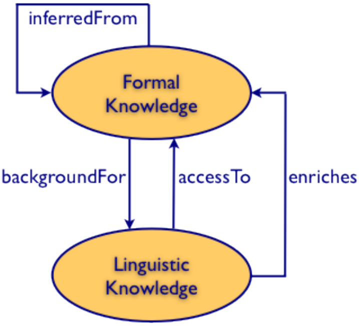


Figure 2: The hybridization cycle of linguistic and logical techniques.

Linguistic and formal interpretation, as well as the integration of related technologies, enable a hybrid methodology that empowers knowledge structuring and querying. A recent spin to that hybridization can be seen in the deep knowledge extraction task, as implemented in the FRED tool.[16] FRED deeply analyzes sentences, produces formally correct structures in RDF and OWL, and enriches the results with entity resolution. Figure 3 shows a fragment of the RDF graph produced by FRED from the EU Parliament sample

[16] http://wit.istc.cnr.it/stlab-tools/fred.

sentence.

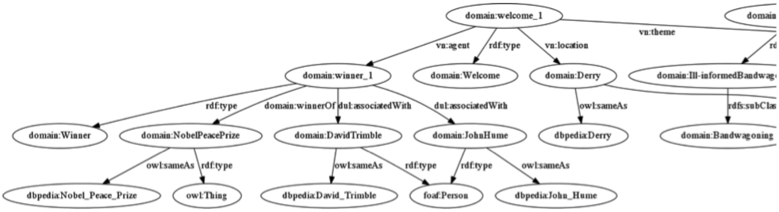


Figure 3: An excerpt from the RDF graph produced by FRED on the EU sample sentence.

A case study at Consiglio Nazionale delle Ricerche (CNR)

A practical application of data designed by following the presented semantic application lifecycle is data.cnr.it, a group of datasets that contain research data from the Italian National Research Council (CNR), enriched by automatic extraction of linguistic knowledge, automatic categorization of person and project profiles, and automated materialization of logical inferences.

The vocabularies for the datasets have been designed as modular ontologies (figure 8 on page 269), which are aligned to reference vocabularies from the LOD Cloud. Part of the taxonomy from the CNR vocabularies is shown in figure 9 on page 269. Where possible, data are linked to public entities, e.g. from DBpedia[17] or GeoNames.[18] Data can be accessed in different ways, depending on who is going to consume them. Information systems will use the data.cnr.it

[17] http://dbpedia.org.

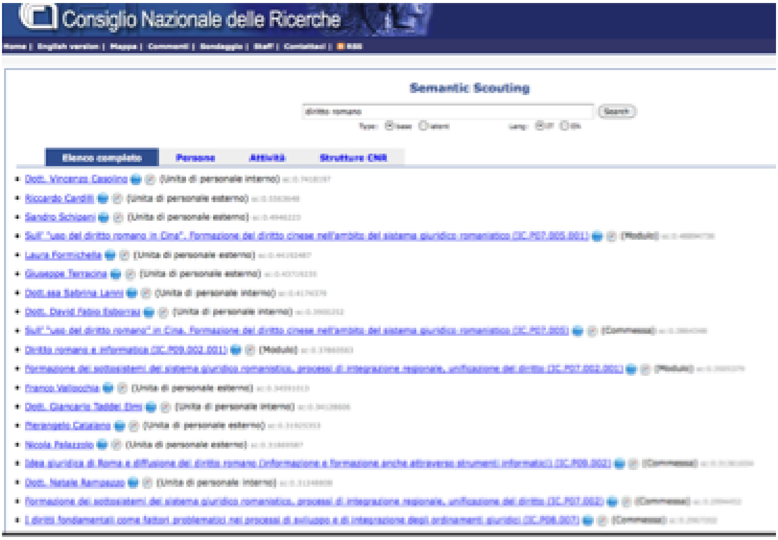
[18] http://www.geonames.org.

SPARQL endpoint to execute queries. Human users have also other ways to search, query, or explore data. We have designed the Semantic Scout,[19] an exploratory browser for human consumption of data: a web application supports the expert finding task on scientific competences existing at CNR, based on how data are represented in data.cnr.it datasets.
Figure 4 on the facing page shows the Semantic Scout keyword search interface for finding who works on Roman Law at CNR. Figure 5 on the next page shows hierarchical browsing in the semantic social network built from knowledge related to people found with the keyword search. Figure 6 on page 268 shows the exploratory search interface to data in a spherical space. Figure 7 on page 268 shows the exporting functionality of results obtained from the choices of the user during the exploratory browsing.

Conclusions

Semantic technologies provide a simple solution to the identity and interoperability issues, exploiting direct access to data, web standards, formal precision of schemas, and easy hybridization between techniques oriented to the extraction of linguistic semantics, and those oriented to the management of formal semantics.

[19] http://bit.ly/semanticscout.



Consiglio Nazionale delle Ricerche
Home | English version | Mappa | Commenti | Sondaggio | Staff | Contattaci | RSS

Semantic Scouting

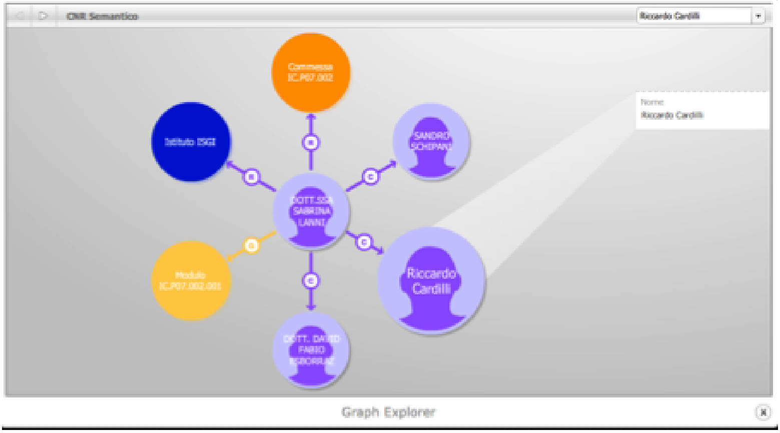
diritto romano

Type: ● base ○ latent   Lang: ● IT ○ EN

Elenco completo
Persone
Attività
Strutture CNR

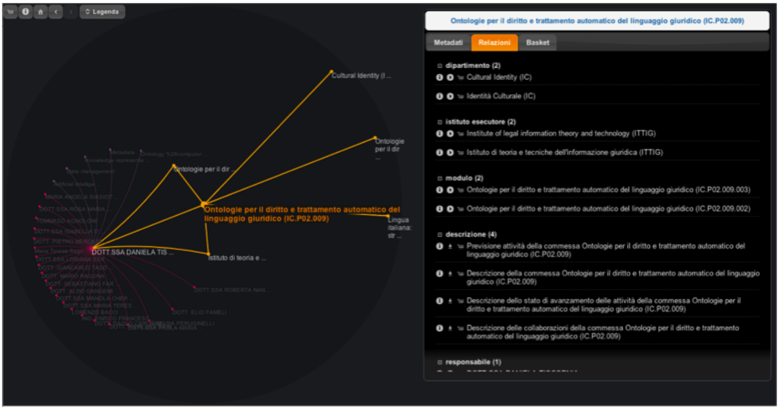
- Dott. Vincenzo Casolino (Unita di personale interno) [illegible]
- Riccardo Cardilli (Unita di personale esterno) [illegible]
- Sandro Schipani (Unita di personale esterno) [illegible]
- Sull' "uso del diritto romano in Cina". Formazione del diritto cinese nell'ambito del sistema giuridico romanistico (IC.P07.005.001) (Modulo) [illegible]
- Laura Formichella (Unita di personale esterno) [illegible]
- Giuseppe Terracina (Unita di personale esterno) [illegible]
- Dott.ssa Sabrina Lanni (Unita di personale interno) [illegible]
- Dott. David Fabio Esborraz (Unita di personale interno) [illegible]
- Sull' "uso del diritto romano" in Cina. Formazione del diritto cinese nell'ambito del sistema giuridico romanistico (IC.P07.005) (Commessa) [illegible]
- Diritto romano e informatica (IC.P09.002.001) (Modulo) [illegible]
- Formazione dei sottosistemi del sistema giuridico romanistico, processi di integrazione regionale, unificazione del diritto (IC.P07.002.001) (Modulo) [illegible]
- Franco Vallocchia (Unita di personale esterno) [illegible]
- Dott. Giancarlo Taddei Elmi (Unita di personale interno) [illegible]
- Pierangelo Catalano (Unita di personale esterno) [illegible]
- Nicola Palazzolo (Unita di personale esterno) [illegible]
- Idea giuridica di Roma e diffusione del diritto romano (informazione e formazione anche attraverso strumenti informatici) (IC.P09.002) (Commessa) [illegible]
- Dott. Natale Rampazzo (Unita di personale interno) [illegible]
- Formazione dei sottosistemi del sistema giuridico romanistico, processi di integrazione regionale, unificazione del diritto (IC.P07.002) (Commessa) [illegible]
- I diritti fondamentali come fattori problematici nei processi di sviluppo e di integrazione degli ordinamenti giuridici (IC.P08.007) (Commessa) [illegible]

Figure 4: Search with the Semantic Scout.



Nome
Riccardo Cardilli

Figure 5: Semantic browsing with the Semantic Scout.



Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009)

dipartimento (2)
- Cultural Identity (IC)
- Identità Culturale (IC)

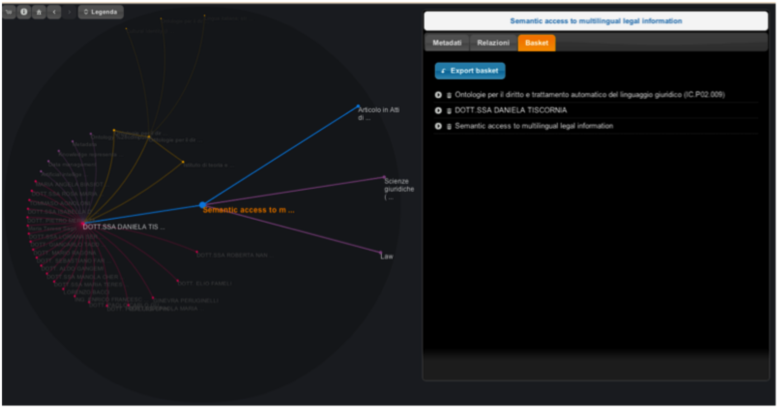
istituto esecutore (2)
- Institute of legal information theory and technology (ITTIG)
- Istituto di teoria e tecniche dell'informazione giuridica (ITTIG)

modulo (2)
- Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009.003)
- Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009.002)

descrizione (4)
- Previsione attività della commessa Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009)
- Descrizione della commessa Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009)
- Descrizione dello stato di avanzamento delle attività della commessa Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009)
- Descrizione delle collaborazioni della commessa Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009)

responsabile (1)

Figure 6: Exploratory search interface of data.cnr.it data in a spherical space.



Semantic access to multilingual legal information

- Ontologie per il diritto e trattamento automatico del linguaggio giuridico (IC.P02.009)
- DOTT.SSA DANIELA TISCORNIA
- Semantic access to multilingual legal information

Figure 7: Exporting results of semantic exploratory search based on user choices.

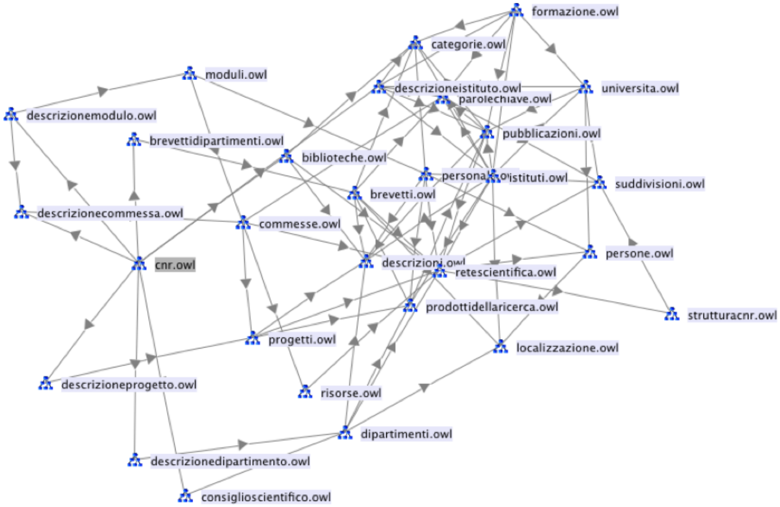


Figure 8: The network of CNR modular ontologies.

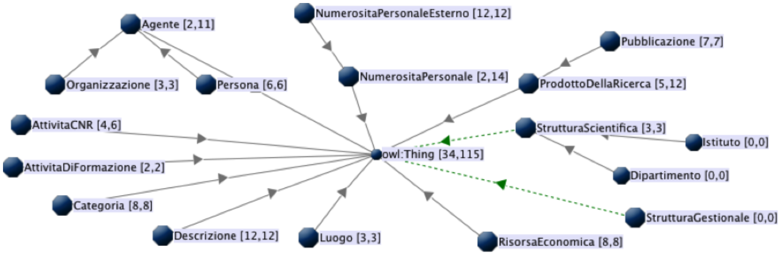


Figure 9: Taxonomy from the core ontology of CNR.

ALDO GANGEMI, Semantic Technology Lab, Institute of Cognitive Sciences and Technologies (CNR-ISTC), Roma, and Université Paris 13 – CNRS –Sorbonne Cité (France).
gangemi@loa-cnr.it

Gangemi, A. ”Semantic technologies and linked data, with a case study at the Consiglio Nazionale delle Ricerche (CNR)”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5457. DOI: 10.4403/jlis.it-5457. Web.

ABSTRACT: Governmental data are being published in many countries, providing an unprecedented opportunity to create innovative services and to increase societal awareness about administration dynamics. In particular, semantic technologies for linked data production and exploitation prove to be ideal for managing identity and interoperability of administrative entities and data. This paper presents the current state of art, and evolution scenarios of these technologies, with reference to several case studies, including two of them from the Italian context: CNR's Semantic Scout, and DigitPA's Linked Open IPA.

KEYWORDS: Library linked data; Semantic web; Governmental data; DigitPA's Linked Open IPA; CNR's Semantic Scout

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

OpLiDaF Open Linked Data Framework: a platform for the creation and publication of linked data

Tiziana Possemato

The ITACH@ project

The purpose of the ITACH@ project for Innovative Technologies And Cultural Heritage Aggregation is to provide innovative tools that will increase the value of the Italian cultural and tourist industries. The system proposed by the project, and currently in development, may be applied to the entirety of the information produced by cultural bodies and institutions such as libraries, archives, museums and tourist organisations and is also intended for use by similar, adjacent or related fields. The project aims to resolve difficulties in a context suffering from:

- a lack of awareness of, and inability to meet, the sector's need for integrated access to data, regardless of the diversity, quantity, distribution or owner of the data itself;
- the necessity for data sharing and for the data to be used (or re-used); the presence of organisations or individuals choosing to

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-6313

share data and who can benefit from the creation of organised and accessible 'ecosystems'.

The fundamental questions to be asked are:

- what is the best way of providing access to data so that it may be easily reused?
- how can the discovery of pertinent data from within a mass of available information be made possible?
- how can applications be made to integrate data from heterogeneous and unknown sources?

These issues place the ITACH@ project within the larger setting of the semantic web, raising questions regarding the publication of data in accordance with the field's standards for good practice and technological declinations, such as linked data.

The OpLiDaF platform

In particular, we will concentrate on one of the system's technological components, the Open Linked Data Framework, or OpLiDaF, drawn up as a framework for the creation, structurization and visualization of data in Resource Description Framework (RDF)/XML format. It is intended to be a specialist platform for the treatment (for example mapping, conversion, cleansing and publication) of linked data for heterogeneously formatted data, through ad hoc tools and procedures, or integrated open source systems, and through the use of standards and languages recognised by the semantic web.
The main functions of the OpLiDaF platform are:

- the selection of ontologies;

- mapping between the data of origin and ontology, or selected ontologies;
- the creation of specific ontologies from within a set of data;
- the production of RDF/XML files;
- data cleansing.

The OpLiDaF system is based on the observation of the composition and typology, despite differences in both content and format, of the data forming the body of information of libraries, archives, museums, tourist and regional organisations and other institutions. We could argue that the list quoted shows a decreasing trend in comparison with the use of recognised standard formats across the board: from libraries, these being without doubt the institutions that have most used standards for the stucturization and publication of their own data in the past, to sectors in which data is collated in Access, Excel or CSV spreadsheets. The libraries themselves, front-runners in standardization, especially in the widespread Machine Readable Cataloguing (MARC), formats, connect this data, relative mainly to bibliographic descriptions and authority files, with a range of other data in different formats, more commonly management-based data such as user profiles, lending and reservation data, acquisitions data, or descriptive and administrative data for periodicals and serials, which are often managed, for ease, convenience or tradition, outside of the centralised bibliographic database. This heterogeneous and facetted composition of information sources becomes even more evident the more one moves away from traditional library contexts towards museums and archives.

The publication of linked data from relational databases

Analysis of this heterogeneous variety of data, much of which is of great public interest, is accompanied by the awareness that, were this data to be converted into linked data, according to recognised and now widespread principles, standards and practices, neither the respective native data management systems, nor business applications, would be abandoned; we would merely see the addition of a supplementary technological layer in the linking of this data to the semantic web.

The diagram in figure 1 on the facing page allows us to analyse a possible work flow for the publication of heterogeneous data in linked data.

Without losing ourselves in different work flow hypotheses, we will focus on the high potential, through different paths and tools, for the transformation of data for the semantic web (both structured data and textual data, another vast wealth of information that is rarely taken advantage of in the traditional web, in relation to its high information potential), with the interesting scenario that we find in relation to the use (and reuse) of data, without necessarily intervening in the legacy systems being used by the organisations (we define as legacy the existing information systems or an application that continues to be used because the user cannot, or will not replace it).

The politics and practices of data publication on the semantic web vary depending on various factors, including:

- the original format of the data (structural or textual);
- the amount of data to be included in a data set;
- the frequency of data updates.

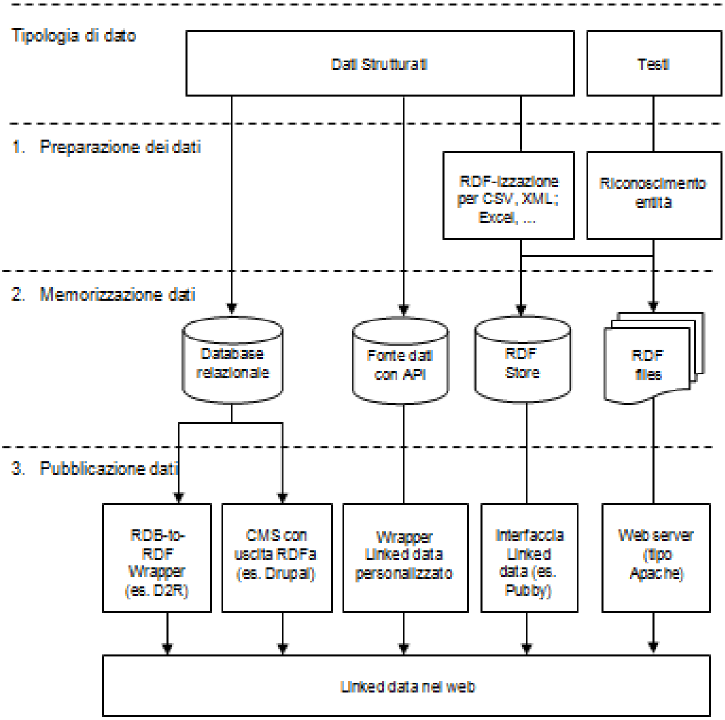


Figure 1: Workflow of the publication of heterogeneous data in linked data.

OpLiDaF concentrates in particular on the first and final of the three factors above, relative to the differing structure of the original data and the need for updates, relying on a technological methodology that produces a transverse layer intended to direct and coordinate the different management requirements for this data. If we focus on the library sector, we cannot avoid the treatment of data in MARC format (in particular, from MARC21 to RDF/XML). It is a known process, supported by a vast literature, and may be considered as the library's first step towards publishing its own data on the semantic web. We prefer, therefore, to deal with a less busy field than that of conversion from MARC21, and will pinpoint the procedures and techniques for the treatment of data contained in relational databases, in order to analyse the potential of the Open Linked Data Framework (OpLiDaF) system, which uses recognised standards and mapping language. Much structured bibliographic data in MARC21 is saved in the memory of relational databases, allowing the data to be recomposed in MARC format during exportation or in cases of external access to the data (for example, by a Z39.50 client). The exercise and study on the translation of data from relational databases to linked data is of particular interest for both bibliographical data and authority files, as this is the relational representation of the separate item of data in MARC. The publication of data from relational databases as linked data is greatly facilitated by the tools now available, which use mapping processes from the relational databases in RDF graphs, before publishing on the web according to the principles of linked data. This possibility becomes all the more interesting if consider about the enormous amount of internal management data, produced and saved in legacy systems and not necessarily destined for the web as an open and public space, but, for example, for company intranets: the same technology as linked data may be destined for internal use but just as useful

and necessary for the controlled diffusion of existing information.
The W3C RDB2RDF Working Group is working on the elaboration of standard languages for the mapping of relational data and outlines of relational databases in RDF and Web Ontology Language (OWL): the two main languages available to date are Direct Mapping (DM) and the RDB2RDF Mapping Language (R2RML). From a technological viewpoint, one of the most widespread and widely-used tools for the publication of relational databases on the semantic web is the D2R Server, which allows RDF and HTML browsers to navigate database contents using SPARQL as a search language.
These are widely recognised standards and technologies for the semantic web, but we are most interested in demonstrating the potential of another mapping language in outlines of relational and ontological databases implemented in RDF(S) or OWL, and used in the OpLiDaF platform: R2O (Relational to Ontology), which allows us to produce a wide-reaching set of primitives with an explicit and recognised semantic. R2O is a high level language separate from the RDBMS (in our case, Oracle), and works with databases that use the SQL standard. R2O is based on D2R, but aims to overcome the two main limits of the latter:

- R2O is more powerful and flexible, and therefore more suitable for the development of complete mapping, providing and a level of expression that DR2 lacks;

- R2O, unlike D2R, is a demonstrative language (that is, it allows us to specify what we want to obtain, without describing how to arrive at the result).

A supposition regarding the use of R20 is that the database and the ontology (implemented in OWL/RDF) are very similar in structure, assuming that both the database and the ontology are pre-existent and do not require modifications to be used. To demonstrate the generational flow of RDF data sets from a relational database within the OpLiDaF platform, we have selected, with the aim of mapping, existing ontologies, rather than generating the ontology semi-automatically from the relational database (another possible strategy that is very useful in contexts where usable ontologies are not available). The ontologies used in this study are:

- Bibtex = http://bibotools.googlecode.com/svn/bibo-ontology/tags/1.3/bibo.xml.owl
- Bibo = http://zeitkunst.org/bibtex/0.2/bibtex.owl

The relational database is Oracle, which contains bibliographic data and is structured, in short, in two different views, created in order to map two different ontologies: BOOK (mapped on the bibtex ontology) and PARTS (mapped on the bibo antology). Other tools used for the study include:

- an open source and multi-platform planning environment for ontologies. It is based on the Eclipse development platform and offers numerous plug-ins that are useful in covering a wide variety of functions linked to the life-cycle of ontologies; one such plug-in is ODEMapster. Neon ToolKit was developed as part of the "NeOn" project[1] and is supported by the NeOn Foundation;[2]
- ODEMapster: plug-in for the Neon ToolKit: allows for guided and extremely simple mapping operations between relational database tables and the selected ontology, as shown in the below illustration, which demonstrates the mapping phase of the BOOK view in the bibtex ontology.

[1] http://www.neon-project.org.

[2] http://www.neon-foundation.org.

Each item of data present in a column of the selected table may be mapped with a class or attribute that has been carefully selected in the ontology used.

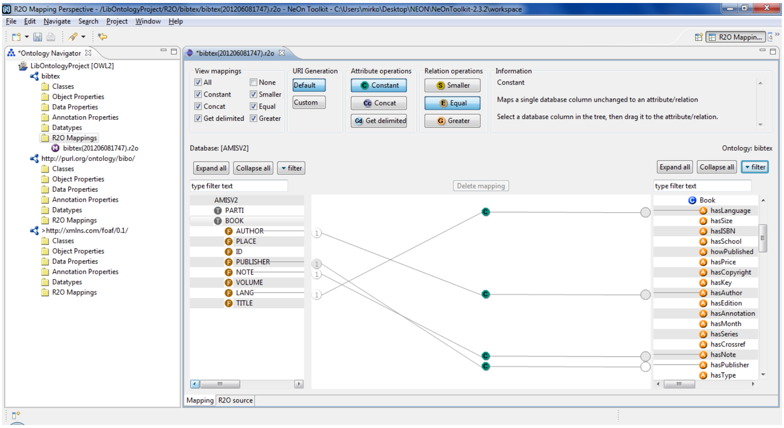
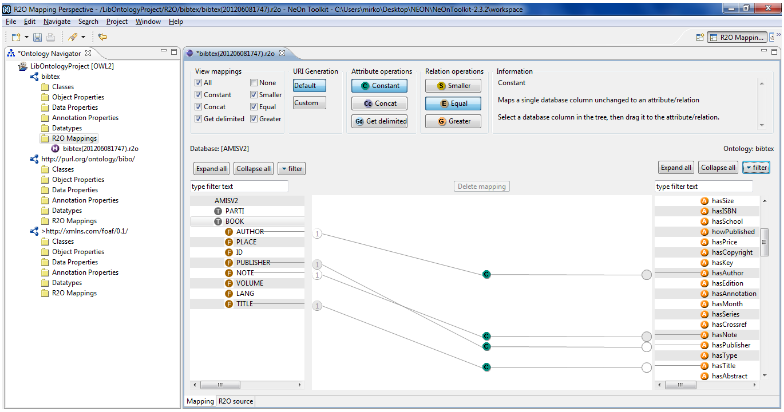


Figure 2

The left-hand section of the figures 2 and 3 on the following page shows the list of ontologies used or available for use (among these, we can see also the FOAF – Friend Of A Friend) vocabulary, which has been included, but was not used in this trial). The left-hand section of the central part of the screen shows the fields that we intend to map in the selected ontology; the ontologies, in turn, are shown in the right-hand section of the central part of the screen, where BOOK represents the class and the yellow dots are the attributes. The selection of database fields to be mapped with the ontology's attributes depends on the institution's willingness to publish and share this data. In our case, we have carried out a simple example of mapping:

- AUTHOR field: Bibtex.hasAuthor



View mappings: ☑ All □ None ☑ Constant ☑ Smaller ☑ Concat ☑ Equal ☑ Get delimited ☑ Greater

Information: Constant

Maps a single database column unchanged to an attribute/relation

Select a database column in the tree, then drag it to the attribute/relation.

Figure 3

- TITLE field: Bibtex.HasTitle
- PUBLISHER field: Bibtex.hasPublisher
- NOTE field: Bibtex.hasNote
- LANG field: Bibtex.hasLanguage

The phase following mapping between the relational database and the ontology is the production of R2O files: the XML that describe the graphic mapping between database and ontology in language form. This is required by ODEMapster to generate the RDF.

Listing 1: Small section of RDF generated by ODEMapster

```
<?xml version="1.0" encoding="UTF-8"?>
  <r20>
    <dbschema-desc name="AMISV2">
      <has-table name="PART1">
      <has-table name="BOOK">
```

```
          <nonkeycol-desc name="AUTHOR" />
          <nonkeycol-desc name="PLACE" />
          <nonkeycol-desc name="ID" />
          <nonkeycol-desc name="PUBLISHER" />
          <nonkeycol-desc name="NOTE" />
          <nonkeycol-desc name="VOLUME" />
          <nonkeycol-desc name="LANG" />
          <nonkeycol-desc name="TITLE" />
        </has-table>
      </dbschema-desc>
      <conceptmap-def name="http://purl.org/net/nknouf/ns/bibtex
          #Book">
        <uri-as type="DEFAULT">
          <operation oper-id="concat">
            <arg-restriction on-param="string1">
              <has-value>http://purl.org/net/nknouf/ns/bibtex#
                  Book</has-value>
            </arg-restriction>
            <arg-restriction on-param="string2">
              <has-column>AMISV2.BOOK.AUTHOR</has-column>
            </arg-restriction>
          </operation>
        </uri-as>
        <default_uri-as>
          <operation oper-id="concat">
            <arg-restriction on-param="string1">
              <has-value>http://purl.org/net/nknouf/ns/bibtex#
                  Book</has-value>
            </arg-restriction>
            <arg-restriction on-param="string2">
              <has-column>AMISV2.BOOK.AUTHOR</has-column>
            </arg-restriction>
          </operation>
        </default_uri-as>
```

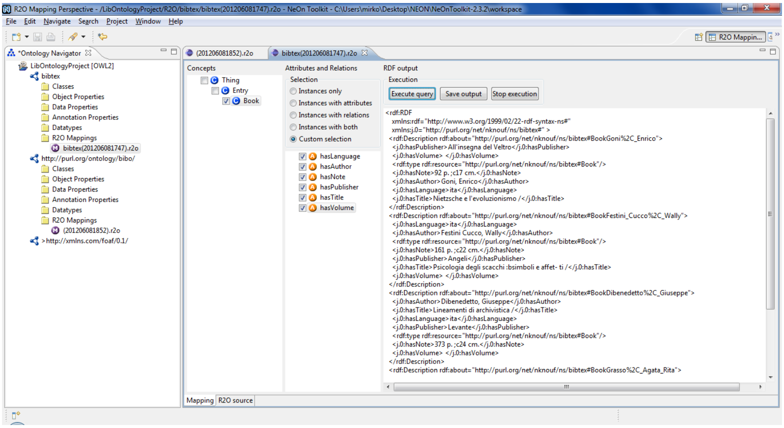
```
<described-by>
  <attributemap-def name="http://purl.org/net/nknouf/ns/
      bibtex#hasLanguage”>
    <selector>
      <aftertransform>
        <operation oper-id="constant”>
          <arg-restriction on-param="const-val”>
            <has-column>AMISV2.BOOK.LANG</has-column>
          </arg-restriction>
        </operation>
      </aftertransform>
    </selector>
  </attributemap-def>
  <attributemap-def name="http://purl.org/net/nknouf/ns/
      bibtex#hasAuthor”>
```

Thirdly, the system interrogates the database, extracts the records and maps them in RDF format according to the guidelines established in the previous phases.

We include in listing 3 on page 286 an extract of an RDF file, to assist reading.

Listing 2: Extract of an RDF file

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns\#”
    xmlns:j.0="http://purl.org/net/nknouf/ns/bibtex\#” >
  <rdf:Description rdf:about="http://purl.org/net/nknouf/ns/
      bibtex\#BookGoni\%2C_Enrico”>
    <j.0:hasVolume> </j.0:hasVolume>
    <j.0:hasPublisher>All’insegna del Veltro</j.0:hasPublisher
        >
    <rdf:type rdf:resource="http://purl.org/net/nknouf/ns/
```



Concepts

□ Thing
□ Entry
☑ Book

Selection

○ Instances only
○ Instances with attributes
○ Instances with relations
○ Instances with both
● Custom selection

☑ hasLanguage
☑ hasAuthor
☑ hasNote
☑ hasPublisher
☑ hasTitle
☑ hasVolume

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:j.0="http://purl.org/net/nknouf/ns/bibtex#" >
 <rdf:Description rdf:about="http://purl.org/net/nknouf/ns/bibtex#BookGoni%2C_Enrico">
  <j.0:hasPublisher>All'insegna del Veltro</j.0:hasPublisher>
  <j.0:hasVolume> </j.0:hasVolume>
  <rdf:type rdf:resource="http://purl.org/net/nknouf/ns/bibtex#Book"/>
  <j.0:hasNote>92 p. ;c17 cm.</j.0:hasNote>
  <j.0:hasAuthor>Goni, Enrico</j.0:hasAuthor>
  <j.0:hasLanguage>ita</j.0:hasLanguage>
  <j.0:hasTitle>Nietzsche e l'evoluzionismo /</j.0:hasTitle>
 </rdf:Description>
 <rdf:Description rdf:about="http://purl.org/net/nknouf/ns/bibtex#BookFestini_Cucco%2C_Wally">
  <j.0:hasLanguage>ita</j.0:hasLanguage>
  <j.0:hasAuthor>Festini Cucco, Wally</j.0:hasAuthor>
  <rdf:type rdf:resource="http://purl.org/net/nknouf/ns/bibtex#Book"/>
  <j.0:hasNote>161 p. ;c22 cm.</j.0:hasNote>
  <j.0:hasPublisher>Angeli</j.0:hasPublisher>
  <j.0:hasTitle>Psicologia degli scacchi :bsimboli e affet- ti /</j.0:hasTitle>
  <j.0:hasVolume> </j.0:hasVolume>
 </rdf:Description>
 <rdf:Description rdf:about="http://purl.org/net/nknouf/ns/bibtex#BookDibenedetto%2C_Giuseppe">
  <j.0:hasAuthor>Dibenedetto, Giuseppe</j.0:hasAuthor>
  <j.0:hasTitle>Lineamenti di archivistica /</j.0:hasTitle>
  <j.0:hasLanguage>ita</j.0:hasLanguage>
  <j.0:hasPublisher>Levante</j.0:hasPublisher>
  <rdf:type rdf:resource="http://purl.org/net/nknouf/ns/bibtex#Book"/>
  <j.0:hasNote>373 p. ;c24 cm.</j.0:hasNote>
  <j.0:hasVolume> </j.0:hasVolume>
 </rdf:Description>
 <rdf:Description rdf:about="http://purl.org/net/nknouf/ns/bibtex#BookGrasso%2C_Agata_Rita">
```

Figure 4

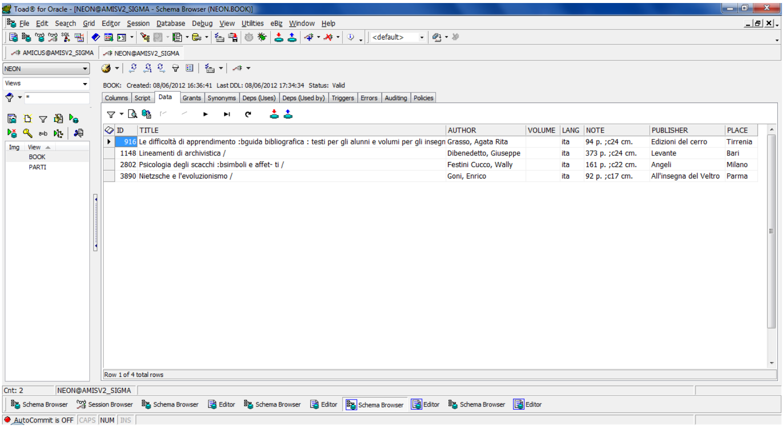
```
          bibtex\#Book"/>
    <j.0:hasLanguage>ita</j.0:hasLanguage>
    <j.0:hasAuthor>Goni, Enrico</j.0:hasAuthor>
    <j.0:hasNote>92 p. ; c17hcm.</j.0:hasNote>
    <j.0:hasTitle>Nietzsche e l'evoluzionismo /</j.0:hasTitle>
  </rdf:Description>
  <rdf:Description rdf:about="http://purl.org/net/nknouf/ns/
        bibtex\#BookFestini_Cucco\%2C_Wally">
    <j.0:hasLanguage>ita</j.0:hasLanguage>
    <j.0:hasAuthor>Festini Cucco, Wally</j.0:hasAuthor>
    <rdf:type rdf:resource="http://purl.org/net/nknouf/ns/
          bibtex\#Book"/>
    <j.0:hasNote>161 p. ; c22 cm.</j.0:hasNote>
    <j.0:hasPublisher>Angeli</j.0:hasPublisher>
    <j.0:hasTitle>Psicologia degli scacchi : bsimboli e affet-
           ti /</j.0:hasTitle>
    <j.0:hasVolume> </j.0:hasVolume>
  </rdf:Description>
```

```
<rdf:Description rdf:about="http://purl.org/net/nknouf/ns/
    bibtex\#BookDibenedetto\%2C_Giuseppe”>
  <j.0:hasAuthor>Dibenedetto, Giuseppe</j.0:hasAuthor>
  <j.0:hasTitle>Lineamenti di archivistica /</j.0:hasTitle>
  <j.0:hasLanguage>ita</j.0:hasLanguage>
  <j.0:hasPublisher>Levante</j.0:hasPublisher>
  <rdf:type rdf:resource="http://purl.org/net/nknouf/ns/
      bibtex\#Book”/>
  <j.0:hasNote>373 p. ; c24 cm.</j.0:hasNote>
  <j.0:hasVolume> </j.0:hasVolume>
</rdf:Description>
<rdf:Description rdf:about="http://purl.org/net/nknouf/ns/
    bibtex\#BookGrasso\%2C_Agata_Rita”>
  <j.0:hasLanguage>ita</j.0:hasLanguage>
  <j.0:hasTitle>Le difficolta di apprendimento: guida
      bibliografica : testi per gli alunni e volumi per gli
      insegnanti/</j.0:hasTitle>
  <j.0:hasAuthor>Grasso, Agata Rita</j.0:hasAuthor>
  <rdf:type rdf:resource="http://purl.org/net/nknouf/ns/
      bibtex\#Book”/>
  <j.0:hasNote>94 p. ; 24 cm.</j.0:hasNote>
  <j.0:hasVolume> </j.0:hasVolume>
  <j.0:hasPublisher>Edizioni del cerro</j.0:hasPublisher>
</rdf:Description>
```

The RDF may be viewed alongside the content of the relational database, as illustrated in figure 5 on the facing page.

Data cleansing

Further analysis of the RDF files produced shows certain limits and errors in the result that contrast with the result intended by the principles of linked data production. Such cases are illustrated



Toad® for Oracle - [NEON@AMISV2_SIGMA - Schema Browser (NEON.BOOK)]
File Edit Search Grid Editor Session Database Debug View Utilities eBiz Window Help
AMICUS@AMISV2_SIGMA
NEON@AMISV2_SIGMA
NEON
Views
Img View
BOOK
PARTI
Columns Script Data Grants Synonyms Deps (Uses) Deps (Used by) Triggers Errors Auditing Policies
Cnt: 2
AutoCommit is OFF CAPS NUM INS
Schema Browser Session Browser Schema Browser Editor Schema Browser Editor Schema Browser Editor Schema Browser Editor

BOOK: Created: 08/06/2012 16:36:41 Last DDL: 08/06/2012 17:34:34 Status: Valid

ID	TITLE	AUTHOR	VOLUME	LANG	NOTE	PUBLISHER	PLACE
916	Le difficoltà di apprendimento :bguida bibliografica : testi per gli alunni e volumi per gli insegn	Grasso, Agata Rita		ita	94 p. ;c24 cm.	Edizioni del cerro	Tirrenia
1148	Lineamenti di archivistica /	Dibenedetto, Giuseppe		ita	373 p. ;c24 cm.	Levante	Bari
2802	Psicologia degli scacchi :bsimboli e affet- ti /	Festini Cucco, Wally		ita	161 p. ;c22 cm.	Angeli	Milano
3890	Nietzsche e l'evoluzionismo /	Goni, Enrico		ita	92 p. ;c17 cm.	All'insegna del Veltro	Parma

Row 1 of 4 total rows

NEON@AMISV2_SIGMA

Figure 5

below. We must note that, in our case, some of these errors result from the lack of content in the data used and therefore to the low availability of expression:

- the file presents a relatively low number of assertions and relations between entity and entity (in the example we reproduced, the only relation is with the Type entity);
- the majority of the assertions have literals as their objects, making the RDF resources "bad" and isolated: the author of our example should be an autonomous entity, with a Uniform Resource Identifier (URI) reference, and not a literal, therefore: non `<j.0:hasAuthor>Goni, Enrico</j.0:hasAuthor>` ma `<j.0:hasAuthor rdf:resource= http://atcult.it/autori/283235467/>`
- some cases show separating characters with relative sub-field codes, inherited from the data structurization saved in the

Oracle tables (these are sub-field codes that are present in the MARC21 record produced in the cataloguing phase), as in the example `<j.0:hasNote>94 p.  ;c24 cm.</j.0:hasNote>` where the sub-field code $c of the record's 300 tag is present, before the field relating to the resource's dimensions.

- some assertions are invalid, as these do not have a object and therefore cannot be espressed as triples (which must be composed as subject-predicate-object).

Listing 3: Example of invalid RDF triple. The question relative is: who is the author?

```
<rdf:Description
rdf:about=http://purl.org/net/nknouf/ns/bibtex\#
    BookDibenedetto\%2C Giuseppe>
  <j.0:hasAuthor></j.0:hasAuthor>
</rdf:Description>
```

On the basis of this analysis of the RDF file produced in OpLiDaF, a series of procedures may be activated to arrive at what can be defined the phase of data cleansing, including:

- the use of cleansing tools to eliminate easily identifiable dirty characters, such as the sub-field codes of MARC21 tags;
- the identification of triple scanning processes for validity control;
- the drawing up of control procedures and the identification of literal triples in contrast to RDF triples;
- the automatic creation of entities that may be identified by URI through the use, for example, of unambiguous identifiers, in the majority of cases already present in the relational databases, or created according to established criteria.

In terms of data sharing, quality is of utmost importance and must be a fundamental characteristic for the selection of any data set produced by third parties wishing to share and link their own data to this.

OpLiDaF and the life-cycle of linked data

To conclude, we offer a summary, starting from the life-cycle of linked data which may be sub-divided into various steps and that we have divided into seven steps ("Methodological Guidelines for Publishing Government Linked Data"), what the OpLiDaF platform is able to cover:

1. identification of data source;
2. modeling of vocabulary;
3. generation of data in RDF format, through the different available mapping languages;
4. publication of the data in RDF;
5. cleansing of the data produced;
6. creation of links between different data sets;
7. making available data, with different steps, including the publication of the data set obtained by the process on the CKAN Registry (Comprehensive Knowledge Archive Network).

The platform appears to be able to completely satisfy steps 2 to 5 and constitutes a useful tool for whoever wishes to produce linked data (regardless of the management system, the data format, the size of the data set and of the mode and frequency of updates), resolving a part of the obstacles and problems that the passage from the traditional web to the semantic web may pose.

References

Villazòn-Terrazas, Boris, Luis M. Vilches-Blàsquez, and Kristin Gòmez-Pérez Asunciòn. "Methodological Guidelines for Publishing Government Linked Data". (2011): 27–49. (Cit. on p. 287).

TIZIANA POSSEMATO, @cult.
tiziana.possemato@atcult.it

Possemato, T. "OpLiDaF - Open Linked Data Framework: a platform for the creation and publication of linked data". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6313. DOI: 10.4403/jlis.it-6313. Web.

ABSTRACT: ITACH@ Project, Innovative Technologies And Cultural Heritage Aggregation, intends to offer innovative tools for the development of the tourism industry and Italian culture. This paper analyzes the particular technological component defined OpLiDaF, Open Linked Data Framework, a platform aimed the creation, structuring and visualizing data in RDF/XML. The paper discusses also the different formats, with special attention to procedures and techniques of processing data in relational databases, according to the instructions provided by the W3C Working Group RDB2RDF. It is working on the development of standard languages for mapping relational data and relational database schemas into RDF and OWL. The paper aims to show the potential of a language mapping between relational database schemas and ontologies implemented in RDF(S) or OWL, used in the platform OpLiDaF: the R2O (Relational to Ontology), which enables the production of data set extensible semantics explicit and well recognized.

KEYWORDS: OpLiDaF; Open Linked Data Framework; Library linked data

Submitted: 2012-06-15
Accepted: 2012-08-31
Published: 2013-01-15

JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

Commercial and cultural sectors: potential for data collaboration?

Graham Bell

The European Commission-funded Linked Heritage project[1] aims primarily at contributing content to Europeana, increasing the quality, richness and reuse potential of that content, and enhancing the network of expertise built up within the heritage sector by previous projects such as Athena and Minerva. But a unique facet of Linked Heritage also seeks to define how commercial organizations might engage with Europeana. This link to the world beyond libraries and other cultural memory institutions is the focus of EDItEUR[2] and its partners within the project.[3]

[1] http://www.linkedheritage.eu.

[2] EDItEUR is the trade standards body for the global book, e-book and serial supply chains. It is a not-for-profit, member-supported organisation based in London, but with a global membership of publishers, distributors, retailers, subscription agents, libraries, and system vendors. It's best known for developing the ONIX and EDItX families of metadata and transactional messaging standards, and is an acknowledged centre of excellence on metadata and identifier issues for the publishing industry. EDItEUR provides management services to the International ISBN Agency and the International ISTC Agencies, and is currently also working on projects supported by WIPO (Enabling Technologies Framework, TIGAR) and the European Commission (Linked Heritage, Arrow Plus). URL: http://www.editeur.org.

[3] Linked Heritage Work Package 4 (WP4) includes EDItEUR Ltd and the following other organizations: ICCU (Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche) – part of the Italian Min-

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5487

When discussing Europeana, of course, 'content' is actually metadata. Cultural objects, and the digital representations of those objects, remain with their host institutes. Europeana aggregates only the objects' metadata, aiming to build a comprehensive cultural discovery portal and to drive researchers, educators and students back to the websites of the originating institutes. And yet there is the 'copyright gap' – a century-long lacuna between creativity and cultural heritage. This was described in the Comité des sages's report (*The new renaissance*) as a 'black hole' of in-copyright and commercial material missing from Europe's digital cultural collections. Copyright – or doubt about copyright – can prevent the digitization of physical objects (for example, the scanning of books in libraries), and prevents institutions making digital representations of the in-copyright parts of their collections available to all via the internet. The material that cultural memory institutions deliver to Europeana is metadata describing more or less ancient objects and artefacts. Any rights and restrictions associated with the original objects, artefacts and digital representations remain in place. On this basis, the Europeana operating model is not fundamentally antithetical to commerce. However, Europeana's Data Exchange Agreement demands

istry of Cultural Heritage and Activities; mEDRA (multilingual European DOI Registration Agency Srl) – an identifier registration agency part owned by the Italian Publishers Association; MVB (Marketing- und Verlagsservice des Buchhandels GmbH) – the leading service company for the German book industry, owned by the Börsenverein des Deutschen Buchhandels, the German Publishers and Booksellers Association; NSL (National Széchényi Library) – the Hungarian National Library; Pintail Ltd – project management consultancy specializing in e-culture, library and internet technology projects; Promoter Srl– provides technical coordination and consultancy in information technology, multimedia, innovation and business development; TIB (Technische Informationsbibliothek) – the German National Library of Science and Technology. The initial report from this workgroup, written by EDItEUR's Michael Hopwood, covers metadata and identifier best practice in the commercial sector, and is available from the Linked Heritage project website, http://www.linkedheritage.eu/getFile.php?id=283.

that any rights in the metadata be waived, to allow Europeana and others to reuse and redistribute the metadata freely. Aside from the obvious difference that commercial metadata describes products that are mostly in copyright – and many of these are in commerce – some other strong contrasts need to be drawn between commercial and cultural sector metadata.

First, commercial and cultural sector metadata often describe different classes. Most cultural sector metadata is concerned with items. This is self-evident for the metadata held by an archive or a museum, as the metadata describes the individual and often unique objects or items within the collection, whether they are archaeological treasures or 19th century ephemera. For libraries, however, this is less clear: a library catalogue contains bibliographic information that is superficially similar to a national bibliography, a books-in-print database or a publisher's catalogue. But at heart, a library holdings catalogue begins as a list of the volumes in the library.[4] In familiar FRBR terms, the catalogued entities are items, with their own accession and call numbers. In contrast, a publisher's catalogue describes classes of items, or manifestations in FRBR terms, with each manifestation identified by an ISBN and comprising many individual instances or items.

Second, commercial metadata often covers a broader, richer range of data elements: a picture of the book cover, synopses of the content, extracts from the text of reviews, and a biography of the author are all common 'marketing collateral' included in ONIX for Books (ONIX is the widely-implemented standard metadata schema used in the global book trade[5]) records produced by a publisher, but not in library MARC records. There is good reason why this is so: data sells – and more data sells more. A 2011 statistical study by Nielsen (*White*

[4] But a MARC record may be more than a catalogue record, see figure 1 on page 298

[5] http://www.editeur.org/83/Overview.

Paper: The Link Between Metadata and Sales), clearly documented the positive effects of enhanced metadata on sales, whether through simple discoverability or through greater engagement with the customer. Products where a standard and very basic set of 11 metadata elements was provided saw a near-doubling of sales – both online and offline – compared with products lacking one or more of these 11 elements, and additional provision of a range of rich marketing collateral raised sales by a further 55%. There are of course other data elements required by the commercial supply chain that have no place in public-sector catalogues. The territorial nature of book rights – where a publisher may have the right to publish a work in one country but not in another – is an obvious example. This may not be familiar where a language is essentially 'national', but in English-language book publishing, it's critical for a global retailer like Amazon or Apple to know whether this product from a British publisher may also be sold in Canada or USA. There could be a different publisher or exclusive distributor who holds rights to the work in North America.

Fourth, commercial sector data is often highly dynamic. Publishers' catalogue data changes frequently. A book might be announced months before publication, and the metadata is, within that intervening period, highly provisional. Planned titles change. Publication dates change. Even author's names change. And post-publication, prices, availability, sales rights and the rich descriptive metadata are all subject to frequent updates. Commercial data is characterised by dynamic data flow rather than by static repositories of data.

Fifth, commercial sector metadata can include copyrighted content. While there is justifiable doubt over whether largely mechanical, factual bibliographic data such as title and authors could possibly be covered by copyright, publishers' metadata often includes sample text – table of contents, sample pages, perhaps even whole chapters

– that allows for no doubt. And a sui generis database right also persists over large collections of bibliographic data.

Like organizations in the cultural sector, commercial organizations commit significant resources of time and money to the creation and, more particularly, the maintenance of metadata. Maintaining rich and accurate metadata in a dynamic business environment with many thousands of new products every year is expensive – but the metadata is a key enabler for the publishing business, a core part of the process, and one that is an asset in its own right. And – somewhat ironically in the light of the growth of the open data movement – the value of that asset is growing rapidly. A decade ago, publishers employed sales teams whose sole purpose was to present books to booksellers. Increasingly, metadata is the publisher's sales team. Given the above, provision of commercial sector metadata is often accompanied by a requirement for some measure of control over the nature and context of any use made of the metadata. ONIX metadata for example often includes elements intended only for internal use within retail organizations, or data that may only be revealed publicly after some embargo date. Many publishers explicitly license use of their metadata to data aggregators or retailers, and impose restrictions on use and service level agreements on those making use of it. This might include commitments over presentation of the metadata, over accuracy and timeliness of metadata updates, over the right to redistribute the data, and above all, over clarity of business process[6] Even for those publishers that provide product metadata but forego explicit licences, an "implied licence" accompanies any metadata, and it can be argued that this limits use of the metadata to trading in, merchandising, promoting and selling the products described, and precludes redistribution. There

[6] For an example, http://www.bic.org.uk/files/pdfs/110721recipients%20best%20practice%20final.pdf.

are of course strong prima facie arguments for open licensing of data where creation of that data was publicly funded, but these do not apply where the data is created by commercial organizations. For the above reasons, commercial publishers – and organizations in other creative sectors – view product metadata as having a commercial value and sensitivity, and waiving rights to this business-critical asset would require extraordinary justification. The alternative is to strip down the range and richness of the metadata to an anodyne – and valueless – minimum, which would meet neither publishers' nor Europeana's needs. It is this issue – in effect, the construction of a business case for release of a commercially-valuable asset where all rights to that asset are waived – that will be the focus of EDItEUR's and its work package partners' effort in the second half of the Linked Heritage project.

Beyond the Europeana context, and aside from the contrasts drawn above, commercial and cultural sector metadata are in many ways complementary. In the face of budgetary pressure, many cultural sector organizations operate at least partly commercially, and publishers have long dealt with memory organizations such as libraries. There is a strong history of using commercial sector data to seed or to enrich cultural sector data.

One well-established example is the use of publishers' product metadata – in the form of ONIX records – to create CIP or MARC records for the library world. In the USA, OCLC has taken pre-publication ONIX data from publishers to construct the basis of its bibliographic records. A small British company, BDS, does the same, as part of its creation of CIP data for the British Library. The aim of these efforts is to create library-grade bibliographic records from the product records that publishers create for quite different purposes.

Carol Jean Godby of OCLC describes the process of mapping from ONIX to MARC21 records in detail, in two papers (*Mapping ONIX* *to MARC* for the latest version of ONIX 3.0; *A Crosswalk from ONIX Version 3.0 for Books to MARC 21* for the earlier ONIX 2.1). These papers present detailed 'recipes' for mapping that assert, for example, the equivalence of the ONIX <ImprintName> XML element with MARC field 260 $b or <ContributorRole> with 100/700 $e, and provide equivalent values for terms in controlled vocabularies used within ONIX and MARC.

However, such mappings are not purely syntactic, and must be constructed carefully, to ensure the maximum semantic value is carried from one record to another, without imbuing a particular metadata element with unjustified meaning and in effect 'inventing' information where nothing is implied. The two metadata schemas, and the abstract data models on which they are based, have different underlying purposes, and are not simply different ways of expressing the same information. Given the similarity of their domains, the level of semantic interoperability between ONIX and MARC is inevitably high, but not every concept in ONIX can be carried across, as many are purely supply chain-related and have no relevance to librarians or library users. Conversely, as figure 1 illustrates, ONIX for Books is not a superset of MARC – it describes only manifestations, and specifically, manifestations that are products.[7] Although an ONIX record can contain identifiers for works (FRBR expressions), this is limited to the extent that it facilitates rights trading and retail customer service.[8]

[7] In the FRBR model, books in libraries are individual items, but marc records often deal with classes of identical items (manifestations) or classes of manifestations with essentially identical content (expressions). the <indecs> model on which onix is based is similar, except that frbr expressions are indecs works. a frbr work is in effect a class of <indecs> works related to each other through revision, adaptation, translation, compilation and so on, but <indecs> models this as an inter-related group of peers rather than as a group descended from a higher-level and entirely abstract entity

[8] There is a separate ONIX metadata format used to characterise <indecs> works –

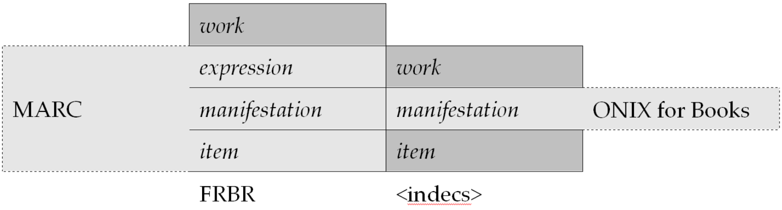


Figure 1: Rough equivalence of MARC and ONIX entities.

Despite these caveats, as Godby writes,

> the outcome of the [mapping] is a MARC 21 record with AACR2 semantics that can be automatically generated from an ONIX 3.0 source, pass a rigorous semantic validation, serve as a rough draft that can be further refined by cataloging best-practices guidelines, and qualify for inclusion in a quality-controlled library database.

Of course, what results from such a mapping is not always a library-grade record, as libraries remain more concerned than publishers with – for example – authority files and cataloguing rules, and ONIX records are not always complete because few data elements are mandatory. But the process of mapping is effective, efficient, and means that cataloguing processes can begin long before the book is available. Mapping from ONIX to MARC21 illustrates how commercial metadata can seed and enrich cultural sector metadata. But interoperability is two-way: cultural data can in principle be used to enrich existing commercial data too. The potential for this can be seen in the new International Standard Name Identifier (ISNI) for public identities of parties involved in creative endeavours.[9]

ONIX for ISTC Registration, used for the registration of ISTC (International Standard Text Code) work identifiers.

[9] http://www.isni.org.

The standard has been launched with around a million identities pre-defined, based on data from national library authority files, and use of the ISNI enables commercial metadata to differentiate between, say Prof. Richard Holmes (ISNI 0000 0001 2147 5396) and the identically-named Prof. Richard Holmes (ISNI 0000 0001 1768 5542), or to state authoritatively that Julian Cope the musician is the same party as Julian Cope the author (ISNI 0000 0000 7725 4712). Linked Heritage's predecessor project Athena[10] delivered a data mapping engine called MINT (Metadata Interoperability Services), a data schema LIDO (Lightweight Information Describing Objects), and a LIDO to Europeana (ESE) mapping. The current focus of EDItEUR and its project partners within the Linked Heritage project is on building mappings within MINT that are conceptually similar to the ONIX to MARC work outlined above. This will enable large volumes of ONIX metadata – and commercial data from other creative sectors, including recorded music (DDEX metadata), film and TV (EIDR metadata), and photography (IPTC metadata) – to be mapped into LIDO, and potentially delivered (either in whole or in part) into Europeana. The appeal of an enriched Europeana record for a van Gogh painting – say The Café Terrace on the Place du Forum, Arles, at Night – with links to the latest commercial biography of van Gogh, a modern travel guide to the city of Arles, a commercial recording of César Franck's Symphony in D minor (completed only a few days before the painting), and perhaps a contemporary photo from a picture library of the café terrace on the Place du Forum, is clear.

Listing 1: Equivalent ONIX and RDF metadata expressions.

```
<Contributor>
  <ContributorRole>A01</ContributorRole>
  <NameIdentifier>
```

[10] http://www.athenaeurope.org.

```
        <NameIDType>16</NameIDType>
        <IDValue>0000000121479135</IDValue>
      </NameIdentifier>
      <PersonNameInverted>Sjöwall, Maj</PersonNameInverted>
    </Contributor>

          http://ns.editeur.org/onix/3.0/reference/Contributor
               genid:A96837
    genid: A96837 http://ns.editeur.org/onix/3.0/reference/
         ContributorRole  http://ns.editeur.org/onix/ codelists
         /17#A01
    genid: A96837 http://ns.editeur.org/onix/3.0/reference/
         NameIdentifier  "0000000121479135" of type http://ns.
         editeur.org/onix/codelist/44#16
    genid:A96837 http://ns.editeur.org/onix/3.0/reference/
         PersonNameInverted "Sjöwall, Maj" of type http://ns.
         editeur.org/onix/ codelists/18#01
```

MARC21 and its associated English-speaking AACR2 cataloguing rules are destined to be replaced by RDA cataloguing and some yet-to-be-defined data format[11] – and this route is likely to followed by other flavours of MARC too. The destination of this journey is 'Linked Data' in some form, and it is this that holds the promise of automatically associating the metadata record for Vincent's painting with that for Cesar's symphony, thereby enriching both. Yet what we have now can best be described as 'data with links': ONIX metadata contains information linking books to people, to places, subjects, dates, other books, and the underlying data could be re-expressed in RDF as illustrated in listing 1.[12] There is an explicit ONIX data model (separate from the XML schema) to guide this re-expression. This

[11] http://www.loc.gov/marc/transition.

[12] The four RDF triples use an arbitrary blank node (a96837) to represent the contributor, and the node has three properties representing the role, name and identifier of the contributor. the use of URIs in the RDF syntax is a more easily machine-

type of adaptation is conceptually similar to mapping between ONIX and MARC, though the first step to mint and promote the canonical URIs necessary for expressing the ONIX as linked data has not yet been taken. The benefit of re-expressing ONIX (or other commercial metadata) as Linked data is that it may be simpler to process the links expressed within the data automatically. But ultimately, this may not be enough. Linked data using industry-specific vocabularies and proprietary identifiers tends to form islands of data, richly linked internally, but ultimately not well linked to the rest of the Linked Data cloud. To increase the density of links between these islands of data, it's necessary to add a semantic mapping layer that says – in effect – this term for a relationship or RDF predicate used in this industry sector is the same as that term used in a different sector. Listing 2 shows how such semantic mappings can be expressed.

Listing 2: Sample RDF showing semantic relationship between onix contributor role (*a01*, meaning 'written by', used in the second triple in figure 2), the exactly equivalent marc relator *aut* and the broadly equivalent *authorwork* term from RDA.

```
<skos:Concept rdf:about="http://ns.editeur.org/onix/codelists
    /17#A01">
  <skos:inScheme rdf:resource="http://ns.editeur.org/onix/
      codelists/17#"/>
  <skos:notation rdf:datatype="http://www.w3.org/2001/
      XMLSchema#token">
    A01</skos:notation>
  <skos:prefLabel xml:lang="en">Written by</skos:prefLabel>
  <skos:exactMatch rdf:resource="http://id.loc.gov/vocabulary/
      relators/
    aut"/>
```

processable variation on onix codelists (controlled vocabularies). Note that canonical URIs for expressing ONIX metadata in Resource Description Framework (RDF) have not been published – this is merely an illustration. the subject of the first triple is omitted, as it is in the ONIX, but could be an identifier for 'the book' such as an ISBN.

```
    <skos:closeMatch rdf:resource="http://rdvocab.info/roles/
        authorWork"/>
  </skos:Concept>
```

Similarly, some agreement on public identifiers used for common entities – people and their public identities, places, organizations etc, is necessary. If each heritage and commercial sector uses a different sector-specific or proprietary identifier for a public identity, for example, then making links between sectors becomes reliant on the error-prone process of matching names. The use of a common, cross-sector public identifier – ISNI in this case – solves this issue.

So when large volumes of data from a range of commercial and cultural sectors are aggregated, the interconnectedness of the data – the degree to which data from one sector can enrich that of another – is dependent on careful semantic mapping and the use of identifiers rather than textual names. It is the use of common public identifiers, interoperable semantics and shared vocabularies that is the glue that allows triples to be bound together automatically, inferences made and implicit connections to be revealed. Without these, disparate databases cannot be bound into a single data space.

References

Breedt, Andre and David Walter. *White Paper: The Link Between Metadata and Sales*. Woking: Nielsen Book, 2011. http://www.isbn.nielsenbook.co.uk/uploads/3971_Nielsen_Metadata_white_paper_A4(3).pdf. (Cit. on p. 293).

Godby, Carol Jean. *A Crosswalk from ONIX Version 3.0 for Books to MARC 21*. Dublin, Ohio: OCLC Research, 2012. http://www.oclc.org/resources/research/publications/library/2012/2012-04.pdf. (Cit. on p. 297).

—. *Mapping ONIX to MARC*. Dublin, Ohio: OCLC Research, 2010. http://www.oclc.org/resources/research/publications/library/2010/2010-14.pdf. (Cit. on p. 296).

The new renaissance: Report of the "comité des sages" on bringing Europe's cultural heritage online. Bruxelles: European Commission, 2011. DOI: 10.2759/45571. (Cit. on p. 292).

GRAHAM BELL, EDItEUR.
info@editeur.org

Bell, G. ”Commercial and cultural sectors: potential for data collaboration?”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5487. DOI: 10.4403/jlis.it-5487. Web.

ABSTRACT: The main goals of the Linked Heritage project (sponsored by Europeana) are to provide qualified content to Europeana from the public and private sector. To this aim is created WP4 (Work Group 4) in which the organization EDItEUR takes part. The ’content’ are metadata for the cultural heritage. Starting by addressing the issue of the “copyright gap”, which can involve metadata, the article notes the differences between metadata types developed by the private sector (ONIX for books) and those defined by the public one (FRBR, MARC, MARC21). The aim is to develop integration of both sectors into a shared references core. Exploring the common reference framework requirement, the article emphasizes the new International Standard Name Identifier (ISNI) potential, which allows to uniquely identify the subjects involved in the creative field. The future outlook can be further enhanced by involving additional metadata mapping that relates books, people, places, data, other books and other references in a possible ’Linked Data’ form, within which priority should be given in common public identifiers, related semantic mapping layers and shared vocabularies.

KEYWORDS: Europeana; ISNI; Library linked data; Linked Heritage Project; EDItEUR

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

Bibliographic standards and Linked Data. Towards a collaboration between cultural and commercial sectors

Patrizia Martini

The new scenarios that are arising in the world of the web of data, with the adoption on the part of private and public institutions of new technology for the interlinking of structured data, such as the linked data model, highlights the great potential and possible development of this instrument in the library world. Giving an identity to data – whether open or not – in order to make them interlinkable and interoperable, and establishing links between objects that can be related to each other, aggregated, and reused, means creating added value for data management and, most of all, offering a new service for the citizenry. Thanks to the actions of the Digital Agenda for Europe,[1] the *Guidelines for semantic interoperability through linked open data*[2] were published. They provide a reference framework for the production of open data that is interoperable between public administrations, thus making data management in the public sector accessible and transparent. In the library sector, relationships with all

[1] http://ec.europa.eu/information_society/digital-agenda/index_en.htm.

[2] http://www.digitpa.gov.it/sites/default/files/CdC-SPC-GdL6-InteroperabilitaSemOpenData_0.pdf

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-8598

information management actors need to be analyzed and redefined. In particular, collaboration and dialogue between all subjects involved in the information chain of production needs to be promoted, in order to analyse different standards and spur the development of innovative initiatives in the cultural field that can become a key factor for economic and social growth. Among the activities involved in this process, a central role is played by those related to the standards used in the library sector and other commercial sectors, and their integration. The adoption on the part of libraries of data models that are more structured and granular compared to traditional catalogue formats, would help enrich the metadata traditionally used. On the other hand, a comparison of commercial standards with well-established, traditional library, museum, and archival standards would ensure the quality, authoritativeness, and sustainability of data. In keeping with its institutional mission, Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane e per le Informazioni Bibliografiche (ICCU) has always been responsible for elaborating national rules for the documentation of all types of materials, and for the definition and diffusion of international standards, with the goal of achieving interoperability among the various cultural sectors. At the national level, through Committees and Working Groups staffed by experts, including some working in other institutions of the Ministry of Cultural Heritage and Activities, it has disseminated international standards and drafted guidelines for the documentation of cultural heritage. Through the new Regole italiane di catalogazione (REICAT) cataloguing code, which is derived from the Functional Requirements for Bibliographic Records (FRBR) model, a data model was obtained that represents bibliographic items in a more granular way. In the National Library Service (Servizio Bibliotecario Nazionale - SBN), the Index database has the bibliographic records structure as a system of ″entities-relationships″, and the application of standards and cataloguing rules refer to within Servizio Bibliotecario Nazionale (SBN) has reflected this structure. Thanks to the new SBN Machine Readable Cataloguing (MARC) format, the representation of ″entities-relationships″ is more complex and flexible; indeed, it is possible to record data on all types of material according to various cataloguing and network participation levels. On the digitalization front, the Institute coordinates the harmonization of national digitization policies and the development of standards for metadata and guidelines for application. The new research and working group on metadata, which sees the participation of museums and archives, has launched the following fields of activity based on linked data:

- recovering SBN′s ″semantics″ through the mapping of SBN-MARC and Resource Description & Access (RDA) elements present in the Open Metadata Registry;[3]
- analysis of the state of the application of persistent identifiers in the linked data field;
- research on the Schema.org [4] initiative for the codification of the Resource Description Framework (RDF) query language in HTML pages; this codification allows search engines to ″understand″ the structure and typology of the documents being indexed;
- strengthening collaborative efforts with the Virtual International Authority File (VIAF), by augmenting the sending of authority records of SBN authors archive;
- the analysis of a crowdsourcing model that uses the contributions of the users of bibliographic services, with the aim of

[3] http://metadataregistry.org.
[4] http://schema.org.

enriching the metadata traditionally produced by professional communities;

- at the international level, ICCU coordinates and participates in European projects for the management and accessibility of digital content, including MINERVA, MICHAEL, ATHENA, DC-NET, Europeana, INDICATE, Linked Heritage, WDL, Partage.

Within these projects, standards play a key role, with the elaboration of best practices for the definition of heterogeneous data sets. Within the framework of the Europeana platform,[5] starting on July 2012, the metadata related to digital objects concerning the European cultural heritage will be presented according to the linked open data model, with a creative commons license, as set out under the Data Exchange Agreement underwritten by the European partners. Collaborations with the public and private sectors include a work with the community of editors to develop a new vocabulary for support, media, and format types, which led to the creation of the RDA/ONIX Framework for Resource Categorization which has merged into the library community, particularly within IFLA's Committees. This work constitutes the broadest grouping framework for the categorization of resources of varying contents and available on all support types used by the various communities. Established with the goal of supporting the needs of libraries and publishers and to facilitate data transfer and dialogue between the communities, this effort has produced a vocabulary of categories upon which the vocabularies of International Standard Bibliographic Description (ISBD). Consolidated edition and the new RDA are based. The controlled vocabularies and the RDA elements are available in the Open Metadata Registry as linked data in RDF. The new descriptive elements for the ISBD Area 0 ”Content form and support type” are

[5] http://www.europeana.eu/portal.

also published in Italian in the Open Metadata Registry. ICCU is currently studying ways to implement these elements in the SBN Index and the Library Management System. Another successful public-private partnership was established through the European ARROW Plus project, coordinated by the Italian Publishers' Association (Associazione Italiana Editori - AIE), which has developed a distributed system for managing information on authors' rights. ICCU, one of the project's partners, is the "National Contact Point" for Italian libraries. The primary goal of the Linked Heritage[6] European project is to foster an increase in contribution to Europeana portal from both the public and private sectors. In particular, activities carried out as part of WP4 "Public and Private partnership" focused on analyzing and comparing the metadata used in the two sectors, in terms of defining and sharing metadata models, integrating data from the commercial sector (books, music, film and TV, photography), developing standards and best practices, data mapping and aggregation, licenses and business cases. The issue of integrating standards from various communities and the adoption of models to represent complex data is also tackled in the recommendations set out in the final report of the working groups established by the World Wide Web Consortium (W3C) for the feasibility study for the Library Linked Data Incubator Group. Final report,[7] published in October 2011. The findings of this working group launched a series of activities whose ultimate purpose is the adoption of the linked data paradigm as an indispensable tool towards an open, interoperable semantic web. Integration with data produced by libraries with other web resources highlights their new role with regards to the broader online information community. Coordination with non-institutional actors involved with production, dissemination, and

[6] http://www.linkedheritage.eu.

[7] http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025.

access to information, means creating added value to information and providing users with a high quality service, but also adopting a new attitude towards the institutional world's tasks concerning the field of bibliographic control and the promotion and diffusion of culture.

Patrizia Martini, ICCU. Department for International projects of cultural heritage.
p.martini@iccu.sbn.it

Martini, P. ”Bibliographic standards and Linked Data.Towards a collaboration between cultural and commercial sectors”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #8598. DOI: 10.4403/jlis.it-8598. Web.

Abstract: The Central Institute for the Union Catalogue of Italian Libraries and Bibliographic Information (ICCU) is the National Bibliographic Agency for the coordination and dissemination of international standards and cataloguing rules within Italian libraries. Linked Heritage, the European project coordinated by ICCU, will contribute large quantities of new content to Europeana, from both the public and private sectors. The Work Package 4 Public private partnership is working to produce an innovative investigation that highlights the differences between metadata creation in the cultural sector (libraries, museums and archives), compared with the metadata of commercial sector. The work which the Group is carrying out, will point out the different roles and purposes between the public and private sectors: how commercial metadata can enrich the data with more information giving potential benefits and new services to Europeana users, while libraries models and standards provide stability, harmonization and an higher level of granularity of the data in order to facilitate the functions of resource discovery. WP4 Public private partnership will offer an opportunity of producing several metadata mappings. This paper analyzes the enriched information that can be offered beyond the traditional bibliographic description.

Keywords: Library linked data; Linked Heritage Project; Europeana

Submitted: 2012-06-15
Accepted: 2012-08-31
Published: 2013-01-15

Linked open data on its way into next generation library management and discovery solutions

Axel Kaschte

Ex Libris is known as an innovative company - we really embrace new technology. And linked data is a very new revolutionary technology. We have seen in the agenda of this event that there are many aspects of linked data and I have tried to put some light on what a commercial company like Ex Libris looks at in this new field. Whenever I look at something new, I remind myself of the Little Prince. He has this special capability of looking at certain things with fresh eyes. You probably remember this question, «What do you see here?»while the Little Prince is showing a drawn picture. Most people will answer immediately that this is a hat. But we all know (from the book) what the answer is – a snake that has swallowed an elephant. Now the next picture I show to you is this curve, looking very similar to that hat, and as you can imagine, I now put the question to you, «What do you see here?»It's not the snake, this much I can tell you. It is a bell curve. To be more precise, it is the bell curve of the adoption of innovation from Rogers Everett (*Diffusion of Innovations*) who invented this way of presenting this information, it demonstrates how many usages of new technology over time will happen. So you see in the beginning of a new tech-

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5492

nology, we have 2.5% of innovators. These are libraries that get new technology going, so really they are trying to invest in new things. Early adapters are the next phase with 13%. It's already a remarkable percentage but it's still called innovative. Then we see the early majority, the late majority and so on. Just to give you some examples of products we know in these internet times, there are a TV and a newspaper. If you are on the right hand side of the bell curve of technology adoption it doesn't mean that it is not used - quite the opposite. It means that everybody already has it and is using it. The bell curve refers to the growth rate of how many users are added each month. Facebook is on top right now, which means that the growth rate is still dramatic but there will be less and less new people coming on board. I mean, there are already 600 million users already. You also see also new websites such as Vimeo, and technologies like the iPhone. Blackberry is a little more on top; the iPhone is newer. Amazon Kindle, the e-book reader, is only in the early adapter phase. It has been on the market for the past three years but the adaption curve shows how much is still ahead for this product. You see also there is profit opportunity mass market. The point is now, as we a commercial company, of course we are interested in money, but also we are interested in serving customers like you with commodity services. Whenever there is a technology and the aim is to make it available to many, to really make it available as a cheap solution, then it is a commercial company that has the best model. How does this translate to libraries? Integrated Library System (ILS), integrated library systems, were invented a long time ago, some 30 years or more, and you see there are still some libraries, especially in Asia, who do not yet have a library system. This explains why the growth rate of new libraries is slow but these are still state of the art technologies and they are still in demand. Then you have things like meta-search. You have things like link resolution.

It's very interesting that, if you look at our statistics, such a large number of libraries are still purchasing link solutions. There are many who have not yet entered into the area of electronic resources but are either in the process of doing so or will in the near future. For that purpose they will need a link resolver, which explains why it is still in this high area of the bell curve. Let's have a look at Discovery, which is the next generation of OPAC, where you see search engine technologies coming into play: discovery is a little bit before the top of the bell curve. We just look at our statistics: Primo, as our discovery solution, is growing rapidly but is not at the top yet.[1] And you can also see central e-resource indexes like Primo Central, a mega-aggregation of electronically-available articles for research, for scholars, and so on. This is something which started just two years ago and it has been adapted very quickly and we are in the phase of rapid growing. These are just product categories. I will now put the technologies next to them. ILS is a technology, which is not growing anymore. Search Application Programming Interface (API), Open URL, search engines, and then cloud, as an e-technology which enables this kind of service to have one central index for every library and also offers increased cost-effectiveness. This highlights why commercial companies are so good at these models; if they can provide a solution very cost-effectively to many libraries, this is the model libraries should use. In this model there is also some more details to share with you. Geoffrey Moore (Moore) introduced the concept of a chasm. There are many products who are very much in the innovation phase and who will never make it as mainstream products. There fall into the chasm. Two of the things Ex Libris invented fell into the chasm: an ERM solution - probably many of you have heard of Verde - and a digital asset management system. If you look, it's not just Ex Libris who failed to deliver these

[1] http://www.exlibrisgroup.com/category/PrimoOverview.

to the mass-market. There are a few hundred using it, compared to four thousand using Aleph worldwide. It's not growing anymore. The chasm is just a way to present the certain technology which is not meeting the needs of the library in the best way: it's never reached to be a best practice. I'm just being very honest. You have to analyse as a commercial company and be able to say, "OK, that was a mistake. It was money spent that was not good for all of us, neither for you nor for us". What we are now introducing, and the whole market is following this idea now, is a solution to the problem of automation in silos. If we go away with the silos, if we go unified - meaning there is one solution for your print management and your e-resource management and probably also your digital assets management - meaning if you introduce one single environment, we find that this is very much what libraries want today. And this is what we are in the initial phase of doing right now; our first customer will probably go live with it next month. It's the software called Alma in our case, and it is cloud technology which allows it to be done. I showed you all of this because I want to bring to you the idea that commercial companies have to look very close when it is the right time to get on board with a new technology in order to make a mass product out of it. Just look at certain other technologies which are established in ILS, we have heard about them today: AACR2, MARC21, Union catalogues, authority files, they have been around for quite a long time. We have also heard a lot about the emerging technologies like RDA. Alan Dunskin from the British Library talked about it and we listened to him asking, «Can you please help us to close the gap and get it used; provide the tool set where the cataloguing happens». In other words, it is the right time and Ex Libris will look into providing the necessary tools in its applications soon. Then we have Resource Description Framework (RDF), which the whole seminar is talking about. And we have open data sources. All of these new technologies are very far on the left of the bell curve. These things are really still in the research phase, as are many aspects of linked open data. We have seen these schematics of objects and their relations; we have books, we have paintings, we have authors and painters. We have these objects and the creators of it – and many more relations. Library data was already highly linked in the past in certain ways. It was not open, it was not using URIs but inside of the systems it was already linked. So if you look at solutions like discovery systems (e.g. from the Austrian Union Catalogue in Vienna) you see things where you can click at the author and get all the manifestations of the works from the author, you have the same for subjects. So you see links are there and you can navigate them but you only stay inside the environment of the library. They also have already a permalink to the manifestation. With this link you will always get to the same point. It's not yet a URI because you don't get to the data, you just get to the same page but it's at least this permanent way to get there. Building an API to just give you the data in a structured form is just the next step. But now there is one point more you can see – this is what they have done in the Austrian Union Catalogue by including Wikipedia as a data source outside the library metadata and they use the authority record with its identifier in there to link into Wikipedia and if you click on it you get some information from Wikipedia. A very simple example, it seems. There are several such examples in various discovery solutions in various libraries. These are not yet using the true URI mechanism. The links are constructed on the fly and it's something which just works because the discovery platform and the data underneath allows to present this to the end user. With this "experiment" in place we can have a look at the acceptability by the end users. Is it something they actually want? Is it something they actually click on? And if they don't, we don't bother. So you see this is the kind of thing we try to do here in this research phase. So what is actually new with linked open data from our point of view? It's not new that you can do linking between manifestations and the authors. We also have subject linking. We can introduce links between many types of library data but the data structure is highly specialised, no one outside the library can actually read it and it's very difficult to exchange and interact with outside the library. So from my point of view if I have to summarise to an outside person who is not from the library business, what is the important thing for libraries in the linked open data theme, it's making the library data available to the outside and maybe even more important, take library data from the outside. In other words, make the library world part of the all-embracing World Wide Web. This is reflected by the work of the World Wide Web Consortium, W3C. It has established an incubator group to look at linked data in libraries and related in software developments in May 2010. This group was looking at real use cases and submitted theLibrary Linked Data Incubator Group Final Report.[2] These use cases are about getting library data into the linked data world. We have seen several of these cases in presentations during this seminar and I will try to summarise them here by putting the various different use cases into a very simple pattern of 3 areas of work. But first we follow what the incubator group did, they categorised all use cases into 8 groups. The first group is about the handling of bibliographic data, bringing it over to a linked data scenario. For example, British National Bibliography, Bibliotheque National France, Bavaria State Library, the Open Library. The second group is about authority data, same institutions but different data. This is just repetition of things you know to get you to a point where you see a pattern here. Third group is work on vocabulary alignment. Many presentations have been done during

[2] http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025.

the seminar on that. Fourth group is about archives working on getting their data into the linked data cloud. Europeana – interestingly enough – is mentioned in the archives group although they are probably working in all the groups. Fifth group is about citation of scientific data sets being expressed in link data, a very new thing, it wasn't done so far. It is now helping to enhance publications, which means that publications come already with metadata about the research data used. The sixths group is about digital objects in the library world. The goal is to provide a digital text repository as linked data so that the metadata, the text and the extra objects the text is referring to are provided in one comprehensive format. Here we have use cases from outside the library world, the UK open government data initiative. It provides many examples where you can draw data from and see how they are interlinked. The sevenths group is about collection building. Librarians have talked about Functional Requirements for Bibliographic Records (FRBR) structures, in which the work level is the highest level. But what if you go even to a higher level and start describing collections. There are already use cases trying to define collections in data sets and more. The eights (and last) group of use cases is about social networks and cross linking environments. The uses cases in this group seem to be not be related to a classical library view but the typical users of libraries are very active in these areas exchanging information especially about the literature they are using, like e.g. via Mendeley. All of this work in the 8 groups, when looking at it from a little distance, translates into three main work areas.

- One area is the data preparation. Creating the data which needs to be there in linked structures to be able to use it. This area of work is about creating tools to be able to handle mass transformation and mass storage with high performance.

- At the same time there is the area of the definition of the rules to transform the data. It's kind of an interaction between the two. There are projects which have tried to put data into link structures but then they get some experience and they need to say "no, you have to change the transformation rules, we have to do it again", so we are in a very frequent iteration process right now.

- At the end, all that matters is what really comes to the end-user interface and how they can they make more use of it than today, which is the third area of work.

Somehow it seems that there are far more projects about getting data into the linked data cloud than there are projects about what to actually do with this data what could not be done before. This and the high frequency of changes to the definition rules of transformation result in our conclusion at Ex Libris that linked data is still in research mode. One of the research examples is Europeana, we have heard about it just today during the seminar. Europeana have a website, it's a productive site - why do I call it research? By looking at the problems they face. The central Europeana portal is not able to deliver state-of-the art performance. The problem arises by sticking to one of the main ideas of linked data – to link data of various sources together. These sources are in fact data silos and to make them discoverable from one central place one has to do a federated search. We do have the experience of the last decade of doing meta-search in library databases, in order to create performing solutions one has to create a central index. Central indexes means harvesting from the various sources and that actually means a variety of source formats, and most important versioning. In the RDF world this is highly problematic and in my view an area of research. This is all known to the very knowledgeable people working at Europeana. In order to make progress in this area a new a European Community funded project has started just this March. It's called

DM2E Digitised Manuscripts to Europeana. The major part is to digitise more materials and to get it done quicker and to easily create metadata. However, work package 2 is about interoperability infrastructure. Because many of the institutions who do the digitisation are libraries, they have library systems and use them to collect the metadata in classical library format like MARC. Because Europeana uses link data structures, a robust RDF transformation toolset will be created as part of work package 2 (WP2). Ex Libris is a partner in WP2, taking part in this research and actually creates products which will be open source and which will be possible to add to an existing library system. This tool will take e.g. MARC-XML and transforms this into RDF. We have already talked about similar examples like the British National Bibliography, we hear this afternoon how the Bavarian State Library has done it. However, these examples are not using common technology; it is something which is in an experimental phase. The tool which is created in WP2 will allow various input formats like MARC21, UNIMARC, DC, MODS, transforming this into a RDF presentation, which in essence is just a different transport format. As a second step, a transformation into the Europeana data model will be done. Both of these steps are based on mapping rules and actually the task is here to make it very easy to change these mapping rules because we are in the phase of defining the vocabularies and that's why we would like to play with it. Currently we still do not know the definite vocabulary that should be used, in every project mentioned in the use case report mentioned above, they use a different ontology. Creating a tool which allows to play with the ontologies is our contribution to the current research phase. To summarize – Why should we as Ex Libris start investing in products using linked open data technology? It is because of interoperability, especially with other domains, in the discovery sector. It's probably reshaping metadata management –

cataloguing will most likely look totally different, it will be more about including external resources as links, as it is typing data.

References

Moore, Geoffrey A. *Crossing the Chasm*. New York: Harper Collins, 1992. (Cit. on p. 315).

Rogers, Everett M. *Diffusion of Innovations*. Glencoe: Free Press, 1962. (Cit. on p. 313).

AXEL KASCHTE, Ex Libris Group.
axel.kaschte@exlibrisgroup.com

Kaschte, A. ”I Linked open data on its way into next generation library management and discovery solutions”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5492. DOI: 10.4403/jlis.it-5492. Web.

ABSTRACT: Library Linked Data Model is an important topic for librarianship and it is equally of interest to the many organizations that provide products and services to that community. Ex Libris, as one of those organizations, frequently gets asked: ”Where do we see it fitting into our plans?”. In order to be able to answer questions like this, we need to ask: ”What exactly are the problems being solved for the profession by this technology that can only be solved with the Library Linked Data model?” What most developers/providers of products analyzing the potential of library linked data would see is that at this stage, this technology is very much in the research stage. The presentation talks about the research ExLibris is involved in and how this can be utilized by innovative libraries to help defining the actual use cases in which the potential of the Library Linked Data Model is indeed exploited.

KEYWORDS: Library linked data; Ex Libris

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

Linked heritage experience in linking heritage information

Gordon McKenna

Introduction

This paper will look at the experience of the EC-funded Linked Heritage project in the area of linked data. It will cover:

- the project in context;
- work package 2 - Linking Cultural Heritage Information;
- the results of research into the use of linked data in the cultural heritage sector;
- a look forward to the further work of the project.

Overview of the Linked Heritage Project

The Linked Heritage project is part-funded by the ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme. The project began in April 2011 and lasts for 30 months. The project has three objectives:

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-6304

- to contribute large quantities of new content to Europeana, from both the public and private sectors;
- to demonstrate enhancement of quality of content, in terms of metadata richness, re-use potential and uniqueness;
- to demonstrate enable improved search, retrieval and use of Europeana content.

Linked Heritage aim to facilitate and deliver large-scale, long-term enhancement of Europeana and its services. It addresses the problems associated with:

- non-standard descriptive terminologies;
- the lack of private sector and 20th Century content:
- the preservation of complex metadata models within the Europeana metadata schema.

Project partners include all the key stakeholder groups from 20 EU member states, with Israel and Russia. They include ministries and responsible government agencies, content providers, aggregators, leading research centres, publishers and SMEs.[1]
The objectives of the project are:

[1] Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche (IT); Università Degli Studi di Padova (IT), Consiglio Nazionale delle Ricerche (IT), Ministère de la Culture et de la Communication (FR), Eesti Vabariigi Kultuuriministeerium (ER), Hellenic Ministry of Culture (GR), National Technical University of Athens (GR), University of Patras (GR), Collections Trust LBG (UK), An Chomhairle Leabharlanna Ireland (IE), Pintail Ltd (IE), Fundacio Privada I2CAT, Internet i Innovacio Digital A Catalunya (SP), Philipps Universitaet Marburg (GW), Stiftung Preussischer Kulturbesitz (GW), Central Library of the Bulgarian Academy of Sciences (BU), Javni Zavod Republike Slovenije za Varstvo Kulturne Dediscine (Slovenia), The Cyprus Research and Educational Foundation (CY), Stowarzyszenie Miedzynarodowe Centrum Zarzadzania Informacja (PL), Riksarkivet (SW), MEDRA S.R.L. (IT), Gottfried Wilhelm Leibniz Universitaet Hannover (GW),

- to contribute large quantities of new content to Europeana, from both the public and private sectors;
- to prepare for the enhancement of the quality of both new and existing Europeana content, in terms of its metadata richness, its re-use potential and its uniqueness;
- to demonstrate improved search, retrieval and use of Europeana content, both within the Europeana portal and by third parties via the Europeana API.

It is doing this by:

- assembling representative stakeholder groups (content providers, aggregators, ministries and policy making bodies, technologists, private sector companies, and associations);
- consultation, consensus building, networking, and the sharing of perspectives and priorities;
- the identification and promotion of best practice. This is the most appropriate and useful standards, specifications and recommendations for the contribution, ingestion and enhancement of Europeana content;
- large scale implementation (including the necessary technology integration in compliance with the Europeana standards)

Editeur Limited (UK), MVB (Marketing und Verlagsservice des Buchhandels) (GW), Orszagos Szechenyi Konyvtar (HU), Koninklijke Musea voor Kunst en Geschiedenis (BE), Institutu Umeni - Divadelniho Ustavu (Czech Republic), Instituto Superior Tecnico (PO), Valsts Agentura Kulturas Informacijas Sistemas (Latvia), PACKED (Platform voor de Archivering en Conservering van Audiovisuele Kunsten) (BE), CORDIA (Slovakia), Universita Degli Studi di Roma La Sapienza (IT), C.T.F.R. SRL (IT), Departament de Cultura i Mitjans de Comunicació (SP),Promoter di Masi Pietro & C S.N.C. (IT), Université de Savoie (FR), Association Dedale (FR), UMA Information Technology (AU), Digital Heritage LBG (UK).

and validations of the identified best practice standards and specifications. These will serve to provide to Europeana 3 millions new objects;

- the preparation of a demonstrator how the improved specifications are to be applied and how to implement the enrichment of Europeana content;
- training and dissemination to build capacity and awareness in the cultural heritage sector, particularly in the use of Linked Heritage technical outputs, but also in Europeana technologies.

The project is split into seven work packages:

WP 1 – Project management and coordination Deals with the basic project management of linked heritage, e.g. monitoring progress and managing the relationship with the Commission. Also manages the setting up and maintenance of working groups, both national and thematic.

WP 2 – Linking Cultural Heritage Information Looks at the potential use of linked data in the cultural sector (see next section).

WP 3 – Terminology Works on the enabling of the use of terminologies with the project and in a wider cultural heritage context.

WP 4 – Public Private Partnership Explores the standards in use in the non-heritage commercial cultural sectors, and the possibilities with integrating this with the cultural heritage sectors, especially with Europeana.

WP 5 – Technical Integration Enables the technical tools and requirements of the project.

WP 6 – Coordination of Conten Manages the process of giving access to Linked Heritage's partners to Europeana.

WP 7 – Dissemination & Training Making the wider community aware of the project's work, and producing learning tools to enable that community to use the results.

Work Package 2 – Linking cultural heritage information

This paper is part of the results of this work package. Its objectives are:

- to explore the state of the art in linked data and its applications and potential;
- to identify the most appropriate models, processes and technologies for the deployment of cultural heritage information repositories as linked data;
- to consider how linked data practices can be applied to cultural heritage information repositories, to enrich them and to allow them to align with other linked data stores and applications;
- to explore the state of the art in persistent identifiers (both standards and management tools);
- to identify the most appropriate approach to persistent identification, e.g. a unique standard or a set of different standards;
- to design a feasibility model and to realised a demonstrator of a flexible, scalable, secure and reliable infrastructure for a network of 'linked data enabled' cultural heritage information repositories;

- to explore the state of the art in cultural metadata models, and in particular their interoperability across libraries, museums, archives, publishers, content industries, and the Europeana models: Europeana Semantic Element (ESE); and Europeana Data Model (EDM);
- to outline the potential benefits that richer cultural heritage metadata could bring to Europeana, and to the other services which will use it.

Linked data in the cultural heritage sector

Partner Survey

As part of the tasks the work package carried a survey of Linked Heritage partners, and providers. This covered, amongst other things, their knowledge of linked data and their experience in using linked data.

Respondent information

Table 1 on the facing page shows that the content being supplied to Europeana through the Linked Heritage project covers all of the cultural domains including aggregators. However there is also significant number of responses from organisations which are not contributing content and therefore they will not appear in the metadata section of the survey. Nearly 60% of the respondents are not one of the ′standard′ types. Therefore it is useful to list what was the response was to the question: ″If you ticked ′Other′ please give organisation type″:

- mediator between providers and Linked Heritage project;

Respondent type	Number of respondents	%
Museum	4	10.3
Library	5	12.8
Archive	4	10.3
Sound archive	1	2.6
Publisher	0	0
Aggregator	10	25.6
Other	23	59.0

Table 1: Here are the figures for the types of organisations that responded to the survey.

- group of museums;
- governmental organisation for the protection of immovable cultural heritage and of the movable and living cultural heritage associated with it;
- National Books in Print;
- technical partner;
- university;
- DOI [Digital Object Identifier] registration agency;
- centre for research and innovation;
- Ministry of Culture;
- company in cultural heritage field;
- scientific research institute with museum collections;
- management and quality services company;
- National contact point;
- SME – consultancy;

- public broadcaster and media archive (video, sound, and photographs);
- publishing standards body;
- theatre documentation (photographs);
- public organisation;
- regional public administration responsible for the cultural heritage information system;
- technology provider;
- association and information centre;
- cultural agency.

Countries

Country	Number of respondents	Country	Number of respondents
Austria	1	Ireland	2
Belgium	4	Israel	1
Bulgaria	1	Italy	6
Cyprus	1	Poland	1
Czech Republic	1	Russian Federation	1
Estonia	1	Slovakia	1
France	3	Spain	2
Germany	4	Sweden	1
Greece	3	United Kingdom	2
Hungary	1		

Table 2: Here are the figures for the countries where respondents are based.

Obviously, figures in table 2 reflect the partners of the project, but there is a spread throughout Europe, with a couple of respondents outside the EU. Taken as a whole, the information about respondents leads the authors of the deliverable to conclude that the sample is fairly representative of the sector.

Linked data

Awareness

Response	Number of respondents	%
Yes	30	75.0
No	10	25.0

Table 3: To ″Are you or your organisation familiar with the concept of linked data?″

The ″No″ surprised the authors, but shows that there is a ′market′ for information and tools about linked data!

Use

Response	Number of respondents	%
Yes	7	17.5
No	33	82.5

Table 4: To ″Have you or your organisation had experience of using linked data in connection with your collections?″

Those who answered ″Yes″ were asked to give details of which source(s) of linked data they use and why they use it′. The sources used were: DBpedia (4); GeoNames (3); Freebase (1); IPTC (1); Thesauri in SKOS (1). Only two respondents gave information as to why they used a source: DBpedia (interesting information source); GeoNames (for place name disambiguation).

Publication

Those who answered ”Yes” were asked to give details. Three respondents gave details: http://data.kunstkamera.ru/sparql and http://data.kunstkamera.ru/; full bibliographic records of OPAC and Digital Library (OSZKDK) in DC. Name authority in FOAF; Thesaurus in SKOS, http://nektar.oszk.hu/wiki/Semantic_web, support RDFa in Digital Library (OSZKDK); the Department for the French Archives had published its thesaurus in SKOS in a linked data reuse perspective. An ongoing national project will bring together all the vocabularies in use in the ministry in order to get a network of concepts that would be connected to other initiatives such as RAMEAU in SKOS.

Response	**Number of respondents**	**%**
Yes	4	10.0
No	36	90.0

Table 5: To ”Have you or your organisation had experience of publishing linked data in connection with your collections?”

Linked data projects and initiatives

Response	**Number of respondents**	**%**
Yes	15	37.5
No	25	62.5

Table 6: To ”Do you or your organisation know of any linked data projects or initiatives in your country in the field of cultural heritage?”

Those who answered ”Yes” were asked to give details. The responses, ordered by country, are listed in table 7.

Country	Project or initiative [2]
France	RAMEAU
	ISIDORE
	Pactols
	BABEL
	COLLECTIONS
	PALISSY
	EROS
	PATRIMOINE LOT
	WIKIMEDIA COMMONS FRANCE:
Germany	Linked data service of the German National Library
	”Several initiatives throughout the country”
Israel	Vocabularies of the Israel Museum Jerusalem (SKOS)
Italy	Linked Open Data Italia
	SPAR ontologies
	Datagov.it
	LinkedOpenCamera
	Spaghetti Open Data
Russia	Open Kunstkammer
Sweden	LIBRIS
Spain	Open Data Gencat
	Euskadi
	Patmapa
	Cantabria’s Cultural Heritage Ontology
United Kingdom	Various government data sets

Table 7: Linked data projects and initiatives - Responses details

Europeana Open Data Agreement

Response	Number of respondents	%
Yes	11	29.7
Not sure	20	54.1
No	6	16.2

Table 8: To ”Europeana's new licence requires that providers will have to agree to have the metadata that they provide to Europeana published as Linked Open Data. This means that any 3rd party use, including commercial, is permitted. Does your organisation agree to this?”

[2] Details of responses listed in table 7. RAMEAU: http://www.cs.vu.nl/STITCH/rameau/index-fr.html, ISIDORE: http://rechercheisidore.fr, Pactols: http://www.frantiq.fr/thesaurus-pactols, BABEL: http://babel.alienor.org, COLLECTIONS: http://www.culture.fr/fr/sections/collections/moteur_collections, PALISSY: http://www.culture.gouv.fr/public/mistral/dapapal_fr?ACTION=NOUVEAU&USRNAME=nobody&USRPWD=4%24%2534P, EROS: http://www.c2rmf.fr/pages/page_id18479_u1l2.htm, PATRIMOINE LOT: http://www.patrimoine-lot.com, WIKIMEDIA COMMONS FRANCE: http://commons.wikimedia.org/wiki/Accueil, Linked data service of the German National Library: http://www.d-nb.de/eng/hilfe/service/linked_data_service.htm, ”Several initiatives throughout the country”, Vocabularies of the Israel Museum Jerusalem that have been migrated to SKOS:http://www.imj.org.il/imagine/thesaurus/allobject.htm and http://www.imj.org.il/imagine/thesaurus/objects/objectTOC.htm, ItalyLinked Open Data Italia: http://www.linkedopendata.it/en-home, SPAR ontologies: http://opencitations.wordpress.com/2010/10/14/introducing-the-semantic-publishing-and-referencing-spar-ontologies, Datagov.it. Associazione italiana per l'Open Government: http://www.datagov.it, http://www.linkedopencamera.it, http://www.spaghettiopendata.org, RussiaOpen Kunstkammer: http://www.kunstkamera.ru, Sweden LIBRIS (joint catalogue of the Swedish academic and research libraries): http://www.kb.libris.se, Spain-Open Data Gencat:http://dadesobertes.gencat.cat/en/index.html, Euskadi: http://opendata.euskadi.net/w79-home/es/, Patmapa: http://patmapa.gencat.cat/, Cantabria's Cultural Heritage Ontology: http://hdl.handle.net/10760/13938, United Kingdom, Various government data sets: http://data.gov.uk.

Respondents were also asked to explain their answer. Those who answered ″Yes″ said (with numbers):

- 1 – Publishing on Web means Open Data;
- 1 – Participated in the ATHENA project;
- 1 – Metadata provided to Europeana specifically selected for open linked data.

Those who answered ″Not sure″ said:

- 4 – Metadata not ours (our providers′ decision);
- 4 – Under discussion;
- 2 – Under discussion (possible legal obstacles);
- 2 – Decision not ours (made at a higher level);
- 1 – Will provide minimal data;
- 1 – Against commercial reuse.

Those who answered ″No″ said:

- 3 – Against 3rd party commercial use;
- 1 – National policy does not allow commercial use;
- 1 – Do not contribute to Europeana.

The Linking Open Data Cloud

The Linking Open Data Cloud[3] (The Cloud) is the best known representation of linked data. It shows ′packages′ of linked data and the

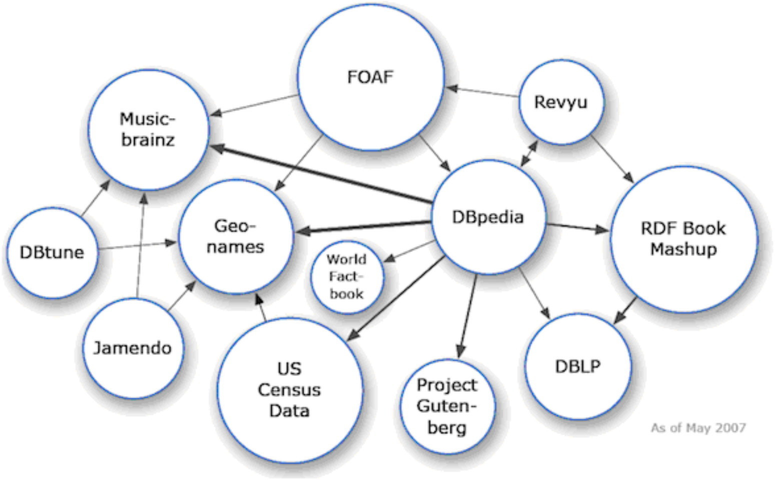


Figure 1: The Cloud in May 2007

links between packages. In May 2007 it looked like in figure 1 (with 12 packages).

By September 2011 the version that is colourised to represent the domain of the package looked like in figure 2 on the facing page (with 311 packages). It can be seen that The Cloud is growing very quickly and, in its latest form, it is becoming very difficult to get a proper overview of what it made up of. Luckily The Cloud is maintained using a wiki which is maintained on The Data Hub website.[4] This effort is part Linking Open Data community project[5] which is part of the W3C's Semantic Web Education and Outreach Interest

[3] http://linkeddata.org.

[4] http://thedatahub.org, http://thedatahub.org/group/lodcloud.

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

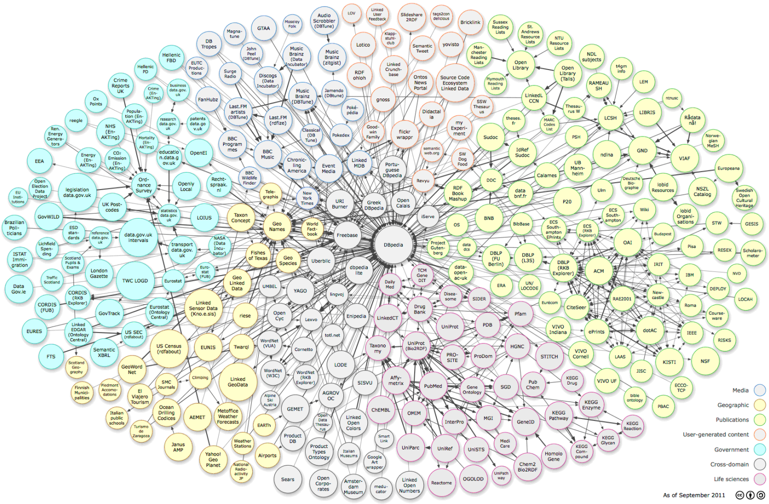
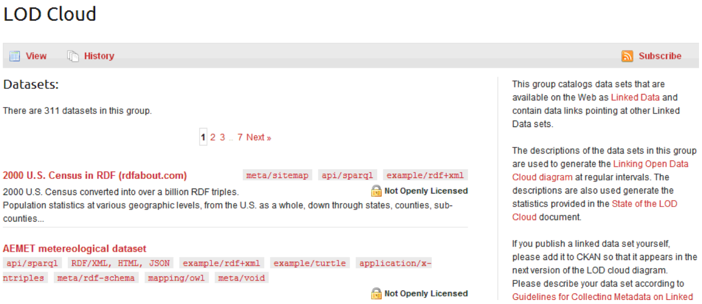


Figure 2: The Cloud in September 2011

Group (SweoIG).[6] Therefore it may be considered as representing a significant proportion of the linked data available. The Data Hub is a registry of open (and not open) knowledge with information on packages and projects (including the LOD Cloud ’group’). Once the LOD Cloud group is chosen a user is presented with the first of a set (currently seven) of result screens, as shown in figure 3. For each package the results screen gives information about:

- name of the package (as a link to the full record);
- description of the package;
- links to the resources (including examples) available for the package;
- IPR status of the package.



LOD Cloud

Datasets:

There are 311 datasets in this group.

2000 U.S. Census in RDF (rdfabout.com) meta/sitemap api/sparql example/rdf+xml

2000 U.S. Census converted into over a billion RDF triples. Not Openly Licensed

Population statistics at various geographic levels, from the U.S. as a whole, down through states, counties, sub-counties...

AEMET metereological dataset

api/sparql RDF/XML, HTML, JSON example/rdf+xml example/turtle application/x-ntriples meta/rdf-schema mapping/owl meta/void

Not Openly Licensed

This group catalogs data sets that are available on the Web as Linked Data and contain data links pointing at other Linked Data sets.

The descriptions of the data sets in this group are used to generate the Linking Open Data Cloud diagram at regular intervals. The descriptions are also used generate the statistics provided in the State of the LOD Cloud document.

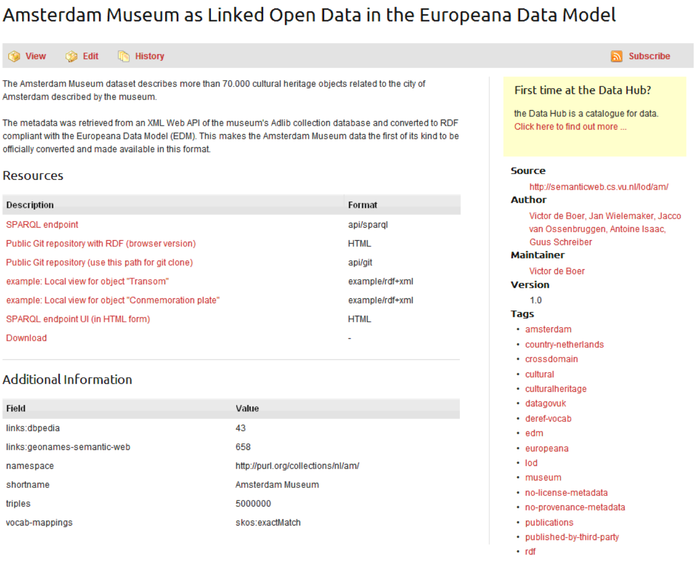
If you publish a linked data set yourself, please add it to CKAN so that it appears in the next version of the LOD cloud diagram. Please describe your data set according to Guidelines for Collecting Metadata on Linked

Figure 3: The Data Hub search results screen

For each package there is a full record, as shown in figure 4 on the facing page. For each package the full record screen includes additional information about:

[6] http://www.w3.org/wiki/SweoIG.

- which other packages are linked to (including number of links);
- the number of ’triples’ in the package (a measure of size)
- further details (not visible in the screenshot) about the IPR situation of the package;
- in Tags:
  - subject information;
  - which ’formats’ are used.



Amsterdam Museum as Linked Open Data in the Europeana Data Model

The Amsterdam Museum dataset describes more than 70.000 cultural heritage objects related to the city of Amsterdam described by the museum.

The metadata was retrieved from an XML Web API of the museum's Adlib collection database and converted to RDF compliant with the Europeana Data Model (EDM). This makes the Amsterdam Museum data the first of its kind to be officially converted and made available in this format.

Resources

Description	Format
SPARQL endpoint	api/sparql
Public Git repository with RDF (browser version)	HTML
Public Git repository (use this path for git clone)	api/git
example: Local view for object "Transom"	example/rdf+xml
example: Local view for object "Conmemoration plate"	example/rdf+xml
SPARQL endpoint UI (in HTML form)	HTML
Download	-

Additional Information

Field	Value
links:dbpedia	43
links:geonames-semantic-web	658
namespace	http://purl.org/collections/nl/am/
shortname	Amsterdam Museum
triples	5000000
vocab-mappings	skos:exactMatch

First time at the Data Hub?

the Data Hub is a catalogue for data.
Click here to find out more ...

Source
http://semanticweb.cs.vu.nl/lod/am/

Author
Victor de Boer, Jan Wielemaker, Jacco van Ossenbruggen, Antoine Isaac, Guus Schreiber

Maintainer
Victor de Boer

Version
1.0

Tags
- amsterdam
- country-netherlands
- crossdomain
- cultural
- culturalheritage
- datagovuk
- deref-vocab
- edm
- europeana
- lod
- museum
- no-license-metadata
- no-provenance-metadata
- publications
- published-by-third-party
- rdf

Figure 4: The Data Hub package record

Is The Cloud ′open′?

This may seem to be a strange question to ask. However when first examining the information on The Data Hub website it became apparent that there is a significant component of The Cloud that is not open. In The Cloud ″Open″ means ″able to be re-used commercially″.

Examining the data showed:

In terms of packages (311)

IPR Status	%
Open	42.6
Not open	57.4

In terms of triples (c38 billion)

IPR Status	%
Open	30.9
Not open	69.1

Table 9

This result is rather surprising as it shows that the majority of The Cloud is not open. One reason for this anomaly may be that The Cloud is rather like a historic landscape with the evidence of many different time periods apparent at the surface. In this case the assumption is that we are seeing many packages which are early components of The Cloud, at time when IPR and having a licence was not considered important. That being said the latest update still has ′Not open′ packages. Other insights can be gained by looking at the licences being used in more detail.

Which IPR licences are used?

Open licenses

Of the 132 packages (c11.9 billion triples) with open licences:

Licence type	**% by Package**	**% by Triples**
Creative Commons Attribution (CC BY)	28.8	45.8
Creative Commons Attribution Share Alike (CC BY-SA)	18.2	10.2
Open Data Commons Public Domain Dedication and Licence (ODC PDDL)	10.6	0.2
Creative Commons CC Zero (CC0)	9.1	2.9
UK Crown Copyright with data.gov.uk rights	7.6	27.4
Other (Public Domain)	6.8	7.0
Other (Open)	5.3	5.0
Other (Attribution)	3.0	0.4
UK Open Government Licence (OGL)	3.0	0.1
GNU Free Documentation Licence (GNU FDL)	3.0	0.0
Open Database Licence (ODbL)	2.3	0.9
GNU General Public Licence (GNU GPL)	0.8	<0.1
New BSD license and Simplified BSD licence	0.8	<0.1

Table 10

The dominant use of CC BY for an open licence is to be expected. It is an obvious choice, together with CC BY-SA and ODC PDDL and CC0. The latter is a relatively new option, and is the choice made by Europeana, and at second hand by its providers, for its publication of linked open data. It is the most permissive of the open licences with attribution being a 'recommendation' rather than mandatory. One national initiative is worth mentioning, is that in the United Kingdom. Much data is being published by the UK government using its own open data licences. At the moment these make up over 10% of The Cloud. The UK Open Government Licence is interoperable with CC BY.

Not open licenses

Of the 178 packages (c26.7 billion triples) with licences that are not open, or with no licence information:

Licence type	% by Package	% by Triples
not given	69.1	89.4
None	14.6	0.3
Creative Commons Attribution Non-commercial (CC BY-NC)	7.3	5.8
Other (Not Open)	6.7	<0.1
Creative Commons Attribution (CC BY)	1.1	0.6
Other (Non-Commercial)	0.6	3.9
Creative Commons Attribution Share alike (CC BY-SA)	0.6	<0.1

Table 11

From the above[7] it can be seen that for over 80% of packages and nearly 90% triples of the 'not open' part of The Cloud or there is no information about the IPRs. It is interesting to note that this situation does not seem to impact on the use of The Cloud, and that some of the newest packages do not have licences. For those who publish their data in The Cloud with a licence, but do not want their data to be open, then one of two options is taken:

- CC BY-NC;
- their own 'non-standard' licence with, presumably, special requirements.

How big is The Cloud?

As mentioned above there are c38 billion triples in The Cloud. There is a large distribution in size. 9 packages (2.89%) have over a billion triples. Nearly a quarter of the packages are relatively small with

[7] Please note that CC BY and CC BY-SA are open but in the data are described as not open. We have preserved this in the table.

less than 100,000 triples. The smallest has only 368 triples. This suggests that there is an element of ’test’ linked data in The Cloud, which is confirmed by some packages being described as ’test’. The average number triples in a package is c124 million. The ten largest packages with open licences are:

Package	**Number of triples**
LinkedGeoData	3.00 billion
UK Legislation	1.90 billion
Linked Sensor Data (Kno.e.sis)	1.73 billion
data.gov.uk Time Intervals	1.00 billion
DBpedia	1.00 billion
Open Library data mirror in the Talis Platform	0.54 billion
The Open Library	0.40 billion
Freebase	0.34 billion
transport.data.gov.uk	0.33 billion
Data Incubator: MusicBrainz	0.18 billion

Table 12

LinkedGeoData (CC BY licence) is a knowledge base of spatial obtained from the OpenStreetMap[8] project. Its aim is to give a semantic element to the Semantic Web. Three packages – UK Legislation, data.gov.uk Time Intervals, and transport.data.gov.uk – are part of an UK Government initiative to publish their public data in an open manner. All of them are published under the ”UK Crown Copyright with data.gov.uk rights”, a UK specific open licence. Linked Sensor Data (Kno.e.sis) (CC BY licence) has data on information on weather stations and observations from a US university-based centre. DBpedia, Open Library data mirror in the Talis Platform, The Open Library, and Freebase are well-known sources of ency-

[8] http://www.openstreetmap.org.

clopaedic information on a wide range of topics. They also have a range of different open licences: CC BY-SA, Other (Open), Other (Public Domain), and CC BY. Data Incubator: MusicBrainz (Other (Public Domain) licence) contains information about music, specifically: albums, artists, tracks, labels and their relationships.

The ten largest packages without open licences are:

Package	**Number of triples**
TWC: Linking Open Government Data	9.80 billion
Data.gov	6.40 billion
Source Code Ecosystem Linked Data	1.50 billion
2000 U.S. Census in RDF (rdfabout.com)	1.00 billion
PubMed	0.80 billion
DBTune.org MySpace RDF Service	0.66 billion
UniParc	0.63 billion
DBTune.org AudioScrobbler RDF Service	0.60 billion
Linking Italian University Statistics Project	0.59 billion
UniProt UniRef	0.49 billion

Table 13

TWC: Linking Open Government Data is the largest package in The Cloud and is an aggregation of US government data. It includes data published in the Data.gov package. The Data Hub does not have any information about the licence for this data. 2000 U.S. Census in RDF (rdfabout.com) is also US government data about population statistics, and has a CC BY-NC licence. The following packages have no licence information on The Data Hub:

- Source Code Ecosystem Linked Data contains structured source code facts from open source projects. It is authored by a Canadian university.

- PubMed is a US-based source of medical publications.

- DBTune.org MySpace RDF Service and DBTune.org AudioScrobbler RDF Service are part of a mini-cloud of nine music-related packages.

- UniParc and UniProt UniRef are parts of life science knowledge bases from US academic institutions.

- Linking Italian University Statistics Project is the publication of Italian Government data about university students.

What are the subjects in the data?

Within the descriptions for each package within The Data Hub wiki are a number of different 'tags'. Some of these tags are obviously subject-based and give an indication of the content of the packages. There does not seem to be a controlled terminology that is being used. So the same subject may be represented by a different tag in different packages. In our analysis we have combined a number of tags which appear to be the same subject. Note also packages can have more that on subject. After this process theten most common subjects in The Cloud are shown in table 14 on the next page. This result generally follows the categories illustrated by the colourised version of The Cloud diagram. It is also a 'snapshot' of the current state of the content. The Cloud is dominated by data in these areas. By comparison there is very little cultural heritage data. This is probably because, until the advent of Europeana, there has been no interest in linked data in this community. The appearance of 'United Kingdom' as a tag shows largely the effect of the UK Government's policy of publishing linked data. The role of the USA is not apparent, but this because packages are not tagged 'United States' even when potentially they could be.

Subject tag	**Number of packages with tag**	**% of packages with tag**
publications	94	30.23
government	54	17.36
life sciences	46	14.79
geographic	40	12.86
media	32	10.29
library	22	7.07
United Kingdom	22	7.07
education	20	6.43
user generated content	19	6.11
bibliographic	15	4.82

Table 14

Which formats are used to encode data?

In order to encode data for The Cloud various formats are used. In most of the literature on linked data the term used for them is 'vocabulary'. We continue to use 'format' here to avoid confusion with the cultural heritage use of vocabulary as being the descriptive terms being used rather than the metadata elements. Also of note is that some of the formats are called 'ontologies'. The most commonly used are listed in table 15 on the facing page.
There seem to be three types of format:

Basic – Those that generally organise the entities in The Cloud, including links between the entities. They are found in use in nearly all the packages in it, as might be expected. Therefore it is likely that any cultural heritage package will also use them. They are: Resource Description Framework; RDF Schema; Web Ontology Language; and XML Schema.

Descriptive – Those whose elements hold descriptive data about the entities for use in many packages. They are generally developed by a set of interested parties who want to publish their information as linked data. Quite often they have their origins

Format	Number of packages using the format	% of packages using the format
Resource Description Framework (rdf)	261	83.92
Dublin Core (dc)	97	31.19
Friend of a Friend (foaf)	84	27.01
Simple Knowledge Organization System (skos)	57	18.33
RDF Schema (rdfs)	42	13.50
Web Ontology Language (owl)	34	10.93
Basic Geo (geo)	25	8.04
Advanced Knowledge Technologies Reference Ontology (akt)	22	7.07
eXtensible HyperText Markup Language (xhtml)	19	6.11
Bibliographic Ontology (bibo)	14	4.50
none given	13	4.18
Music Ontology (mo)	13	4.18
DBpedia Ontology (dbpedia)	12	3.86
vCard (vcard)	11	3.54
Semantically-Interlinked Online Communities (sioc)	10	3.22
Creative Commons (cc)	8	2.57
Functional Requirements for Bibliographic Records (frbr)	6	1.93
GeoNames Ontology (geonames)	6	1.93
XML Schema (xsd)	6	1.93
Event Ontology (event)	5	1.61

Table 15: The abbreviation in brackets after a format's name is the 'namespace' for that format.

in a specific project or initiative.They are: Dublin Core (for web resources); Friend of a Friend (persons); Simple Knowledge Organization System (terminologies); Basic Geo (geographical); Bibliographic Ontology; Music Ontology; vCard (business cards); Semantically-Interlinked Online Communities (social networks); Creative Commons (IPR); Functional Requirements for Bibliographic Records and Event Ontology.

Package specific – Those whose elements represent the specific data held in a particular package. They were developed in the context of the publication of a single package as linked data. However they can be used in the publication of other packages which may lead to them becoming de facto standards. They are: Advanced Knowledge Technologies Reference Ontology, DBpedia Ontology, and GeoNames Ontology. That there are some formats of this type that are used by more than one package is significant. It suggests that these 'parent package' is playing a significant role in The Cloud. Obvious examples of this are DBpedia and GeoNames, and we shall see a similar pattern when we look at linking in The Cloud in the next section. It is surprising, when Berners-Lee suggests using a 'standard' format, to find that 75 formats are used by two or less packages. What we are seeing is perhaps, taking a biological analogy, is an evolutionary explosion in 'species' in a new environment. For the sake of interoperability it may be hoped that 'survival of the fittest' will begin to act. It seems that linked data is still in an experimental phase.

How is The Cloud linked?

The most important part of The Cloud is how the packages are linked together. The Data Hub site allows us to see the detail of the links. The ten most commonly linked to packages, in terms of the number of packages linking, are:

Package being linked to	**Number of packages linking**	**Number of links**
DBpedia	158	31,531,365
GeoNames Semantic Web	42	9,353,935
(none)	34	0
DBLP Computer Science Bibliography (RKBExplorer)	27	1,338,927
Association for Computing Machinery (ACM) (RKBExplorer)	26	1,487,410
ePrints3 Institutional Archive Collection (RKBExplorer)	26	281,385
Freebase	25	10,452,728
CiteSeer (Research Index) (RKBExplorer)	24	805,921
School of Electronics and Computer Science, University of Southampton (RKBExplorer)	24	37,996
ReSIST Project Wiki (RKBExplorer)	24	408

Table 16

The clear 'winners' are DBpedia, GeoNames Semantic Web, and Freebase. These are linked to by 50.8%, 13.5% and 8.0% of the other packages in The Cloud. It is supposed that this success is due their being well-known. The six packages in the list with '(RKBExplorer)' at the end of names are part of a mini-cloud of about 50 packages. RKBExplorer[9] is a system for publishing linked data, developed during the EC-funded ReSIST[10] project. It has a browser that allows users to explore the interlinked data sets. It is interesting, and perhaps at first glance surprising, to note that over 10% of the packages in The Cloud do not link to other packages. They are generally linked to, or have been published in order to be linked to. Included in this group are some of the largest packages, e.g. Data.gov, 2000 U.S. Census in RDF (rdfabout.com), data.gov.uk Time Intervals, UniParc, The Open Library, and GeneID. The ten most commonly linked to packages, in terms of number of links, are:

[9] http://www.rkbexplorer.com.
[10] http://www.resist-noe.org.

Package being linked to	Number of packages linking	Number of links
UniProtKB Taxonomy	6	46,630,898
MARC Codes List	3	42,409,958
QDOS	1	40,000,000
UniProtKB	10	33,447,122
DBpedia	158	31,531,365
Ordnance Survey Linked Data	16	29,717,902
UniParc	1	27,534,215
IdRef: Sudoc authority data	3	20,040,000
Sudoc bibliographic data	1	20,000,000
flickr™wrappr	4	16,358,998

Table 17

DBpedia is the only package to appear in this and the previous list, which reinforces its 'popularity'. flickr™wrappr is extensively linked from DBpedia to provide images for its concepts. Packages with 'UniProt' at the beginning of their name, and the UniParc package, are part of a mini-cloud of the subject of proteins. Sudoc is the French academic union catalogue, and the links here are between packages related to it. Ordnance Survey Linked Data is geographical data for the UK, and linked to by packages from that country, especially UK government data packages. QDOS is connected to a package dealing with popular music. This analysis shows that the linking of packages is not something that is, at least at the moment, growing in an 'organic' way. There are initiatives which are responsible for creating large parts of The Cloud. The implication is that for the cultural heritage sector that such an initiative needs to happen too. Europeana is taking a leading role in such an initiative.[11]

[11] http://version1.europeana.eu/web/lod.

Cultural Heritage data in The Cloud

There are 18 packages in The Cloud that could be identified as having 'cultural heritage' as their subject or related to it:

Package	IPR	Number of triples
VIAF: The Virtual International Authority File	(not given)	200,000,000
Europeana Linked Open Data	(not given)[12]	185,000,000
British National Bibliography (BNB)	CC0	80,249,538
Hungarian National Library (NSZL) catalog	(not given)	19,300,000
Amsterdam Museum as Linked Open Data in the Europeana Data Model	CC BY-SA	5,000,000
Library of Congress Subject Headings	(not given)	4,151,586
Swedish Open Cultural Heritage	Other (Open)	3,400,000
Calames	[not given]	2,000,000
RAMEAU subject headings (STITCH)	[not given]	1,619,918
data.bnf.fr - Bibliothèque nationale de France	(not given)	1,400,000
National Diet Library of Japan subject headings	(not given)	1,294,669
Gemeenschappelijke Thesaurus Audiovisuele Archieven – Common Thesaurus Audiovisual Archives	ODbL	992,797
Gemeinsame Normdatei (GND)	Other (non-commercial)	629,582
Archives Hub Linked Data	CC0	431,088
Thesaurus for Graphic Materials (t4gm.info)	CC BY-SA	103,000
Italian Museums (LinkedOpenData.it)	CC BY-SA	49,897
Thesaurus W for Local Archives	(not given)	11,000
MARC Codes List Open Data	Other (Public Domain)	8,816

Table 18

Two of the packages are directly related to Europeana: Amsterdam Museum and Europeana itself. There is evidence of a French effort with linked data, especially terminologies: Calames, RAMEAU subject headings (STITCH), data.bnf.fr - Bibliothèque nationale de France, Thesaurus W for Local Archives. This was also seen in the Linked Heritage partners' survey. Sweden is also doing something similar with Swedish Open Cultural Heritage. Italy is also starting to follow the same path. There is an additional terminology and authority file component with: VIAF: The Virtual International Authority File, British National Bibliography (BNB), Library of Congress Subject Headings, National Diet Library of Japan subject headings, Gemeinsame Normdatei (GND), Thesaurus for Graphic Materials

[12] This will eventually be published as CC0.

(t4gm.info) and the MARC Codes List Open Data. Finally there is a contribution from the domains of libraries (Hungarian National Library (NSZL) catalog), archives (Archives Hub Linked Data), and audio-visual archives (Gemeenschappelijke Thesaurus Audiovisuele Archieven – Common Thesaurus Audiovisual Archives). The part of The Cloud from cultural heritage is still rather small (c500m triples or <1.5%). However developments from Europeana are planned to significantly increase its size. Linked Heritage will be a significant component of it. Let us further explore further details about the cultural heritage mini-cloud. Cultural heritage packages use formats listed in table 19.

Format	**Number of packages using the format**
Resource Description Framework	13
Simple Knowledge Organization System	11
Dublin Core	7
eXtensible HyperText Markup Language	4
Friend of a Friend	3
Basic Geo	1
Bibliographic Ontology	1
DBpedia	1
Music Ontology	1
Object Reuse and Exchange	1
RDF Schema	1
vCard	1
Web Ontology Language	1
XML Schema	1

Table 19: Formats used

The general picture is similar to The Cloud as a whole, except that the use of SKOS is much more significant, indicating the importance of terminological resources and authority files in the sector; Of note is the absence of a format for museum information specifically. Also the Europeana Data Model is not mentioned in The Data Hub, but from other sources was used by Amsterdam Museum, and probably by the Europeana packages.

Cultural heritage packages in The Cloud link to targets listed in table 20.

Package being linked to	Number of packages linking	Number of links
DBpedia	5	82,308
Library of Congress Subject Headings	4	108,135
VIAF: The Virtual International Authority File	2	1,820,684
GeoNames Semantic Web	2	510,658
Dewey Decimal Classification (DDC)	2	200,543
RAMEAU subject headings (STITCH)	2	83,530
Swedish Open Cultural Heritage	1	100,489
Gemeinsame Normdatei (GND)	1	20,000
IdRef: Sudoc authority data	1	10,000
(DCMI Type Vocabulary – not in The Cloud)	1	10,000
UK Postcodes	1	5,000
AGROVOC	1	700
Hungarian National Library (NSZL) catalog	1	136
(none)	1	0

Table 20: Targets of links in The Cloud

As one might expect DBpedia is the most popular package to link to. Another 'general' package linked to is GeoNames Semantic Web. Both of these were also identified in the Linked Heritage survey, and represent well known sources of cross-domain and geographical information to link to this. Apart from this the rest of the linked packages are mainly other cultural heritage packages, and especially standard terminologies and authority files. Looking at the use of serialisations listed in table 21 on the next page. RDF/XML is used by all but two of the packages: Europeana Linked Open Data uses mentions only N-Triples, and the Calames Package do not mention

Serialisation	Number of packages using (%)
RDF/XML	16 (88.9%)
N-Triples	5 (27.8%)
Turtle	1 (5.5%)
(none given)	1 (5.5%)

Table 21: Serialisations

any serialisation. N-Triples are usually published together with RDF/XML. The one occurrence of Turtle is in combination with RDF/XML. This suggests that cultural heritage linked data should be, at least, published as RDF/XML and possibly as N-Triples in order to be compatible to existing data. However there is no reason why all the serialisations cannot be used.

Best practice recommendations

The publication of linked data is still at the experimental stage. Best practice can only be said to be emerging. Therefore the recommendations given in this section are based on:

- common practice in the general linked data community, as represented by The Cloud;
- the practice of cultural heritage organisations that have published linked data;
- the general practice of the cultural heritage sector.

Some of the recommendations offer a range of options, with no ′right′ choice. The choice an organisation makes is dependent on individual circumstances, and may be affected by legal and ethical considerations. The recommendations can be separated into three ′choice areas′:

What information to publish as linked data

Looking at what kind of information is being published as linked data in The Cloud, and especially the relatively small part which is about cultural heritage, two main types of information should be considered:

Collections information

This will be the bulk of the information that will be published by cultural heritage organisations. However they should also consider publishing information about:

- surrogates – the results of digitisation;
- supporting material – including exhibition catalogues, books, history files, and learning units;
- user generated content – reactions to the collections (permissions having been gained to publish).

Terminological information

Looking at The Cloud a large component is from terminological resources being used by cultural heritage organisations. These can be the result of international, national, thematic, organisational initiatives. The effort to do this is strong in the library and archive domains. It includes the publication of name authorities. Also this work gives the opportunity for cooperative, possibly international and multilingual, publication, perhaps in the context of EC-funded projects. Topics for terminological publication include: object types; event methods (e.g. creation method); places; organisations; events;

materials; iconography; and many others. The primary advice in choosing what kind of data to publish as linked data is:

- consider publishing information about all aspects of collections and their related materials;
- consider publishing terminological information, and seek partners to cooperate with in order to avoid duplication.

What licence should there be for the linked data

This section deals with the licensing arrangements that are associated with the publication of linked data. Choices made in this are affected by general considerations of how much control the publisher of linked data wants to have over its data, but are also affected by what kind of data is being published. As was seen by the analysis of The Cloud a large part of published linked data does not seem have a licence for its use. The result is that it is unclear what can be done with this data. In these litigious times users are particularly careful not to do anything that will leave them exposed to a possible loss of organisational reputation or even a lawsuit. The primary advice about licensing is:

- any publication of linked data must be accompanied by a licence which makes it clear what uses can be made of the data;
- the licence may be standard, e.g. provided by Creative Commons, or one created specifically by the publisher.

In general terms the two classes for the licence are:

Open licence – This allows any use of the data, especially including commercial use, sometimes with restrictions about attribution and misuse.

Not-open licence – This restricts uses to non-commercial only, with similar requirements for attribution and misuse.

How to publish the linked data

In this area a potential publisher of linked data has three choices to make:

- Which format standards to use;
- RDF serialisations to publish;
- How to link the package into The Cloud.

Which format standards to use

It is inconceivable that they will not use the basic standards like: RDF, RDFS, and OWL. However for the ′descriptive′ formats it is advised to:

- not to create a proprietary format which is only intended to be used for your package;
- use standard format(s) appropriate for the type of data being published. Looking at what is being used a few formats seem to be good suggestions:
  - Web resources: Dublin Core;
  - Persons: Friend of a Friend;
  - Terminological resources: Simple Knowledge Organization System;

  - Bibliographic resources: Bibliographic Ontology;
  - Music: Music Ontology.

These recommendations are based on the current, in-use, formats. However there is a ’gap in the market’ for a format for cultural heritage linked data.
Consider[13] using a cultural heritage specific format for linked data. Possible candidate formats, ones based on: EDM, CIDOC CRM, and LIDO.

RDF serialisations to publish

On the basis of the common practice it is advised that to publish the linked data in the RDF/XML and N-Triples serialisations.

How to link the package into The Cloud

One issue that was brought out by discussions of the WP 2 Working Group was: Which are the ’trusted’ packages in The Cloud? A measure of trust is if one knows the publisher of a package. This type of linking seems to be very common in all parts of The Cloud and leads to the formation of mini-clouds of interlinked packages. There seems to be a cultural heritage mini-cloud forming. A possible reason for this formation is the Europeana initiative. Other very important issues are:

- the identification of resources. Are the identifiers you use compatible with the identifiers used in a potential package to link to;

[13] The Linked Heritage project gives the community an opportunity to look at these possibilities. In particular if offers the possibility of using LIDO. See next section.

- how compatible are the semantics of the packages. For example, if one wishes to identify ’personas’ (public identities), is that the same as FOAF, which says it identifies people.
- a package has to be accessible to queries of it.

Therefore we advise:

- link to packages, of a general nature, which are often linked to: DBpedia; GeoNames Semantic Web; national sources of terminology (e.g. UK Postcodes);
- link to known packages in the cultural heritage, e.g.: Library of Congress Subject Headings; VIAF: The Virtual International Authority File; and Dewey Decimal Classification);
- provide a SPARQL endpoint to the package.

Obviously the final task is to make an entry for the package into The Data Hub registry!

Future Work on linked data

In the next stage of the project work package 2 will be working on two tasks which will show the potential of linked data:

Task 2.3 – Technical specifications

This will specify how cultural heritage information can be enriched by, and can enrich, the ’Cloud’. We will identify: models, processes and technologies which offer the best potential. Selection criteria will include:

- existing use of linked data in cultural heritage and the humanities;

- the use of standards;
- being able to interoperate with other linked data stores. These will include ′major actors′ already identified, such as DBpedia and GeoNames;
- show integration with the technologies selected in other Linked Heritage thematic work packages (i.e. public private partnerships and terminologies);
- maturity and quality of a technical implementation, documentation and support.

Task 2.4 – Enabling linked cultural heritage data

This will demonstrate how to extend existing ingestion procedures to enable content providers to publish their content as linked data, in addition to publishing it in Europeana. The demonstrator will:

- enable content providers to contribute content to the linked data repository and maintain their existing linked data information;
- enhance the ingestion processes with tools for:
  - browsing the linked data repository and its connections to external sources;
  - creating and editing links between entities;
  - extending retrieval to include preferred sources for links and textual information.

GORDON MCKENNA, Collections Trust.
gordon@collectionstrust.org.uk

McKenna, G. ”Linked heritage experience in linking heritage information”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6304. DOI: 10.4403/jlis.it-6304. Web.

ABSTRACT: The Linked Heritage Project started in April 2011. It is funded by the European Commission under the IST Policy Support Programme (ICT PSP), and runs for 30 months. Its main objective is to contribute a large quantity of new content to Europeana, from both the public and private sectors (c3 million items). In addition, the project will show how: 1. To enhance the quality of Europeana content, in terms of its metadata richness, its re-use potential and its uniqueness; 2. To enable improved search, retrieval and use of Europeana content. The author is working in this project, specifically as lead partner in the work package Linking Cultural Heritage Information. This will, amongst other things, be exploring best practice report on cultural heritage linked data and metadata standards. This paper will give some results of the research that has been undertaken. Questions that will be answered include: ”What is linked data? Is all of it Open? Which standards are being used? What use is being made of linked a data in cultural heritage at the moment?”.

KEYWORDS: Library linked data; Linked Heritage Project; Europeana; Cultural Heritage Project

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

DataCite and linked data

Jan Brase

The late Jim Gray from Microsoft Research has introduced the fourth science paradigms in his late work (Hey, Tansley, and Tolle). Thousand years ago science was empirical, describing natural phenomena. The last few hundred years saw a theoretical branch evolving, where models and generalizations were used to understand what was behind these natural phenomena, thus making the shift from the first paradigm to the second, with the scientists no longer being a passive observer, but actively trying to fins out, why things are like they are. The digital revolution in the last few decades allowed a computational branch to grow with the opportunity to use the developed theories to simulate complex phenomena. This was of course the shift from the second paradigm to the third, allowing the scientist to test in detail their theory against their empirical observation.

Today now with the next paradigm shift, we encounter what is labelled as enhanced science or eScience: Data intensive science that unifies theory, experiment, and simulation. This is what Jim Gray defined as the fourth paradigm. Now why is this important for libraries and what are the consequences of this for them? Libraries have a strong mandate of offering access to scientific information and knowledge. The German National Library of Science and Technology has a national mandate to provide scientific information to Academics and Industry in Germany. Secondly libraries have a long

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5493

tradition and experience in doing this, as they are doing so for thousand of years now. This makes libraries trustworthy organization that also have a tendency to be persistent. Especially in the digital age, where more on more information is only available in electronic formats this has become more and more important. While there is always a great risks that current projects and initiatives that create information will not longer be around after a decade, or to be more precise after the funding stops, the chance that the libraries will still be around are much higher. Following the paradigm shifts, information nowadays is more than article or books or any kind of textual information. If we take our mission seriously, we have to widen our mandate to any kind of information that might be relevant for our customers. This includes for example primary data, graphs videos, source code, power point slides, chemical structures among others. And other consequence directly effects us, the dramatic change in the definition of a library catalogue. Traditionally a library catalogue can be seen as a window to the library's holding, a structured summary of what can be brought easily to the shelf. Due to the growth of the internet in the last decades, this has slowly changed and more and more catalogues offer direct access to pdf-versions of document, but the principle has been the same throughout the centuries. Now in the fourth paradigm it becomes more and more impossible for a library to actively store all these kinds of information that are important for its user. Nevertheless the great chance with the growth of the internet is that the library does not have to store this information, when it is available somewhere else in the internet. The libraries job in the future is to know where the information is, if the content provider is trustworthy and to have a distinguished description of the content in its catalogue to offer the service of answering queries from user. In a nutshell, the library of the future should be able to answer the query of a user with the statement: «We do not have

what you are looking for, but we now where it is, and we can offer you a link to it». This implies many aspects: The library has to able to understand what the user is looking for. It has to able to have enough distinguished information about content in its catalogue to know what ideal results would be for the query. The library has furthermore to know where this content is stored and has to provide a persistent link to it to.
Today Technische Informationsbibliothek (TIB) is a global supplier for scientific and technical information, mostly traditional text-based documents. In the last years TIB as the German National Library of Science and Technology has started to actively open its catalogue or to be more precise its GetInfo portal to answer this challenge[1]. Nowadays you can use TIB's central information portal GetInfo as a search tool to access primary data, architectural models and chemical information.

Move beyond text - example

TIB is the German National Library for all areas of engineering as well as architecture, chemistry, information technology, mathematics and physics. It ranks as one of the world's largest specialist libraries, and one of the most efficient document suppliers in its subject areas. GetInfo, a portal for science and technology developed by TIB, bundles access to leading subject databases, publishing house offerings and library catalogues with integrated full text delivery. In doing this GetInfo offers a worldwide unrivalled supply of technical and natural scientific information. At present GetInfo is the only major library portal in Europe to include scientific datasets. The aim is to include all sorts of non-textual information into GetInfo.

The following two examples show data already included:

[1] https://getinfo.de

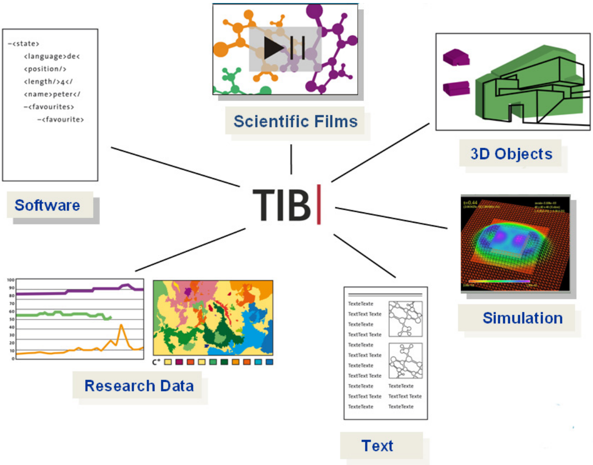


Figure 1: Contents and different ways to access to information

- Library catalogues are classical sources for information (Inger and Gardner). When querying for a certain topic, users might not be interested in only receiving all relebvant publications as a result, but also additional datasets collected by the corresponding researchers. The assignment of persistent identifiers allows this research data to become directly accessible through library catalogues. Nowadays a selection of more than 5.000 datasets that are part of scientific publications are directly accessible through GetInfo (Brase). When the persistent identifier of the dataset is resolved, the user does not directly download megabytes of data but is linked to a preview page where the data center provides metadata and download links to different parts of the data. This workflow is similar to the use of Digital Object Identifier (DOI) names in scholarly journals, where the resolution of a DOI name of an article directs you to a

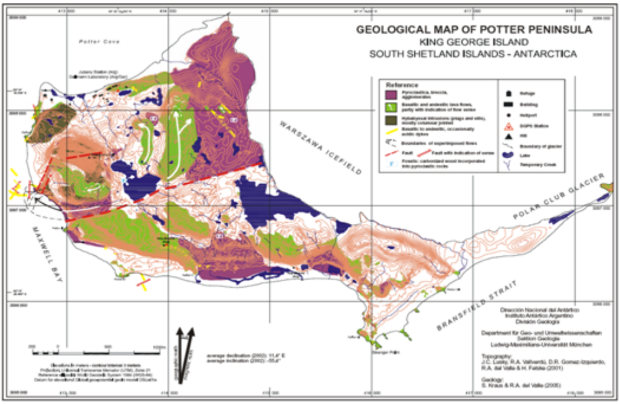


Figure 2: A geological map: non-textual information as rearch result.

publisher's page, including the metadata of the article.

- Another example of non-textual content in GetInfo can be seen in figure 2, where a geological map is displayed as a search result. Again the resolving of the identifier leads towards the preview provided by the specific data that includes the relevant information to this scientific object and displays the download link to the map.

As described earlier, the use of persistent identifiers for stabile linking between the catalogue and the external content is a fundamental requirement for the inclusion of non-textual information in our portal GetInfo. The registration of DOI names for scientific content especially scientific data has furthermore another consequence. Data sets than can persistently be linked to by DOI names become independently citable by other scientists.
Data integration with text is an important aspect of scientific collaboration. It allows verification of scientific results and joint research

activities on various aspects of the same problem. Only a very small proportion of the original data are published in conventional scientific journals. Existing policies on data archiving notwithstanding, in today's practice data are primarily stored in private files, not in secure institutional repositories, and effectively are lost This lack of access to scientific data is an obstacle to international research. It causes unnecessary duplication of research efforts, and the verification of results becomes difficult, if not impossible. Large amounts of research funds are spent every year to re-create already existing data. Encouragingly the "Brussels declaration on STM publishing"[2] focuses on dataset identification as a key element for allowing citation and long term integration of datasets into text as well as supporting a variety of data management activities. It would be an incentive to the author if a data publication had the rank of a citeable publication, adding to their reputation and ranking among their peers.

TIB developed and promotes the use of Digital Object Identifiers (DOI) for datasets. A DOI is used to cite and link to electronic resources (text as well as research data and other types of content). The DOI System differs from other reference systems commonly used on the Internet, such as the Uniform Resource Locator (URL), since it is permanently linked to the object itself, not just to the place in which the object is located. As a major advantage the use of the DOI system for registration permits the scientists and the publishers to use the same syntax and technical infrastructure for the referencing of datasets that are already established for the referencing of articles. The DOI system offers persistent links as stable references to scientific content and an easy way to connect the article with the underlying data. For example:

The dataset: G.Yancheva, N. R. Nowaczyk et al (2007) Rock magnetism and X-ray flourescence spectrometry analyses on sediment

[2] http://www.stm-assoc.org/brussels-declaration/.

cores of the Lake Huguang Maar, Southeast China, PANGAEA doi:10.1594/PANGAEA.587840 is a supplement to the article: G. Ycheva, N. R. Nowaczyk et al (2007) Influence of the intertropical convergence zone on the East Asian monsoon Nature 445, 74-77 doi:10.1038/nature05431.
Since 2005, TIB has been an official DOI Registration Agency with a focus on the registration of research data. The role of TIB is that of the actual DOI registration and the storage of the relevant metadata of the dataset. The research data themselves are not stored at TIB. The registration always takes place in cooperation with data centers or other trustworthy institutions that are responsible for quality assurance, storage and accessibility of the research data and the creation of metadata.

DataCite

Access to research data is nowadays defined as part of the national responsibilities and in recent years most national science organisations have addressed the need to increase the awareness of, and the accessibility to, research data. Nevertheless science itself is international; scientists are involved in global unions and projects, they share their scientific information with colleagues all over the world, they use national as well as foreign information providers.
When facing the challenge of increasing access to research data, a possible approach should be global cooperation for data access via national representatives:

- a global cooperation, because scientist work globally, scientific data are created and accessed globally;
- with national representatives, because most scientists are embedded in their national funding structures and research or-

ganisations.

The key point of this approach is the establishment of a Global DOI Registration agency for scientific content that will offer to all researchers dataset registration and cataloguing services. DataCite was officially launched on December 1st 2009 in London to offer worldwide DOI-registration of scientific data to actively offer scientists the possibility to publish their data as an independent citable object. Currently DataCite has 16 members from 11 countries:
The German National Library of Science and Technology (TIB), the German National Library of Medicine (ZB MED), the German National Library of Economics (ZBW) and the German GESIS – Leibniz Institute for the Social Sciences. Additional European members are: The Library of the ETH Zürich in Switzerland, the Library of TU Delft, from the Netherlands, the L'Institut de l'Information Scientifique et Technique (INIST) from France, The technical Information Center of Denmark, The British Library, the Swedish National Data Service (SND), the Conferenza dei Rettori delle Università Italiane (CRUI) from Italy. North America is represented through: the California Digital Library, the Office of Scientific and Technical Information (OSTI), the Purdue Univerversity and the Canada Institute for Scientific and Technical Information (CISTI). Furthermore the Australian National Data Service (ANDS) is a member.
DataCite offers through its members DOI registration for data centers, currently over 1.3 million objects have been registered with a DOI name.

References

Brase, Jan. “Using digital library techniques - Registration of scientific primary data”. *Lecture notes in computer science* 3232. (2004): 488–494. (Cit. on p. 368).

Hey, Tony, Stuart Tansley, and Kristin Tolle, eds. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft Research, 2009. http://research.microsoft.com/en-us/collaboration/fourthparadigm. (Cit. on p. 365).

Inger, Simon and Tracy Gardner. “How readers navigate to scholarly content”. (2008). http://www.sic.ox14.com/howreadersnavigatetoscholarlycontent.pdf. (Cit. on p. 368).

JAN BRASE, National Library of Science and Technology.
jan.brase@tib.uni-hannover.de

Brase, J. ”DataCite and linked data”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5493. DOI: 10.4403/jlis.it-5493. Web.

ABSTRACT: Science is global, it needs global standards, global workflows and is a cooperation of global players. But science is carried out locally by local scientists that are part of local infrastructures with local funders. DataCite is an international consortium, founded in 2009 of currently 17 institutions from 12 countries worldwide. Its mission is to allow a better re-use and citation of data sets. Over 1 million datasets have been registered with a DOI name as a persistent identifier, so they can be published as independent scientific objects to allow stabile citation of data. Citable data sets can be crosslinked from journal articles, their usage and citations can be measured therefore helping scientists gain credit for making their data available. DataCite offers a central metadata repository with additional linked data service for persistent access to RDF metadata.

KEYWORDS: DataCite; DOI; Data sets

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

Trust and persistence for internet resources

Maurizio Lunghi, Chiara Cirinnà
Emanuele Bellini

Introduction

Internet radically changed our way of working, communicating, living, producing and accessing information, interacting with institutions and bodies, buying things and managing resources. Now everything is available on an open and flexible infrastructure, often freely accessible to all the users: contents are usable by many services tailored to the user requirements. The web has probably been the killer application for the internet. In the past few years, the web moved from a web of documents towards a web of data where information is no more packaged in fixed documents but is available in a de-structured way and usable in a more flexible way by users. The recent developments on the web witnessed the emergence of the semantic web technologies and the linked open data[1] approach, associated with an increasingly large amount of data available for publishing and connecting structured data on the web. Linked data best practices, supported by W3C,[2] are now ready to be endorsed

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

[2] W3C - http://www.w3c.it.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5494

by a relevant number of data providers, leading to the creation of a global data space - the web of data. Unfortunately, the LOD 5 stars[3] are mainly oriented towards the usability and standardization of data published on the web without considering the trustability and persistence of the data and the URI used to refer to them. In fact the objective of the LOD approach seems to be oriented to make a huge number of data on the web accessible in a non-proprietary format (e.g. CSV instead of Excel) and to link these data to other datasets (e.g. Genomes[4] or DBpedia[5]) to disambiguate content and to provide a context. However, in some cases, and especially in the cultural and educational domains, besides retrieving the needed data or their relations, it is also equally important to get information about their authenticity, integrity and provenance. Systems for certification using PIs for digital objects, for authors and for institutions can be of great help in order to refine the quality of information retrievable from internet and to largely increase its usability and the development of potential new services. This paradigm based on the identification and interconnection of data offers solutions to many of the actual library issues, like enhanced web searching, authority control, classification, data portability and disambiguation. In the web of documents identification and trust were provided by web sites and institutions supporting them, in the web of data they are integrated in the single piece of data. The evolution of this paradigm is increasingly important in a vision for the long term curation of the digital resources.

[3] http://www.w3.org/DesignIssues/LinkedData.html.

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

[5] http://en.wikipedia.org/wiki/DBpedia.

Requirements for the long term curation of digital resources

Presently the number of scientific and cultural heritage digital resources made available on the internet throughout digital library applications is constantly growing and it is now crucial to guarantee persistency, authority, reliability and wide dissemination of resources while supporting their long term curation. One of the main requirements to tackle this issue is to adopt credible and PI systems within the life cycle of these resources. A PI should be assigned only to resources that are stable, significant for the related user community and suitable with the scope of the identification system. A number of initiatives, standards, technologies are available, but it may be difficult for an institution to understand which of these are more appropriate for their digital objects. The PI technologies help make stable the reference to a digital resource, even if it is well-known that persistency isn't only a technical issue. In fact these technologies are not obviously reliable per se, no technology can exist indefinitely or guarantee services without a trustable organization and clearly defined policies. In our vision PI systems are meant as the available technology plus a trustable organization and precise policies for digital preservation, implemented by the managers of the related user community. The concept of persistence moves from the commitment of an institution/registration authority to a commitment of the entire user community served by PI. A PI system can be considered as a contract between the final users and the service-providers responsible for the implementation and maintenance of the PI-service and the functionality of the system. From this point of view, the persistence of a PI depends also on the commitment of the community that promotes and uses the identification system for their own resources. This happens when the standard adopted is effectively oriented to the community requirements and the authority in charge to manage the system is recognized by the community itself. It is well known that the structural instability of simple URLs (e.g. domains no longer available) and related resources (relocation or updating) is one of the main issues that prevents the use of internet as a trustworthy platform for the research and the dissemination of digital contents. The current use of the simple URL approach used as persistent digital object identifier brings many and documented risks in a long term vision not only for retrieval and access of resources but also with respect to the loss of reference to the digital documents or the lack of guarantee of authority and provenance. These risks affect:

a) the cultural heritage and research domains, preventing the implementation of reliable citability services, research evaluation, digital preservation, access, etc.,

b) the business domain, preventing the use of purchase services provided on these objects,

c) the public domain (e-gov), slowing down the dematerialization process of public administrations.

It is clear that the problem is not only to face the HTTP 404 error, but it is moving towards identification systems able to support authority, reliability, preservation, certification, exploitation and wide dissemination of these resources. A trustworthy solution is to associate a trusted PI with the digital resources.

The challenge of trust

Trust, broadly speaking, concerns the assessment and management of the risks perceived by each actor entering into a relationship. In other words, ”trust entails risk”. According to the ISO definition, the risk can be defined as the combination of the probability of an event and its consequences (ISO/IEC Guide 73). There are a number of events with bad consequences that could occur during the lifetime of the PI service, with different degrees of probability but all with high costs in case of failure. Examples of these risks are:

a) failing to determine the initial and recurring costs and the pricing of service (risks associated to the financial sustainability),

b) adopting technologies no longer available (risk associated to the standards adoption),

c) the object identified is no longer available on the network (risk associated to the agreement between content and service providers),

d) to lose the support of the community (risk associated to the community mandate), etc.

These factors can determinate the decrease (lowering) of trustworthiness in the PI service by the content provider and affect the dissemination and exploitation of digital resources. The various digital repositories store intangible objects and entities and make them available to users through telematics networks: we access our bank account as well the hospital or the municipality for official documents, we download tons of files and chat with avatar actors. But who certifies the identity of actors and guarantees our privacy? How can we rely on the authenticity of the documents we download? And also how can we trust the institute issuing an ‘official’ document? What is the risk if we cannot demonstrate that a document is not valid for our expected purposes? Which are the risks? A good amount of trustworthiness is necessary to live in this virtual and artificial world. A PI service must address at least the following core requirements:

1. global uniqueness: the PI is clearly part of a name domain and it is unique and associate to a unique resource.

2. persistence: it refers to the permanent lifetime the significant properties of an identifier, for example, it is not possible to reassign the PI to other resources or to delete it.

3. resolvability: it refers to the possibility of retrieving information regarding a resource or to access it directly on the internet.

Currently, there are different technologies and standards for the implementation of PI systems, but there isn't a general agreement on their adoption, often because some of these systems were born as technical solutions, without the support of the community of users who need specific levels of PI services. Systems like the PURL or Cool URIs(Berners-Lee) have considerable advantages in supporting the web of data implementation thanks to the their immediate dereferenceability through the protocol HTTP, but on the other hand, there are several limitations due to the fact that their persistence is not guaranteed in principle by an independent and trustable third party. It is well known that the Cool URI approach to persistence is based on the URL design. This approach, even if it is considered a best practice for the implementation of the semantic web in general and linked data in particular, is mainly based on technical solutions. The basic assumption is that a correct design of the URI should reduce the need to change them in order to ensure their stability over time. An example of this best practice is to avoid the explicit extension of web pages as .php or .asp so that changes in technology implementation do not affect the URI form (e.g. from PHP to ASP). In this perspective, the persistence is based uniquely on the commitment of individual institutions establishing a trusted relationship directly with the final users, without the mediation of a third party. Unfortunately, it is well known that the commitment of a single institution is no longer sufficient to ensure neither long term persistence of URLs nor the trustworthiness of the resources in terms of provenance, authenticity, integrity, conservation, and so on. In practice resources are moving on the network, they can be changed or deleted due to a multitude of factors that cannot always be predetermined or regulated by the content management policies of institutions or governed by best practice techniques. A typical case occurs when an institution runs out because it has been absorbed by another institution, or it is suppressed, or simply its official name has changed. In these cases, the digital objects can be renamed to be adapted to the workflow of the new institution, or transferred to other institutions, or at worst deleted because they are no longer relevant to institutional goals. It is clear that all these actions can cause the breaking of the old URLs independently of how they were built. This may not be a problem if the institution does not handle scientific, cultural or administrative resources but it becomes a critical issue if these changes affect institutions like scientific datastore, libraries, archives, governmental dataset, and so forth. In these cases, for example, bibliographies based on simple URL or even cool URI referring to resources that were present in the archives of these institutions, can no longer be used to check the scientific work or to calculate bibliometric indexes. Another critical issue is related to the connection of datasets which have been updated several times. In such cases, it may be difficult or even impossible to verify the validity of the scientific outcome presented in a related paper. What is most critical, however, is the impossibility to implement systems to check the authenticity, provenance and integrity of these resources because of the absence of a third party able to guarantee the association name - resource. In this scenario, most benefits of a wide access to linked dataset are dissolved by the lack of their reliability.

NBN:IT service as a support of trust LOD

To tackle the challenge of trust in LOD, a possible solution could be to adopt a URN based PI solutions.[6] Presently, to implement a PI system, the main approach is to separate the identification from the localisation of the resources. As shown above, Tim Berner Lee advises that adopting clear and stable policies and implementation guidelines is sufficient to manage the persistent identification of resources on the internet. Even if this suggestion is reasonable and appropriate in some domains, it is evident that we cannot delegate this responsibility to each institution, in particular in the scientific and cultural heritage domain for two main reasons:

1. many institutions fail to decide the approach and the strategy to be adopted in terms of content selection, formats, naming, etc.;
2. many institutions fail to decide the approach and the strategy to be adopted in terms of content selection, formats, naming, etc.;

In any case, Uniforum Resource Identifiers (URIs) are widely used in the semantic web context to identify any type of resources or any real, digital, abstract, virtual object, trying to harmonise in a semantic vision all the user communities applications. For instance, to address this issue, the info-URI scheme[7] was developed by libraries and publishing communities for "URIs of information assets that have identifiers in public namespaces but have no representation within the URI allocation". It is clear that, in order to refer to a certified digital object in a trustable way, the use of URN or identifiers

[6] APARSEN DE22.1 Persistent Identifiers Interoperability Framework - http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=D22.1+Persistent+Identifiers+Interoperability+Framework.

[7] RFC 4452: http://info-uri.info.

that implements the RFC 1737 (Functional requirements for Uniform Resource Names) is today a best practice. The purpose of a URN is to provide a globally unique, persistent, location-independent resource identifier which can be used for the identification and access to the characteristics of a resource or for the access to the resource itself. The URN specification is part of the IETF family of specifications encompassed by the URI framework. This framework also includes URLs, which specify both a protocol and a location in order to give access to resources on the web. IANA is the registration authority for URN namespaces. URNs are designed to enable heterogeneous namespaces mapping onto a URN-space, and therefore enable the reuse of well-known identifiers. Unlike URLs, URNs are not directly actionable (browsers generally do not know what to do with a URN) because they have no associated global infrastructure that enables resolution (such as the DNS supporting URL). Although several implementations have been made, each proposing its own means for resolution through the use of plug-ins or proxy servers, an infrastructure that enables large scale resolution has not been implemented. But single implementations of namespace, like the URN-NBN or the DOI, offer a resolution-service available on internet. The NBN namespace, as a namespace identifier (NID), has been registered and adopted by the Nordic Metadata Projects but is being separately implemented by individual systems with no reference implementation which enable the coordination of information sources. In fact, several national libraries have developed their own NBN systems within national projects; several implementations are currently in use, each with different descriptive metadata or granularity levels. According to this, it is clear that the PIs, cannot support the LOD trustworthiness successfully. The NBN-Italy service supports at least three levels of persistence:(Bellini et al., "The National Bibliography Number Italia (NBN:IT) Project. A persistent identifier supporting national legal deposit for digital resources")

1. *Persistence of the identifier NBN.* If the resource is no longer available online, the URN identifier will be maintained (e.g. as proof that at some point that resource has existed);

2. *Persistence of the association URNs and URLs.* It is a commitment that ensures that in the long term URN is resolvable (which leads at least to an address of URL type). The accessibility to the resource is not guaranteed but is assured the access to the the so-called "Tombstone" if the resource is no longer available on the network (e.g. "This ebook is no longer on the market");

3. *Persistence of the resource referenced by NBN.* Ensuring long-term existence and accessibility the resource referenced by URN. This is the level of persistence of NBN made possible thanks to the storage (statutory or voluntary) at the national libraries and authoritative description of the national bibliography.

Thanks to these levels of service, NBN-Italy names represent a clear added value if used in the LOD architectures to support the trustworthiness of the assertions (RDF triple). This proposal goes towards the integration of the LOD and PI systems, by exploiting the on-going initiatives and projects as outlined in the next paragraph.

Next steps: Den Haag Manifesto 2.0 and Florence Agenda

The forthcoming event "Cultural Heritage On Line 2012" that will be held in Florence in December 2012 aims to improve and make effective the "Den Hague Manifesto" through the union of several ongoing related initiatives, projects and stakeholders like: APARSEN

NoE,[8] Datacite,[9] EPIC,[10] and PersID[11]/URN-NBN, W3C5, Knowledge Exchange,[12] and so forth. Two of the major objectives that we are going to achieve are:

1. a review of the Den Haag manifesto and its improvement towards the 2.0 version.
2. the definition of a Florence Agenda to define a common strategy for a Trusted LOD implementation

Den Haag Manifesto 2.0

In the recent developments some initiatives are merging the open approach of the linked open data and the potentiality of the semantic web with the added value of identification, authenticity, and provenance offered by the PI systems. The Knowledge Exchange organised a seminar[13] on persistent object identifiers inviting various current practices to compare services and explore future cooperation and convergence. This seminar took place on 14-15 June 2011 at the DANS offices in The Hague and was hosted by PersID, SURF foundation and DANS. Three major players in the persistent object identifiers area, Datacite/DOI, EPIC/Handle and PersID/URN-NBN, informed each other about recent developments, shared user experiences and discussed trends and policies. In break-out sessions participants discussed the benefits and challenges in operating multiple

[8] APARSEN - http://www.alliancepermanentaccess.org.

[9] Datacite - http://www.datacite.org.

[10] European Persistent Identifier Consortium (EPIC), http://www.pidconsortium.eu.

[11] PersID- Building a persistent identifier infrastructure, http://www.persid.org.

[12] Knowledge Exchange http://www.knowledge-exchange.info.

[13] Knowledge Exchange, http://www.knowledge-exchange.info/Default.aspx?ID=440.

PI systems and the relation of PIs to linked open data communities: there was a clear interest in connecting the PI systems to the linked data standards. This led to the ”Den Haag Manifesto”(DHM), which outlines a series of concrete actions to join the PID and Linked Open Data communities. FRD has participated in the working group to define opportunities for collaboration between LOD and PI systems. During the meeting a sort of ”cultural gap” between the LOD and the PI community came up. The major differences concerned the concepts of identification, persistence and trustworthiness. In fact, the LOD approach is strongly oriented to the representation of the information flow on the web. In this view the resource can change over time according to the workflow of the publication. For instance, a dataset can be updated on the web several times while its URI can remain the same. With an opposite vision, the PI domains are more oriented to identify stable resources managed by systems of trusted digital repositories. During the work we tried to identify the main characteristics of the IP systems that can be imported in LOD. The results of this first assessment was the definition of a 5- point manifesto that morally committed the institutions working in the domain of PI and LOD to ascertain their possibility of integration. The points raised are:

1. A PIs can be an http URIs including content negotiation.

2. Using LOD vocabularies for diagram elements.

3. Identifying a minimum set of common elements across space identifiers in scholarly (examples are DOI kernel metadata, DataCite kernel, etc.).

4. To use ‘same as’ to help PI interoperability.

5. To use PIs for subjects and objects in the RDF triples.

Since then, the DHM is used as the basis for a co-ordinated approach to identifier issues across the PI and LOD communities, but starting from these points, the DHM has to be revised, specified and extended according to present trends and solutions. Moreover, it has to be supported by a shared agenda able to guide the forthcoming LOD and PI implementations, in order to have harmonized and interoperable solutions: the Florence Agenda.

A proposal for a Florence Agenda

Presently FRD is leading a specific work package (WP22) that is dealing with PIs interoperability and LOD within the APARSEN EU project. The APARSEN is a Network of Excellence of 34 institutions and is co-funded by the European Commission in order to fight the fragmentation of digital preservation of scientific records in Europe. In the first year the WP22 developed a reference model for interoperability of PI existing systems. The work started with identifying some basic user requirements for identifiers for digital objects, persons and bodies, then some criteria for trusted PI systems have been agreed. Finally an interoperability framework has been proposed where any trusted PI system can expose its data through a shared schema; the model proposes an ontology for interoperability of PI systems in line with the LOD approach. The Italian NBN initiative follows the same flow. The NBN project is leaded by the Italian legal deposit[14] consortium that has defined some criteria and guidelines to assign the PI. This defined workflow in conjunction with the commitment of the national libraries of Florence, Rome and Venice that manage such service assure the level of trust to the PI generated that, through its reuse in the LOD domain, enables the T-LOD implementation. The Florence Agenda aims to identify some

[14] www.depositolegale.it.

milestones, guidelines and criteria that can be adopted by the PI and LOD communities to cooperate to build a more reliable web of data.

References

Bazzanella, B., et al. *Persistent Identifiers Interoperability Framework – Alliance for Permanent Access to the Records of Science Network APARSEN*. 2012.

Bellini, E. and M. Lunghi. *Persistent Identifiers for cultural heritage*. Briefing paper – digitalpreservationeurope (DPE - EU project). 2007.

Bellini, E., et al. "Persistent Identifier Distributed System for Digital Libraries, Information Technology". *World library and information congress: 75th IFLA general conference and council*. Milan, Italy, 2009.

Bellini, E., et al. iPres2008 conference proceedings. 2008.

——. "Semantics-Aware Resolution of Multi-part Persistent Identifiers". *Emerging Technologies and Information Systems for the Knowledge Society*. Springer Berlin-Heidelberg, 2008. 413–422.

Bellini, E., et al. "The National Bibliography Number Italia (NBN:IT) Project. A persistent identifier supporting national legal deposit for digital resources". *JLIS* 3.1. (2012). (Cit. on p. 383).

Berners-Lee, Tim. "Cool URIs don't change". *W3C Design Issues* 1. (2009). <http://www.w3.org/Provider/Style/URI.html>. (Cit. on p. 380).

CENL. *CENL Task Force on Persistent Identifiers, Report 2007*. 2007.

——. *The National Libraries Resolver Discovery Service (RDS) - CENL Recommendation*. 2007.

Daigle, L., et al. *URN Namespace Definition Mechanisms*. 1999.

DPE. *Winer Dov, Persistent Identifiers systems in the Public Administration sector, Digitalpreservationeurope (DPE) Briefing Paper*.

Hakala, J. *Using National Bibliography Numbers as Uniform Resource Names "NBN". RFC 3188, 2001*. 2001.

ISO 3297:1986: Documentation – International standard serial numbering (ISSN). Geneva: International Organization for Standardization,

Luhmann, N. "Trust: a mechanism for the reduction of social complexity". *Trust and Power*. Wiley, 1979. 4–103.

Lynch, C., C. Preston, and R. Daniel. *Using Existing Bibliographic Identifiers as Uniform Resource Names*. 1998.

Masinter, L. and K. Sollins. *Functional Requirements for Uniform Resource Names (RFC 1737)*.

Moats, R. *URN Syntax*. 1997.

NESTOR. *Catalogue of criteria for assessing the trustworthiness of PI systems*.

NISO/ANSI Z39.56-1997 Serial Item and Contribution Identifier. Baltimore, MD: National Information Standards Organization,

NISO/ANSI Z39.9-1992 International standard serial numbering (ISSN). Baltimore, MD: National Information Standards Organization,

NISO/ANSI/ISO 2108:1992: Information and documentation - International standard book number (ISBN). Geneva: International Organization for Standardization,

Sollins, K. *Architectural Principles of Uniform Resource Name Resolution (IETF RFC 2276)*.

Wilson, Frank. *Think Paper 11: Trust and Identity in Interactive Services: Technical and Societal Challenges*.

MAURIZIO LUNGHI, Fondazione Rinascimento Digitale.
Lunghi@rinascimento-digitale.it

CHIARA CIRINNÀ, Fondazione Rinascimento Digitale.
Cirinna@rinascimento-digitale.it

EMANUELE BELLINI, Fondazione Rinascimento Digitale.
Bellini@rinascimento-digitale.it

Lunghi, M., C. Cirinnà, E. Bellini. ”Trust and persistence for internet resources”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5494, p. 375–390. DOI: 10.4403/jlis.it-5494. Web.

ABSTRACT: Internet has changed our way of working, communicating, living, producing and accessing information, everything available on an open and flexible infrastructure accessible to all the users mainly free of cost. However in some cases, it's not only important to find information but also having information about its authenticity, integrity, provenance and relations with other pieces of information. Systems for certification using URN technology like the persistent identifiers for digital objects, for authors and for bodies can extremely help in order to refine the quality of information retrievable from Internet and to increase largely its usability and potential development.

KEYWORDS: NBN:IT; Persistent identifier; National Bibliography Number — Italy

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

Linking library metadata to the web: the German experience

Gabriele Meßmer

«What is the value of a catalogue of more than 23 million records?»was one of the questions we discussed when starting the linked open data project at the Bavarian State Library. Many generations of librarians have been doing a good job creating machine-readable catalogue records, nowadays called metadata, with the purpose of describing books, printed music, manuscripts and maps, of building up authority files, listing holdings and more. To increase the value of these expensive data, it is essential today not only to offer catalogue interfaces to retrieve information, but also to open the catalogue databases and to give free access to the records.

In Germany some libraries of the North-Rhine-Westphalian library network were the first to publish their records as linked open data (LOD) in March 2010. At the same time, the hbz created a LOD website. Since then more and more libraries have started to discuss the issue of open and linked open data as well as the question of making their data freely accessible. In Germany there are six library networks running five different union catalogs. Although the German cataloging networks and the libraries have worked closely together for many years, sharing common authority files as well as exchanging records for re-use, there are still in many cases different records for one and the same book in the various union

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5507

catalogs.

Because of the special situation of five different coexisting library network catalogs, there is no uniform identifier for catalogue records in Germany. Many records already have an OCLC number that serves almost as such an identifier, but this is by no means the case for all records. Sometimes records describing the same book may have two or even more different identifiers. After starting the LOD projects in Germany, it became immediately obvious that a common and persistent identifier was needed for every title record and ideally only one single identifier for different records describing one and the same resource. So Culturegraph[1] was born, a linked open data service with the aim to generate a specific identifier for all kinds of objects held by libraries in Germany. This identifier should be used to reference the description of various objects. It should have a defined syntax, must be unique and persistent.

In the first step, the German library networks provided records for monographs and multi-part works published after 1945 to be ingested into the Culturegraph database. Then the records were compared and bundles (clusters) were created with records which – although slightly differing – were supposed to belong together and described the same object. A number of identifiers served as match criteria such as ISBN, ISMN, OCLC number and others. By now a resolving and look-up service is available to retrieve these single titles or bundles in Culturegraph. The next step will be to establish a larger database with all records held in the German library networks, a task which is especially challenging when it comes to early printed books with long baroque titles.

The North-Rhine-Westphalian Library Service Center (hbz) commissioned the Berlin lawyer Dr. Till Kreutzer to analyse the legal aspects of open library data. In 2011 he published a guideline (*Open Data*

[1] http://www.culturegraph.org.

– *Freigabe von Daten aus Bibliothekskatalogen*) which is a good basis for all legal questions around open data. The guideline contains chapters about the protectability of data, especially of single data fields. It examines databases and collections of data such as catalogs respectively and it discusses the possibility to release catalogs and publish records as open data.

The first linked open data service of the German National Library was the publication of the authority file for personal names and the authority file for subject headings in 2010. Today, the German National Library makes available three data sets: title records of the main collection (without records for printed music), the German Union Catalogue of Serials (Zeitschriftendatenbank, ZDB) and the complete German integrated authority file (Gemeinsame Normdatei, GND). The data model for the bibliographic data is documented in a paper (*The Linked Data Service of the German National Library: Modelling of bibliographic data*), which is available on the web also in English. All data sets are published under a Creative Commons Zero (CC0) license.

The B3Kat project – open data

In 2010 the Cataloging and Metadata Commission of the Bavarian and the Berlin-Brandenburg library networks also started to discuss open and linked open data. The research libraries of Bavaria, Berlin and Brandenburg use a common catalogue database called B3Kat. This catalogue contains more than 23 million records of 180 member libraries. The records are held in MAB, the special German data exchange format, and are linked to records of the German authority file. A small working group was established in order to achieve quick results. At first the working group identified reasons for having open catalogue metadata.

- to make local and regional data visible worldwide, i.e. to have them no longer hidden in thedeep web;
- to complete and increase the value of already established web presences such as Wikipedia by giving links to authoritative resources;
- to provide data for newly developed web-based services;
- to integrate data into the semantic web with the possibility to re-use the completed and enriched data one's own catalogue environment;
- to contribute to and to actively promote the open access movement.

In the project part A an OAI PMH repository was established in the Aleph500 environment of B3Kat. The title records held in the Aleph system and structured according to the German exchange format MAB were converted to MARCXML and then provided as open data. The particular challenge was on the one hand to map as many MAB fields as possible to MARCXML, in order to include a maximum of information, and on the other hand to include the basic information about the owning libraries. As the data pool contains more than 23 million records – a huge amount of data – two different methods are offered to pick up the records: there is the complete data set, split in three parts, frozen at a certain date, and an OAI repository comprising all continuously ongoing updates of records or newly created records.[2]It is possible to download the whole data set or to select only the records of a specific library, to select a single record, if the B3Kat ID number is known, or to obtain defined sets of records

[2] Information about the open data pool can be found here: https://opacplus.bib-bvb.de/TouchPoint_touchpoint/help.do?helpContext=opendata_en.

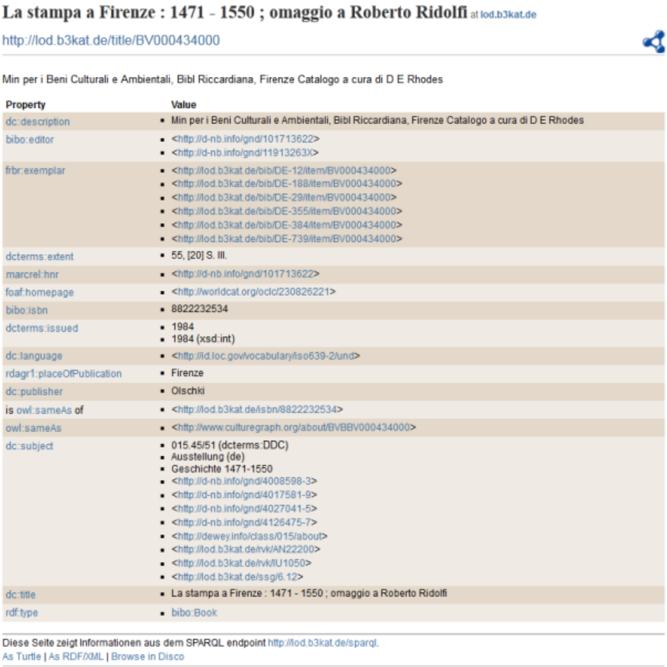
which can be identified because they include certain codes or fields. The complete data set will be published twice a year. As of March 2012 the complete set or parts of the set have been downloaded by this approach more than 400 times.

The B3Kat Project – linked open data

This MARCXML based open data pool serves as a basis for the linked open data pool (part B of the project). This was the easiest way for the next step, the transformation to RDF, because there are already tools for this process. Many fields provided in the MARCXML format had to be mapped to the RDF data model and published as RDF data. Wherever possible URIs are being used. Therefore for every title record a uniform resource identifier (URI) was created, based on the B3Kat identity number (starting with BV) and the name space reserved especially for B3Kat `lod.b3kat.de`. This name space was registered at DENIC, a registry for German domains under the top level domain `.de`.

To link the data to other data as well as possible many more links were implemented which were provided by the particular content of our records: for example links to the German integrated authority file, to WorldCat, to the language code ISO 639-2, to the Library of Congress subject headings and to the Dewey Decimal Classification. Currently, the RDF data pool consists of about 600 million RDF triples. As of March 2012 the pool has been downloaded more than 680 times and had more than 7600 visits. For the time being the linked open data set is published only bi-annually, while the updates for the open data are continuously provided. Information about the RDF data set, the data model, the used ontologies and the SPARQL endpoint can be found on the project webpage.[3]

[3] http://lod.b3kat.de. Currently this page is only in German.



La stampa a Firenze : 1471 - 1550 ; omaggio a Roberto Ridolfi at lod.b3kat.de

http://lod.b3kat.de/title/BV000434000

Min per i Beni Culturali e Ambientali, Bibl Riccardiana, Firenze Catalogo a cura di D E Rhodes

Property	Value
dc:description	Min per i Beni Culturali e Ambientali, Bibl Riccardiana, Firenze Catalogo a cura di D E Rhodes
bibo:editor	<http://d-nb.info/gnd/101713622> <http://d-nb.info/gnd/11913263X>
frbr:exemplar	<http://lod.b3kat.de/bib/DE-12/item/BV000434000> <http://lod.b3kat.de/bib/DE-188/item/BV000434000> <http://lod.b3kat.de/bib/DE-29/item/BV000434000> <http://lod.b3kat.de/bib/DE-355/item/BV000434000> <http://lod.b3kat.de/bib/DE-384/item/BV000434000> <http://lod.b3kat.de/bib/DE-739/item/BV000434000>
dcterms:extent	55, [20] S. Ill.
marcrel:hnr	<http://d-nb.info/gnd/101713622>
foaf:homepage	<http://worldcat.org/oclc/230826221>
bibo:isbn	8822232534
dcterms:issued	1984 1984 (xsd:int)
dc:language	<http://id.loc.gov/vocabulary/iso639-2/und>
rdagr1:placeOfPublication	Firenze
dc:publisher	Olschki
is owl:sameAs of	<http://lod.b3kat.de/isbn/8822232534>
owl:sameAs	<http://www.culturegraph.org/about/BVBBV000434000>
dc:subject	015.45/51 (dcterms:DDC) Ausstellung (de) Geschichte 1471-1550 <http://d-nb.info/gnd/4008598-3> <http://d-nb.info/gnd/4017581-9> <http://d-nb.info/gnd/4027041-5> <http://d-nb.info/gnd/4126475-7> <http://dewey.info/class/015/about> <http://lod.b3kat.de/rvk/AN22200> <http://lod.b3kat.de/rvk/IU1050> <http://lod.b3kat.de/ssg/6.12>
dc:title	La stampa a Firenze : 1471 - 1550 ; omaggio a Roberto Ridolfi
rdf:type	bibo:Book

Diese Seite zeigt Informationen aus dem SPARQL endpoint http://lod.b3kat.de/sparql.
As Turtle | As RDF/XML | Browse in Disco

Figure 1: Esempio di un record in B3kat.

Both the open data and the linked open data pool of B3Kat went live in the first days of December 2011. At the moment it is the largest bibliographic record or title set available in Germany.

Legal aspects

An important topic in the discussion about open data is the legal aspect. The working group had to consider questions such as:

- Will service providers agree with their records being published?
- Will all libraries accept the publication of their records?
- Are there fields/tags that should not be published, e.g. subject headings or URLs?
- What about catalogue enrichment, e.g. abstracts or table of contents integrated in the records?
- Under which license should open data be provided?

Some German libraries decided not to publish the full records as open or linked open data, but to omit from the open data some fields such as URLs. Some librarians believe that particularly expensive parts of a record, like subject headings, should not be published for free. The Bavarian-Berlin-Brandenburg working group however recommended to publish the records as completely and fully as possible – in order to make them really meaningful for all interested parties and to make sure that this service is also beneficial for the general data exchange between libraries and networks. For the time being only the URLs linking to table of contents purchased from commercial service providers cannot be published for copyright reasons.

Before publishing metadata it is also necessary to consider and to define the legal conditions for their reuse. There are two models:

- to publish data under a special license or
- to waive any rights.

To really comply with the concept and prerequisites of Linked Open Data (LOD)it is necessary to provide these data without any restrictions under a completely free license. The B3Kat records are therefore published under CC0 Universal Public Domain Dedication, which is also used for the metadata in the Europeana context. This allows the maximum use of the records and the provider has no administrative overhead to control the licenses of the users.

Conclusions

The project has come to an end, but it is not finished. There are still things to do: both project webpages must also be published in English. The license information must be integrated into the records, the MARCXML records as well as in the RDF ones. Furthermore, an update process is needed to keep the linked open data up-to-date as well, which implies to handle corrected and deleted records. A request often articulated is to publish the complete set of 23 million records in the MARCXML based OAI PMH repository, not split it as it is in three parts. Up to now this could not be realized because of performance and hardware issues. Both parts of the project were successful. The implementation was realized quickly, the sets are frequently in demand and we learned a lot about publishing metadata in MARCXML (our experience so far has been predominantly with the German MAB format) and about doing this in an OAI repository. Technicians and catalogers had to work hand in hand to get the best out of the existing data. Publishing open data is no job to do on the fly, alongside the daily work. It needs time and money, because a sound calculation is necessary to account for staff cost, for the hard- and software needed and above all for the time required to hold the data up-to-date.

Providing open data may also mean a shift in data management inside the library community. Until now, delivery of data for different purposes was up to the database provider and a lot of work had to be done for different file definitions and transfers, always up to the sender, not the recipient. With Z 39.50 and even more with OAI it was still up to the data provider to define the method for the delivery of data and thus the structure and the fields to be supported. With open data and even more linked open data it is now up to the user to make the relevant choices and selections and even a re-modelling of records on their side. Nevertheless the learning curve in analyzing the data provided in order to make best use of them is still to be followed and a lot of standardization and harmonization of formats and contents is still to be done in order to make the use of open data a smooth method of library cooperation and record reuse.

Until late summer 2012 yet another OAI repository will be established in the context of the Europeana Libraries project in which the Bavarian State Library is one of the partners. One of the outcomes of the Europeana Libraries project is the *Report on the alignment of library metadata with the Europeana Data Model* (*EDM*). This repository will also use the open MARCXML data, but it will only contain metadata of digitized objects. These metadata will be enriched with links to thumbnails. Not least this repository – also in MARCXML – will serve as a data pool for German and European portals which present metadata for all kinds of digital materials. The expected advantage of EDM is to enrich the records in Europeana and thus make them interoperable and fit for the semantic web. With EDM for Europeana and CIDOC CRM, the common data format for the future German Digital Library we can clearly see which further requirements come up when dealing with digitized information. It is not only a thumbnail – it may be all images (surely not in the original high resolution, but in a lower one) and the structural metadata as transported in METS/MODS files and finally the full text information.

Linked open data will lead to new services which will be developed in the near future. They heavily rely on a lot of basic knowledge of librarians: metadata, structures, normalization, quality control, standard numbers etc. The librarian of today is no longer a cataloger, but a metadata specialist and a manager – working on how the rich information contained in library records can be most usefully exploited and integrated into the web.

References

Kreutzer, Till. *Open Data – Freigabe von Daten aus Bibliothekskatalogen*. Köln: Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen, 2011. http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/open-data-leitfaden.pdf. (Cit. on p. 392).

Report on the alignment of library metadata with the Europeana Data Model. 2011. http://www.europeana-libraries.eu/documents/868553/1eade085-34ac-487f-82af-d5cd2545e619. (Cit. on p. 399).

The Linked Data Service of the German National Library: Modelling of bibliographic data. Leipzig: Deutsche National Bibliothek, 2012. http://www.dnb.de/SharedDocs/Downloads/EN/DNB/service/linkedDataModellierungTiteldaten.pdf?__blob=publicationFile. (Cit. on p. 393).

GABRIELE MESSMER, Bayerische Staatsbibliothek.
messmer@bsb-muenchen.de

Meßmer, G. ”Linking library metadata to the web: the German experience”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5507. DOI: 10.4403/jlis.it-5507. Web.

ABSTRACT: One of the major challenges of libraries today is to make metadata available for the usage and re-usage by researchers and the scientific community. Therefore it is necessary to open the cataloguing systems for non-restricted and completely free access. Libraries of Bavaria, Berlin and Brandenburg decided in 2011 to publish their shared network catalogue with nearly 23 million records as open data and as linked open data. In March 2012 this data pool won the second prize in the first German-wide programming competition ”Apps for Germany”. The paper presents the steps of the project and the versatile experiences in publishing the data of more than 170 libraries. In addition it will introduce the Europeana libraries project in that more than 5 million records among them 600.000 records of the Bavarian State Library will be ingested in Europeana and be published as linked open data.

KEYWORDS: Library linked data; Semantic web; RDF; B3Kat Project

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

Linked open data for new library services: the example of data.bnf.fr

Romain Wenz

Library catalogues were designed to locate books and to handle collections. They are used by librarians collecting books and by users finding them. Yet, it can be hard for a user to reach library information on the web, especially as there can be several catalogues for one library. Indeed, different kinds of tools are required for different kinds of collections. For instance, a collection of archives and manuscripts needs a hierarchical structure, to describe documents together, as they were produced and received during the activities of a person. Therefore managing documents can be different goal from making access to them. Web users have new expectations and new habits in a changing web environment. Library data should meet these needs and truly belong to the web. Libraries try to make their data really useful on the web. We will focus on the use case of data.bnf.fr, a project from the Bibliothèque nationale de France that relies on efficient links, automatic techniques and semantic web tools.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5509

New expectations

Online catalogues make things different. With the world wide web, researchers have access to plenty of resources from a single computer, even from home. There is now some kind of competition between document providers, since it is a lot easier to switch from one to another. For instance, a copy of book will be less needed once it is digitized and available online. The theories of Walter Benjamin showed that content is losing its value once it is copied with industrial processes, which is very true in the digital world. On the other hand online access creates another kind of value, at least for cultural and educational resources, which are meant to be spread. The resources provided by libraries have to be easy to find, because they become part of a more general "web search". There are always more documents online: many specific websites provide information which can be compared to what can be published in books. Moreover, digital collections published by libraries become part of the web. For readers who are looking for resources on the Internet, texts from digitized books provide information, like other web pages. That is why digital collections, but also online references of physical books, have to be easy to find and accessible through automatic programs. The general public can find some documents without even knowing they exist. Typically, with the use of powerful search engines, users now commonly search with keyword associated with the final document. This habit was spread in a decade. It has made all kinds of online information always easier to find, through the use of search engines. This implies that users tend to search with keywords associated together, as opposed, for instance, as using a series of fields as in catalogues. It also means that results sorted by algorithms are commonly accepted. We are all familiar with sentences such as "results 1 to 10 on 120000": noise is not a problem, if it comes after the relevant results, found automatically and presented first. How should libraries take advantage of these evolutions? Several sources of information can help us, in order to decide how to adapt. First, the statistics of our local search interfaces provide accurate and free feedback on what our traditional users search. For instance, some years ago people used to search for "complete works" of writers, knowing in advance what they would find in the book. Now, we mainly have searches for the books themselves, typically with the title of the book and name of the writer. Public surveys from by the libraries or other institutions show that, using search engines for browsing the web has now become a habit. Internet users usually find bibliographic references online before going to the physical library. This is confirmed by all the user surveys made those past few years, for students as well as for researchers, as the ones made by OCLC,[1] and by the Bibliothèque nationale de France.[2] Therefore, book references that are impossible to find online are almost useless for most people. If librarians want them to be found, they have to put those references on the web. Most catalogues are available online with a specific portal. But they are usually not accessible from web link, and impossible to crawl for search engines. Those new expectations from the public are essential for libraries, because of the size of the content owned by libraries. The amount of content and information available is so huge that the most recent techniques have to be used to handle them. For instance, the Bibliothèque nationale de France displays 1,5 million objects in Gallica,[3] which is the biggest French-speaking digital Library, and 12 million bibliographic records, thanks to the legal deposit of the French edition. Thousands of manuscripts and archives are also available, with all types of resources, from medieval manuscripts to archival fonds of modern writers. Handling

[1] For instance http://www.oclc.org/reports/onlinecatalogs.
[2] For instance http://www.bnf.fr/documents/enquete_gallica_2011_rapport.pdf.
[3] http://gallica.bnf.fr.

this kind of resources creates several scale issues, as we are dealing with millions of documents. There are always duplicates, and the quality of the data is irregular, as a result of the long history of our catalogue. Moreover, printed books and manuscripts are usually described with various logics, inside the catalogues. Records from the main catalogue describe a physical book, usually in a MARC format. They are structured deliberately around a collection which was constructed on purpose, with a series of books that would be shelved together and make sense for the end-user. On the other hand, archival were produced and received during the activities of a person, and considered in a way as "by-products" of the life and activities of some person or organisation. The documents were gathered according to this logic, which is not always obvious for the end-user today. Therefore, the documents cannot be described with simple "records", but with the model of a hierarchical tree, which makes it possible to understand the original logic of the archives. The format which is commonly used for this kind of resources at the Bibliothèque nationale de France is XML-EAD (Encoding Archive Description). The digital collections, available in Gallica, are described with a simple format: Dublin Core. All digital items are accessible with a persistent identifier (ARK), given and maintained by the Bibliothèque nationale de France. Between these catalogues, efficient links have to be provided, so that the users can browse quickly and go simply from one document to another. Machines are not intelligent, so it is necessary to provide structured information in the catalogues, with efficient links between the documents.

Importance of efficient Links : principles in data.bnf.fr

If we want the resources to be truly part of the web, in the sense that users can quote them on sites, blogs, pages, and e-mails for instance, and access them by following links, we have to give them proper identifiers, and to comply with web standards. Thus it is also possible to link resources from our different datasets. As big libraries often have several catalogues, making links between them makes it possible to find resources without having to learn how to use all the different tools, just by "following one's nose". It makes it possible to handle library data at large scale, with different types of documents. This is very important since many distinctions between documents were made before the web. For instance, for the end-user, a digitized "regular book" and a digitized medieval manuscript can be equal, in the sense that the same user can access them in the same way if they are online. The very notion of "special collections" can change if they are digitized and available on the web. This form of openness is accomplished through digitization processes. In the context of digitization, many resources which were interesting only for specialized scholars have become relevant for a broader public. For instance, medieval miniatures are surprisingly used by a very broad public once they are online. The way to search has to be simple for these resources, open to the web with digitization. They have to be easy to find inside databases but also available through links on the world wide web. In general catalogues as well as for digital collections metadata, the data describing documents has to be technically available, but also legally re-usable if we want it to be broadly spread. This is why many libraries move to the techniques of the semantic Web, together with licenses of "open data". This way, some libraries are part of the "linked open data" movement and are involved in the development of the "semantic web". The Bibliothèque nationale de France develops a new project, bringing together data from catalogues (MARC), archives (EAD) and digital resources (DC). All the data are extracted and gathered automatically. This project, called data.bnf.fr,[4] is still a young project, as it has been online for a year. Data.bnf.fr gathers descriptive information, and links directly to online catalogues and digital documents. There are several aspects: a first goal is to make the information compliant with the "semantic web" standards, by providing persistent identifiers for the resources, with a RDF view on the available information. For the library, gathering information around concepts of works, writers and subjects also implies to work on modelling issues. In fact, it is a first opportunity to implement the FRBR model, and to use it with automated matching and alignments. To do so, we use a free software, called CubicWeb.[5] This is not only a technical issue, but also a way to get a first feedback on what is possible, and how users react. Therefore, it is very important to publish both structured data for computers, and web pages, quick and easy to use for humans.

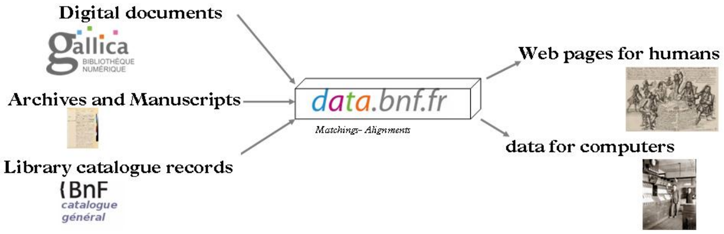


Figure 1: data.bnf.fr.

[4] http://data.bnf.fr.

[5] Site and documentation at http://www.cubicweb.org and http://docs.cubicweb.org.

So as to comply with the "semantic web" principles, it is essential to provide information which is described with common vocabularies, with a strict structure, the way they would be inside a database. This can only be done with identifiers given to all the concepts that have to be handled. The Bibliothèque nationale de France uses persistent identifiers for web URIs: the "ARK"[6] identifiers, which are used to identify catalogue records, archival resources, digital objects from Gallica, and authority records. These ARK identifiers are also useful for quoting these resources, with a common "resolver": for instance, a digital object[7] will also be accessible with a persistent link.[8] In order to gather information about Works, Writers and Subjects, data.bnf.fr relies on the authority files. The ARK identifiers that have been given to the authority records are also used for the pages in data.bnf.fr, so as to build reliable URIs, and efficient links between the data.

Use cases for Linked Open Data (LOD)

Concepts from authority files are basically used for reliable identification and description, for writers, books, and subjects. Then, the difficulty is to provide relevant links to editions of the books, manuscripts, archives, images. Some online examples can explain how data.bnf.fr copes with these issues. For instance, when searching for information about a writer like Goldoni,[9] the user will find all his main works with pages gathering all the editions of the books. There are links to the references of the editions of the catalogues, to the digital items, and to the manuscripts such as the letters which

[6] Archival Resource Key.

[7] http://gallica.bnf.fr/ark:/12148/bpt6k134521m.

[8] http://ark.bnf.fr/ark:/12148/bpt6k134521m.

[9] http://data.bnf.fr/11905320/carlo_goldoni.

Goldoni received. The authority files have been extracted, and the identifiers are used to make persistent links, and to avoid duplicates. The "Work" pages are created at the FRBR level. This means that the book is described as a concept and not as a particular edition of this Work. All the editions are gathered around this concept. For instance, on the page about the *Trionfi* by Petrarcha,[10] it is possible to find a list of the manuscripts and of the printed books, with links to all the digital items when they are available. Writers are of course linked with their works, with the associated documents, and with other writers. This kind of information is extracted from the documents linked to both of them. For a writer such as Leonardo Bruni, a page[11] gathers links to all editions, manuscripts, and digital items. The texts he wrote, translated, edited or commented are available separately, depending on his role on the document. The user can look for the translations he made, for instance. There are also links to the pages of the associated writers, such as Cicero and Aristotle. Since he was an editor of Cicero, we provide both the references of the edited books, and a link to the page of Cicero.[12] The web semantic tools and the reliable links enable us to build pages around the common properties and infer new relation from our RDF graph. For instance one can find:

- data.bnf.fr pages associated with the date 1515;[13]
- data.bnf.fr pages associated with the date 1789;[14]
- or all the authors who have been making coins, such as Louis XIV.[15]

[10] http://data.bnf.fr/11953648/petrarque_les_triomphes.
[11] http://data.bnf.fr/12027636/bruni_leonardo.
[12] http://data.bnf.fr/11885977/ciceron.
[13] http://data.bnf.fr/what-happened/date-1515.
[14] http://data.bnf.fr/what-happened/date-1789.
[15] http://data.bnf.fr/vocabulary/roles/r370.

Data.bnf.fr makes links and publishes web pages containing already about 2.5 million linked resources. The complete data is also displayed in RDF and available by clicking on the RDF icon, at the bottom of the pages, by adding the following suffixes to the URL: NT, N3, RDF-XML, according to the format needed,[16] via content negotiation, using a RDF web browser, from the URL, or by bulk downloads.[17] As yet (in the summer 2012), the complete available dataset is 6.3 million RDF triples, which is not to massive considering the 2.5 million resources, thanks to the proper links that avoid us too much redundancy. All the raw data is also displayed in RDF and available with an open license. Allowing all kind of uses, also for commercial purposes, was not obvious.

Why we use an open license for data.bnf.fr. Legal and technical requirements

Displaying information on the web means that the institution is responsible for publishing documents. Legally speaking, the library becomes responsible for the content which is displayed. From a "marketing" perspective, publishing information on the web is an incentive to focus on what you can do best, and to let others take care of the things they do better than you, because the users will prefer to use their resources anyway. For instance, library catalogues are describing resources, and handling "concepts" that have to do with documents. The strengths and weaknesses are not the same, for instance, as in an encyclopaedia. It would be no use to try to insert universal encyclopaedic knowledge inside library catalogues, just as it would be useless to provide full lists of documents in-

[16] Example http://data.bnf.fr/11928016/jules_verne/rdf.xml.

[17] From http://data.bnf.fr/semanticweb-en.

side general information encyclopaedias. On the web, libraries are bringing a long-time perspective. They have been collecting books for centuries and data for decades. The manpower provided has no equal in terms of describing books. Besides, the data has been structured quite early, with international standards since the 1960s. This "descriptive data" was not produced in a "marketing" perspective: all elements are accurate, and meant to be interoperable, even though several formats actually exist. The rules for producing the bibliographic descriptions remained stable, and were strictly followed by trained cataloguers. Through several formats for various types of documents such as books (MARC formats), archives (EAD), and digital resources (DC), the standards were respected. Therefore, the information can be trusted and processed automatically, even on a long period of time, and through huge amounts of data. The library catalogues are already machine-readable, even if it is not yet necessarily with web standards. Displaying them on the web with web standards implies to use identifiers (URIs) so that people can quote the resources. If we want to let people use these web resources, we must provide reliable links. Because this is a great opportunity to share our cultural materials, the Bibliothèque nationale de France decided to make the structured data from data.bnf.fr freely available, with an Open Licence. Opening library data on the web is a way to take part in the "open data" movement, and to give access to the information to the broader public, by using the most recent technologies. It is also an incentive for others to use this material and to give access to culture. By making the RDF data free, this project also take part of the international experimentations of "Linked Open Data" (LOD) that have popped up among national libraries. As Gildas Illien puts it, «Transforming pre-existing MARC records and authority vocabularies in RDF triples; starting to implement the FRBR model ; playing with the semantic web standards ; building applications and datasets of a new, linked data-friendly type: this is what looking at LOD means to them at this stage» ("Are you ready to dive in? A case for open data in national libraries"). Because libraries are working on a long-term perspective, data.bnf.fr also tries to experiment on solutions that can be used in the original library catalogues. First, when developing "matching" and algorithms for gathering data around "Works", we try to provide honest information for displaying data on the web, for instance, to avoid having duplicates, to avoid displaying keywords that would not match with the content of the documents, or any other information that would in fact not be useful for the end-user. This is why we keep so many links to all the resources in the original catalogues, inside the pages of data.bnf.fr. We also try to build routines and mechanisms that can be used inside the original catalogues, in the long-term, for instance for automatically generating "Work" pages inside our authority files, according to FRBR. Besides, after a first year of presence online, we can already have a feedback from some users, on the kind of content that is being used. Some of them are re-distributing the dataset and referencing it for others to re-use, starting with data.gouv.fr,[18], the official open data portal of the French State, but also other sites such as CKAN[19] OKF[20] and open data directory.[21] Other users are data specialists from the cultural sector, who use a part of the data for specific purposes in their local applications, such as the *Institut français*.[22]

Some are developers who want to build timelines for research pur-

[18] http://www.data.gouv.fr/donnees/view/Donn%C3%A9es-compl%C3%A8tes-du-contenu-de-la-BNF-30383137.

[19] http://thedatahub.org/dataset/data-bnf-fr.

[20] http://en.wikibooks.org/wiki/Open_Metadata_Handbook/Technical_Overview#Biblioth.C3.A8que_Nationale_de_France_.28BnF.29.

[21] http://open.mflask.com/dataset/data-bnf-fr-bibliotheque-nationale-de-france.

[22] http://ifverso.com.

poses, such as "Yokafun",[23] or for Smartphone applications.[24] This broad range of uses of the "raw data" shows us that library information can be useful for broader communities, even if the first purpose remains to describe collections and to give access to them. When gathering all available resources at the level of intellectual "works", the dataset is not a "catalogue" in the traditional sense, because it is not necessarily used for identifying a document or handling a collection. It becomes part of the web, in a new way. The authority files and identifiers are more important than ever to build this kind of service, but the dataset itself is something else than a traditional catalogue. Besides, the web tools allow us to keep trace of the behaviour of users. We can of course collect the direct feedback on how people have been reacting to it, what kind of content is being used and what has to be improved, which leaves a wide range of possible improvements for the future.

Works cited

Illien, Gildas. "Are you ready to dive in? A case for open data in national libraries". *Libraries now! Inspiring, surprising, empowering. IFLA World Library and Information Congress. 78th IFLA General Conference and Assembly*. 2012. http://conference.ifla.org/sites/default/files/files/papers/wlic2012/181-illien-en.pdf. (Cit. on p. 413).

[23] http://plindenbaum.blogspot.fr/2011/07/drawing-svg-timeline-with-httpdatabnffr.html;https://gist.github.com/1093853.

[24] For instance https://sites.google.com/site/catbnf;http://www.appforcash.com/section/item/id/41491.

ROMAIN WENZ, Bibliothèque nationale de France.
romain.wenz@bnf.fr

Wenz, R. ”Linked open data for new library services: the example of data.bnf.fr”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5509. DOI: 10.4403/jlis.it-5509. Web.

ABSTRACT: The Bibliothèque nationale de France (BnF) develops a new project, bringing together data from catalogues (MARC), archives (EAD) and digital resources (DC). It makes links and publishes web pages, available at http://data.bnf.fr, with already about 750.000 linked resources. All the raw data is also displayed in RDF and available with an open licence. The presentation will explain the importance of authority files and identifiers to build this kind of service, an give a first feedback on how users have been reacting to it: what kind of content is being used.

KEYWORDS: data.bnf.fr; Gallica; Bibliothèque nationale de France; Library linked data; Authority file

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

Cataloguing in the open: the disintegration and distribution of the record

Martin Malmsten

Background

LIBRIS, the Swedish Union Catalogue saw first light in 1972 with the mission to rationalize work in libraries through the use of information technology. To this end a cooperative cataloguing environment was created where each member library would describe it's collection. The core idea, much in the spirit of the times, being that every library contribute what was unique to them, but still gain access the sum of descriptions from all member libraries. Access to the database was restricted to member libraries with access to a 3270 terminal and a modem connection the the LIBRIS mainframe. In 1997 a service was created to search and display the data on the web, in effect making the information in LIBRIS accessible to anyone, anywhere provided they had access to the internet. At around the same time a Z39.50[1] target was made available making it possible for copy-cataloging clients to remotely search and download data from LIBRIS. This was a important also since it allowed machines to access information about the collections described in

[1] http://www.loc.gov/z3950/agency.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5512

LIBRIS. Though Z39.50 was deemed sufficient for copy cataloguing, assuming the same format is used, it does lack in other areas such as userfriendlyness and has proven a problem when trying to connect to others outside the library community.

In 2006 a new catalogue frontend, a ″next generation catalogue″, was launched and with it a number of Application Programming Interfaces (APIs) meant to make it easier for anyone to create services built on top of LIBRIS′ data. A consequence of this new catalogue was that a URI, or a URL rather, was minted for each bibliographic record making it possible for the system itself and external applications to link to individual records in a persistent and easy way. While this was all good and well given that you knew (a lot) about the APIs and could understand the formats provided (MARC21, DC, MODS, etc.), everything was still centered around descriptions contained within records. Also, few links to resources outside LIBRIS were present in the records once they had been found, even tough links to sites such as Google Books were provided to a human user of the frontend.

In 2008 LIBRIS as a whole was released as linked data (Malmsten), including authority data describing persons, organizations and subject headings. Links to external resources such as those described by Library of Congress Subject Headings (LCSH), Wikipedia and the Virtual International Authority File (VIAF) were added and in an instant LIBRIS was part of quickly expanding graph of metadata generated by a number of entities, mostly outside the GLAM sector. This move garnered a lot of interest especially from other government entities and other organizations that wanted to either link to or download parts of the authority data. This is unsurprising since an identifier for, e.g, a famous author is useful both for libraries, archives and other cultural heritage institutions.

Cataloguing in the open

Starting September 2011 the National Bibliography and Swedish Authority file, two subsets of the LIBRIS database, are made available in the same format that they are being created (MARC21). This decision to not only expose Resource Description Framework (RDF)/linked data derived from the records, but also the records in their original form is a strategic one. By doing so anyone can see, evaluate, reference and ultimately contribute to the work done by the National Library, the assumption being that visibility and openness will in the end lead to higher quality data.

To avoid any restrictions when it comes to re-use the National Library has chosen[2] CC0[3] for the National Bibliography and the Authority File, effectively putting the datasets in the public domain. The only exception is MARC field 667 (Nonpublic note) which is filtered out due to reasons of personal integrity. The license was chosen because we see a problem with attribution licenses such as ODB-BY and CC-BY when it comes to re-use of data over time, for example so called Cattribution stacking". The goal is to release to whole dataset in the original format with CC0, though since some records or part of records in LIBRIS originates from a number of other organizations (LC, BNB, DB, OCLC, etc), this will take some time. Anyone wishing to access the data can do so in two ways: either through Atom feeds and/or using the OAI-PMH protocol. The feeds are, essentially, getting the data from the cataloguing system in real-time. This means that anyone can get changes made to the Swedish Authority File within mere milliseconds of the change being made. The work of the cataloguers is being made available essentially as they type. As an aside the choice not to provide complete files for download (a

[2] http://librisbloggen.kb.se/2011/09/21/swedish-national-bibliography-and-authority-data-released-with-open-license.

[3] http://creativecommons.org/choose/zero.

″dump″) of the data is to signal that the dataset is live, whereas a dump is essentially stale and/or obsolete in the same instant you download it. However, both Atom and OAI-PMH can easily be used to download the whole dataset, so the distinction is perhaps somewhat academic.

Consuming linked data

To actually reap the benefits of linked data, however, we must also use it as an integral part of our systems, not only expose it. This has a number of interesting consequences. First, since linked data allows us to relate to any data, wherever it is created, the distinction between internal and external datasets disappear. This has a profound impact on systems design since you have to rely on protocols normally used for external datasets internally. Secondly, the matter of control then becomes a matter of trust rather than technology. If you cannot control the information you must decide who to trust, and perhaps even cooperate with them though that is surely a small price to pay for a world of data. Often those we trust will also be consumers of the data we produce. Thirdly, as more and more information in our records relate to some resource outside of our control, be it a person in Wikipedia or a subject heading in id.loc.gov, the idea of a record becomes somewhat less interesting. A lot of what was the record is then actually controlled by descriptions that live elsewhere, in datasets merely linked to rather than owned.

Seamlessness

Often we do not need to aggregate data produced by others, but rather react to the fact that something that we link to has changed. Again this makes complete downloads of datasets at discrete intervals problematic since

1. the purpose a change may not be apparent and
2. having multiple batch imports that relate to same data will definitely cause problems.

A situation where synchronisation of datasets is done through download of the whole dataset simply does not scale. A goal for a truly linked system must be to be able to signal changes, or information about changes, seamlessly in near real-time to interested parties. While creating feeds makes it possible for clients to ask for updates, more often than not the answer will be that no, no updates have been made. This makes for a very inefficient system where a lot of requests have to be made to a ensure that the datasets are in sync. There are at least two efforts that deal with this issue: pubsubhubbub[4] and ResourceSync[5]. By using hubs to which a publisher can signal a change and a consumer can subscribe, an efficient network is created through which information can be sent.

Conclusion

Actually using linked data, as opposed to only exposing it, somewhat removes the distinction between internal and external datasets. Control becomes at matter of trust, not technology. The record disintegrates as the data becomes distributed.

[4] https://code.google.com/p/pubsubhubbub.

[5] http://www.niso.org/workrooms/resourcesync.

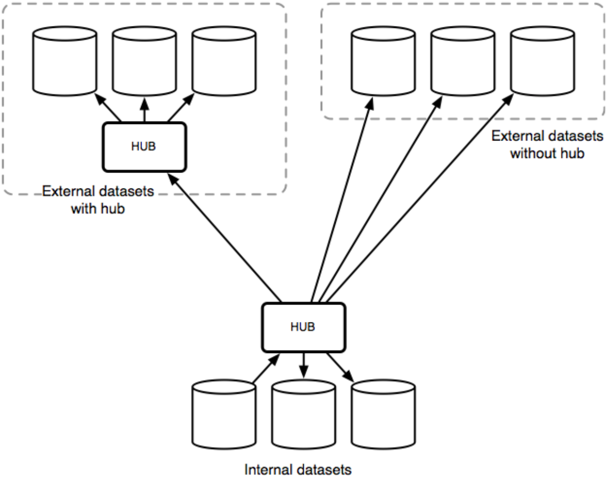


Figura 1: A change in one dataset is propagated through hubs, both to internal and external datasets

Riferimenti bibliografici

Malmsten, Martin. «Making a Library Catalogue Part of the Semantic Web». *International Conference on Dublin Core and Metadata Applications, DC-2008–Berlin Proceedings*. 2008. (Cit. a p. 418).

MARTIN MALMSTEN, National Library of Sweden.
martin.malmsten@kb.se

Malmsten, M. "Cataloguing in the open: the disintegration and distribution of the record". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5512. DOI: 10.4403/jlis.it-5512. Web.

ABSTRACT: As part of a strategic investment in openness the Swedish National Library has released the National Bibliography and accompanying authority file as open data with a Creative Commons Zero license effectively putting it in the public domain. The data has been available as linked open data since 2008 but is now also released in its original, complete form making it fit for re-use by other library systems. An important principle of linked data is to link out the other datasets. However, as data becomes more interconnected and distributed the need for ways to track and respond to changes in other datasets, even ones outside our area of control, becomes bigger. The issue of who to trust of course becomes vitally important. This paper details the motivation behind the release as well as the technology used to support it. Also, a consequence of exposing and using linked data is that the idea of the record as a self contained and delimited entity starts to fall apart.

KEYWORDS: Library linked data

Submission: 2012-04-25
Accettazione: 2012-08-31
Pubblicazione: 2013-01-15

Metadata framework and application profiles in the global structure of catalogs and digitization projects of the Vatican Library

Paola Manoni

The issue of technical interoperability unfolds in the informatics field of the Vatican Library in an area covering the aggregation of bibliographic data (derived from systems with different metadata structures) and the interconnection of structured data (provided through sharing of cooperative programs) within the recent digitization project regarding collections of manuscripts and incunabulum. This presentation aims to show the context, as well as the application models and systems currently in use in the Vatican Library, focusing attention on the metadata framework involved in the overall structure of the new Library's catalogues inaugurated into the web on the 16th of May 2012. Moreover, there will be illustrated the organization of the Digital Library to be soon published on the Web. As a preliminary discourse it would be appropriate to mention something about the current state of informatics and cataloging development of the Vaticana Library, in order to locate more precisely the library environment into a broader design scenario of the global interoperability of the library sector. Starting from the

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5516

present and going backwards, the BAV (Biblioteca Apostolica Vaticana)[1] has warned the need to interact with different bibliographic databases, stipulating an aggregation method for heterogeneous data and syntax regarding the description of various collections composing Vaticana's patrimony. It was therefore designed a system to establish connections between various catalogs, in order to enable an integrated (and not federated) search within the platform of a general catalog. Without aiming to retrace the history of Vaticana's electronic catalogs but only to give some introducing information, here it is appropriate to mention that at the BAV the consultation databases are available, particularly, in relation to:

- publications (monographs and text periodicals): the first electronic catalog built in the middle of 80's years of the past century on the first group of current cataloguing, continued with the complete retrospective data entry and then subjected to revision (especially with regard to the access points);

- graphic materials (prints, drawings, photographs): the project "Stampe on line" ("Prints online"), started in 1998, includes the analytical cataloging of prints as well as drawings. Since 2001, it has been proceeded with the digital scanning of images and their hypertext links together with the bibliographical cards;

- coins and medals: the project in progress that has been started in 2001. This project includes numismatic descriptions until the sixteenth century. As for graphic materials, almost for all data digital images are available: for each unit description the photo shots of the obverse and reverse of a coin are got;

- incunabula: newly established electronic catalog (available since 2009) that combines full descriptions of the inventory

[1] http://www.vaticanlibrary.va.

published in 1997 and maintained by P. William Sheehan. The project BAVIC (Bibliothecae Apostolicae Vaticanae Incunabulorum Catalogus) consisting of analytical cataloging of specimens is in progress.

The mentioned above catalogs are structured under the MARC21 record syntax. Regarding the organization of electronic catalogs of manuscripts and archival documentary collections (the latter ones are in the Archives Section of the Department of Manuscripts), we refer to another type of computer processing and cataloging regarding both data format, system architecture and their interoperability.

- Manuscripts: since 2002 a project of retrospective conversion of catalogs and inventories of paper fonds has been operated. The project was divided into several phases. Preliminarily, the definition of criteria for data processing has been chosen according to the standard TEI (Text Encoding Initiative Consortium). The electronic cataloguing is therefore conforms to the elements set out in the form of TEI-MS - specifically for the description of manuscripts. From a management perspective, a database able to handle its structure in XML was designed. The system, developed by the Coordination of computer services of the Vaticana is called InForMA (Informatics For Manuscripts and Archives), which manages the production and research data on manuscripts and on archive's papers.

- Archival material: formalized in the same XML language, but according to the XML EAD (Encoding Archival Description) standard. In effect, the system can handle different collections of data or documents that refer to different metadata schemas, at the same time allowing them to be managed separately (different formats for different cases) or in the related way: in cases of shared lists of authority items (for names, titles and se-

> mantic descriptors) and the connections between bibliographic records and files of digital images.

Starting from 2011 there has been initiated a study for the adoption of a new management system OPAC able not only to aggregate in a general catalog the contents of individual catalogs but also to manage structures of separated and referenced indexes to each type of catalog. OPAC enables to link bibliographic records to other information resources available in the Library and to manage information useful for researchers, alongside the mere querying of the catalog. The tool which was built allows the regular population of data from different catalogs to a repository in which the different data formats or MARC21, TEI-MS and EAD are stored, each for the semantics of reference. The system conforms to the interoperability protocols such as OAI-PMH and OpenURL. Now it would be appropriate to mention a few words on the use of these two protocols at the BAV, with particular regard to the first experience of the aggregation of data for the general catalog. As it is known, the main goal of the protocol is to allow two or more archives of separated data to exchange metadata. OAI-PMH defines the communication protocol through rules and methods for transferring metadata, established according to the two classic basic entities: data providers and service providers. The data provider provides the OAI-PMH to publish its metadata. The service provider sends through the OAI-PMH requests (via http) to data providers and collects metadata. The data provider responds with an XML message. Considering the problems of data harvesting (given the extensive testing on which it stands) rather than the use of MARCXML for cataloging data in MARC 21 (which obviously did not present any problems), it is worth dwelling on the handling of metadata for archives and manuscripts involved in managing Vaticana's catalog. The regular population of the general catalog for descriptions of manuscripts and archives encoded in EAD and TEI formats occurs according to a conversion format designed to maintain similar structural organization and granularity, thus giving way to interaction via OAI-PMH, considering the difficulty in managing crosswalk producing satisfactory results, which avoid the reduction or loss of information in data representation. One can think about the hierarchical organization of the descriptive units in EAD, about the richness of attributes and complexity of nested elements in both EAD and TEI-MS encodings. The literature on the topic abounds with the assessment on the Dublin Core/DC (also in qualified version) wherewith the protocol represents the aggregated data. Anyway, the DC format is too general to mediate the types of referred metadata. In the construction of the catalog, there was made a computers choice of the hybrid type. Particularly, besides the harvesting protocol functioning for MARC 21catalogs there was implemented a Web service according to the peer-to-peer interaction model transferring the EAD and TEI-MS data from the InForMA system to the general catalog, ensuring the interoperability between heterogeneous systems. Beyond the harvesting technique, the system (based on the technological solutions offered by Infor) controlling the new catalogs offers an indexing system that allows the uniform representation of the results of a search. In practical terms, it means that the interoperability framework for communication among the catalogs has the function of the device for the collection of heterogeneous data converging within the same system in which, in real time, the bibliographic descriptions in different formats and from the specific catalogs are gathered. In other words, the interaction occurs within a database, whose conceptual model is based on the notions of information objects, classes, attributes and methods. The classes and attributes are used to describe the structural aspects (the metadata for each application domain), while methods are used to represent the functions of objects (information units treated) derived from different cataloging representations. If within the problem of the interaction it would be desirable to situate other types of aggregation of information, in the evaluation of the front-end structure of new Vaticana's system one can see how – for each information unit dedicated to the various catalogs – the OPAC querying is provided, through an interface between various proposed widgets offering textual, visual, multimedia documentaries relevant to different document areas. The application presents to users a functional organizational complex that can be related to the model of a portal, consisting of web pages dedicated to each cataloging activity, to OPAC queried through searchable indexes and separated for each type of catalog and material typology. The search presents also the possibility to perform post-coordinated selections as the association of similar categories and concepts, through an automatic network of links offering to the researcher documents available of the same author or publisher in another language, another edition or another support, as well as techniques for scrolling through lists of results via progressive refinement categories, properties and attributes Moreover, there are also searches on fuzzy[2] logic allowing the analysis of the search term in its different parts (roots, prefixes, suffixes, spelling variants, etc.) thus suggesting to the researcher performing a search other allied results. The search, staring from a bibliographic record, is further extended to the collection of electronic databases available at the BAV or other resources available via the web through a link resolver that conforms to the standard OpenURL. This last is a protocol for metadata exchange, aimed at managing services of the so-called "linking in context", widely used in bibliographic databases, in academic citation systems and in the open archive. It's also possible to manage the interaction mecha-

[2] Polyvalent Logic (literally 'fuzzy'), extension of Boolean logic or the categorization of a continuous variable in the subranks.

nisms with RSS feed (for now not activated), as well as the function of word cloud. Both of them will be eventually allowed to meet the specific user information needs permitting the access to the system with authentication (after the activation profile is determined). The option for consultation of the application by entering user's credentials will also allow the user to customize the information display and OPAC functionalities (e.g. elimination of widgets, sorting or reduction of the search indexes, etc.). From a point of view of data aggregation and interoperability between systems at the BAV, there can be further extended the treatment with the interaction between catalogs and the RFID implementation created at the Vaticana as well as with the security system[3] tracking the movement of books and persons within the possible paths in the Library. Anyhow, in order to not abandon the central theme of the present discussion, the work is propelled to analyze the issue of interoperability in view of digitization projects of the BAV. Regarding this topic, the Library is putting into practice the digitization both in a view of long-term storage, and in the implementation of a digital library accessible via the Web and through links to the catalogs. Leaving aside the first aspect, whose organization or specific use of metadata is evidently not directly involved with the subject discussed in this section, it would be appropriate to consider the structure of the second goal. Hereby there should be immediately added that, while the first project (concerning the long-term preservation) has already started, the second one has not yet produced any public evidence, even the presentation of the first group of data and digital objects in the Web over the next years is expected. Without any fear of contradiction, there can be said that treating a library of digital objects means, necessarily, handling of its metadata. The choices, which have been already accomplished in this direction at the BAV, are related to the

[3] System for internal use, which protects the privacy.

management of digital objects according to the METS (Metadata Encoding and Transmission Standard) model which represents an information container to encode in the XML metadata required for managing digital library objects. In addition, to each digital object accessible through METS profile a persistent URL is assigned. For the projects originated within cooperative agreements, stipulated with other international institutions considering the importance of the creation of digital libraries shared with the BAV, the digital objects are expressed in RDF graphs. These last represent an approach for structural interoperability, aiming at managing a connections between digital objects belonging to collections in various libraries. From this scenario there can be obtained a description of digital composite objects formed by a set of distributed resources in the web independent of their allocation. For this task each resource aggregation will be described in a resource map expressing the semantic relationships between existing aggregations of resources. It is expected that each map has an associated identifier (URI) that will be invoked (differentiated) with an http request, thus providing the serialized representation according to the RDF/XML standard format. In the first instance, these technical specifications make reference to the important work of digitization of the Vatican Library in partnership with the Bodleian Libraries of Oxford. This activity is made possible thanks to the contribution of the Polonsky Foundation committed to support initiatives providing access to and knowledge of cultural heritage of mankind preserved in the world's great library collections. Within this specific work plan, in a five-year period, it is expected the digitization of a million and a half pages in total regarding volumes (Greek and Hebrew manuscripts, and incunabula) chosen by both institutions. Thus, among the first treasures connected via the web differentiated URI there will be (among the incunabula) the famous *De Europa* di Pio

II Piccolomini, and the Latin Bible of 42 lines of Johann Gutenberg. While for the Hebrew manuscripts of the Vaticana, as for one of the most important collections of the existing Jewish codes (even if not the most extensive) there will be chosen the *Sifra*, written between the late ninth and mid tenth century and, probably, representing the oldest Jewish code come up to our days. Moreover, there will be chosen a whole Bible written in Italy around 1100 as well as biblical commentaries, Halakhah and Kabbalah, Talmudic commentaries, and writings from the liturgy, philosophy, medicine and astronomy fields. The earliest Greek manuscripts that will ultimately enter into the inferential logic of Linked Data will be important witnesses of the works of Homer, Sophocles, Plato, Hippocrates, as well as codes of the New Testament and the Fathers of the Church, many of which are richly decorated with Byzantine miniatures.

PAOLA MANONI, Biblioteca Apostolica Vaticana.
manoni@vatlib.it

Manoni, P. ”Metadata framework and application profiles in the global structure of catalogs and digitization projects of the Vatican Library”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5516. DOI: 10.4403/jlis.it-5516. Web.

ABSTRACT: This paper will focus on the application profiles recently implemented in the new Vatican Library’s discovery tool that interacts with interoperability standards and manages different metadata. The presentation will also take into consideration what the Library is planning for accessing web-based digitized manuscripts collections..

KEYWORDS: Vatican Library-Manuscripts collections; Library linked data

Submitted: 2012-06-01
Accepted: 2012-08-31
Published: 2013-01-15

Open data in the Italian Government: the experience of the City of Florence

Giovanni Menduni, Gianluca Vannuccini,
Giacomo Innocenti

The City of Florence, Italy, has been carrying out a large data integration and consolidation process since 2003, that was even more improved and targeted to an open and federated approach during 2009. Results of such a process are now being exploited in several current hot eGovernment fields, among which business intelligence, and, more recently, the open data movement. During the year 2011, the City of Florence lead an internal structured assessment process in which each department named an open data referee, and was called to analyze which available public data stores were eligible to be opened up in a suitable website section. It is worth noticing the strategic choice made by City of Florence, that decided to publish, when possible, information at its atomic level, in order to provide raw data, which can be directly analyzed by the final users. This approach has made the open data particularly suited for business intelligence purposes, but on the other hand the hidden informative richness makes such documents harder to read without proper inquiring tools. To this end, a child project has born. Its aim is to provide the common citizen with simple visual and graphical tools, designed to investigate the data via comparisons defined by

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2012).
DOI: 10.4403/jlis.it-5500

the user through parametric selectors. At the same time, the best practices to publish on the internet the open data have been studied, taking into account the different needs which can characterized the possible users. Based on the results of this research, it has been decided to export the open data made of collections of single electronic documents according to open and free formats, in order to not compromise the user experience because of the need of proprietary viewers and reading tools. Among all the possible choices, particular attention has been put on the more popular formats; specifically, for data directly deriving from database tables the CSV formats has been preferred to the others, while for geospatial data the choice has fallen to the KMZ format. City of Florence has also decided to encounter the needs of the local industry by starting an other project for the publication of its data as services. According to this paradigm the information is provided to the user according to parameters that the latter can specify at the moment of the request. This approach is particularly fruitful for software applications, because it reduces both the computational burden and the weight of the data transmission. To this aim the City of Florence has started a series of dialogs with independent developers, which have led to the production of several software application for mobile phones based of the open data. At first, a number of ”apps” for Android and iOS smart phones has been realized by embedding the information directly into the software; then, new apps accessing the open data via specific online services has been designed and they are actually in testing phase. In the above framework, the City of Florence has also decided to used the data-as-service paradigm to federate its collections with the ones provided by other subjects. In this respect the Linked Data approaches been pursued, with particular attention to the W3C star-rating scheme suggested by Tim Berners-Lee, From a technical point of view, the linked data are made of rich entities interconnected via semantic properties and the common method for their representation is a list of clauses describing both the features of each entity and their linking properties. To this aim it is used the RDF language and its derivations along with a specific storage system. Even though specific software and tools exist to handle these elements, a preliminary analysis drove to the adoption and experimentation of an alternative solution consisting of a wrapper that maps the data in a DBMS into an RDF graph at run-time. This choice has been preferred in order to not duplicate the open data with another storage system. Hence, the D2R server open source project has been choice as a basis for DBMS to RDF mapping. The first three datasets that were mapped in RDF are the municipal road network naming system, the corresponding house number system, and the rank of streets with more street-accidents. The first open data section of the city of Florence website was published in October 2011, to which followed the publication of RDF scheme, and a data-as-service connector, i.e. a SPARQL endpoint, for the three above datasets. In February 2012, a brand-new specific open data portal was published, whose main sections were: the dataset catalog; the ″open data for all″ section, with human-readable data visualizations helping the user to experience simple Business Intelligence tools; the linked data section; and, eventually, the GeoPortal section, giving access to geospatial data through standard catalog, storage and web portal, which are based on the open source GeoNetwork and GeoServer projects. Inside the open data project, the City of Florence has promoted the citizen participation by means of communication tools as the social networks. In particular, a specific hashtag `#opendatafirenze` was used for citizens engagement via Twitter, and for the City of Florence as a channel to publish the daily news. Since the new website launch, in February 2012, more than 120 tweets with this hashtag have been registered, and the ″daily dataset" is twitted every working day.[1] The main efforts of the City of Florence on this field are now particularly focused on the improvement of the RDF-mapped portion of the whole data store (the museums dataset was recently published), and on the improvement of the adopted dictionaries. Indeed, due to the lack of specific and easy-to-use semantic standards for public administration, we adopted a home-made dictionary to model the above RDF-mapped datasets. The early draft of the dictionary was intrinsically linked to the nature of the open data; however, this is not the best practice to define a semantic ontology. Hence, a further development has led to define a new dictionary, this time considering all the opposing needs: on the one hand, to be completely fruited, its logic has to be close to the data itself; on the other hand, it has to be sufficiently generic in order to be re-used in several different contexts. To this aim the new release of the dictionary propose has been internally organized in a deep structure, where the upper layers provide the generic property, while the lower ones are devoted to describe the details of the considered data. In this way, a user can adapt the dictionary by over-writing only the lower layers, when needed. To simplify this approach, the dictionary itself has been divided into a series of connected sub-dictionaries. Even though the related ontology is not providing yet a complete coverage of the Public Government needs, hopefully further collaborations with other public bodies (such as central national bodies, for Government ontology standardization) will better off the dictionaries.

[1] https://twitter.com/i/#!/search/?q=%23opendataFirenze&src=hash.

GIOVANNI MENDUNI, Area di Coordinamento Programmazione, Sostenibilità e Innovazione - Comune di Firenze.
giovanni.menduni@polimi.it

GIANLUCA VANNUCCINI, Servizio Sviluppo Infrastrutture Tecnologiche - Comune di Firenze.
g.vannuccini@comune.fi.it

GIACOMO INNOCENTI, Dipartimento di sistemi e informatica - Università degli Studi di Firenze.
giacomo.innocenti@unifi.it

Menduni, G., G. Vannuccini, G. Innocenti. ”Open data in the Italian Government:the experience of the City of Florence”. *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2012): Art: #5500, p. 435–440. DOI: 10.4403/jlis.it-5500. Web.

ABSTRACT: The City of Florence, Italy, has been carrying out a large data integration and consolidation process since 2003, that was even more improved and targeted to an open and federated approach during 2009. Results of such a process are now being exploited in several current hot eGovernment fields, among which business intelligence, and, more recently, the open data movement. During the year 2011, the City of Florence lead an internal structured assessment process in which each department named an open data referee, and was called to analyse which available public datastores were eligible to be opened up in a suitable website section. The first Open Data section of the City of Florence website was published in October 2011, to which followed the publication of RDF scheme, and SPARQL endpoint for the three above datasets. In February 2012, a brand-new specific open data portal was published, whose main sections were the dataset catalogue, the ”Open data for all” section, with human-readable data visualisations, the Linked Data section, and the GeoPortal section, giving access to geospatial data through a standard catalogue and a portal which are based on the open source GeoNetwork and GeoServer projects.

KEYWORDS: City of Florence; Open data; eGovernment

Submitted: 2012-04-25
Accepted:
Published: 2012-05-16

JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

Legal Interoperability: Making Open (Government) Data Compatible with Businesses and Communities

Federico Morando

Introduction

This paper offers a view of ″legal interoperability″ amongst (open) data licenses that I understand as the possibility of (legally) mixing data coming from different sources (e.g. government data, user generated content, corporate data) and using them within a broad range of business (and community) models. I will discuss license interoperability from a point of view at the intersection between law and economics. Notice, however, that this is just one of the possible perspectives on this issue. For instance, some authors (e.g. Fujita and Tsukada or Krötzsch and Speiser) attempted a formalization of the analysis of license interoperability from the disciplinary angle of software engineering, mathematical logic and formal languages. The rest of this paper is organized as follows. In section 2, I describe the reasons why licenses are needed in order to open up (government) data. Section 3 sketches a description of the open data licensing landscape. Section 4 represents the core of mine contribution and

JLIS.it. Vol.4, n.1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5461

includes a table summarizing the license interoperability scenario. Finally, section 5 concludes.

The Legal Background: Open Data Need "Copyright" Licenses

It is acknowledged (Krötzsch and Speiser) that the distribution of data also requires their licensing. In other words, the terms under which data can be reused and republished should be explicit (Bizer, Heath, and Berners-Lee; Miller, Styles, and Heath). This is the case because of the current "copyright default", i.e. the set of rights that the current regime of copyright protection automatically grants to authors. (Notice that, in this paper, I use the term "copyright" in a broad sense, encompassing copyright strictu sensu, droit d'auteur and the database sui generis right.[1]) This "copyright default" implies that "all rights are reserved" for the maximum duration allowed by the law (typically, the life of the author plus 70 years). Moreover, no formalities are required to enjoy these rights, not even a statement that a certain work is protected.[2] And, even if in principle the protection granted to non-creative databases through the sui generis database right alone is shorter than copyright protection (i.e. 15 years), also this exclusive right is automatically granted and it is very difficult to entirely rule out the possibility that a layer of copyright protection also applies to any given dataset. In a few words, in the absence of a clear statement about the legal status of a dataset,

[1] For the sake of brevity, I remand to Aliprandi for an introduction to the legal protection of databases in Europe, with an approach focused on open data related issues (Aliprandi).

[2] La diffusione delle note del tipo "Tutti di diritti riservati" è solo un fossile dell'art. 3 della Convenzione di Buenos Aires del 1910, che richiedeva un'esplicita affermazione della riserva all'autore dei suoi diritti.

it is safer to assume that data are legally locked-up preventing any kind of reuse (or copy). In conclusion, to open data, we ″also need to clearly communicate our basic intention: that the data is available for reuse. And we need to be clear on what forms of reuse we expect or want to support.″(Dodds)

A Bird′s-eye View on Open Data Licenses

When you actually try to choose a license for your data, you have to weight various elements, including the opportunity of adopting the most standard tools and the legal suitability of the selected tool for the licensing of data in general and for the kind of data you are opening up in particular. A first option could be to use Free and Open Source Software licenses. However, this is not a very widespread approach, since FLOSS licenses are very specialized tools and using them for things which are not pieces of software is typically suboptimal. Another option could consist in using one of the licenses from the Creative Commons (CC)suite, which are general purpose licensing tools. These licenses offer to right-holders a menu of elements/modules from which they can pick their favorite combination and including: ″Attribution″ (BY); ″Non-Commercial″ (NC); ″No Derivative Works″ (ND), meaning that only verbatim copies could be produced; and ″Share Alike″ (SA), meaning that the author requires the creators of derivative works to adopt the same license used by him/her (the so-called ″viral″ or ″copyleft″ effect).[3] The (meaningful) combinations of the previous elements generate six different licenses, two of which can be defined as ″open

[3] You may find more practical information about the CC licenses at http://www.creativecommons.it. For a more theoretical and impartial commentary about CC licenses, see (the first part of) (Elkin-Koren).

licenses":[4] CC BY and CC BY-SA. On top of these standard licenses, CC also offers a right waiver or dedication to the public domain (with a fall-back clause to a very permissive license in jurisdictions where some rights cannot be waived): Creative Commons Zero (CC0). Also notice that CC licenses are "ported" (i.e. translated and adapted) to each national legislation. To date, the latest versions of the CC licenses (3.0) for EU countries include special provisions about the sui generis database right (since this right is peculiar of these jurisdictions), consisting in a waiver of the right.[5] Until the release of their (EU) 3.0 version, it was unclear if the CC licenses where an appropriate legal tool for the licensing of databases (potentially) protected by the sui generis database right. This was one of the reasons because of which, in 2006, Talis[6] published the first public license specifically targeting open data, the Talis Community License (Miller, Styles, and Heath)[7]and then funded the lawyers J. Hatcher and C. Waelde to draft the Public Domain Dedication and License (PDDL). This activity then triggered the creation of the Open Data Commons (ODC) project, which is currently part of the Open Knowledge Foundation project portfolio.[8] To date, the ODC licensing suite includes the PDDL, the Open Database License (ODbL) - which is a copyleft license - and an Attribution license. All these licenses concern the rights covering a database as such (as opposed to the data it contains). Despite the availability of standard public licenses, such as the ones from CC and ODC, several national governments decided to draft their own licenses for the

[4] According to the Open Knowledge Definition:http://opendefinition.org.

[5] Rectius (and mainly for license-geeks), the licensor waives the right of using the sui generis database right as a tool to legally enforce the license clauses.

[6] Talis is a firm developing Semantic Web solutions and, in particular, consulting and training services in this domain (http://www.talis.com/corporate).

[7] http://www.talis.com/tdn/tcl.

[8] http://opendatacommons.org/about.

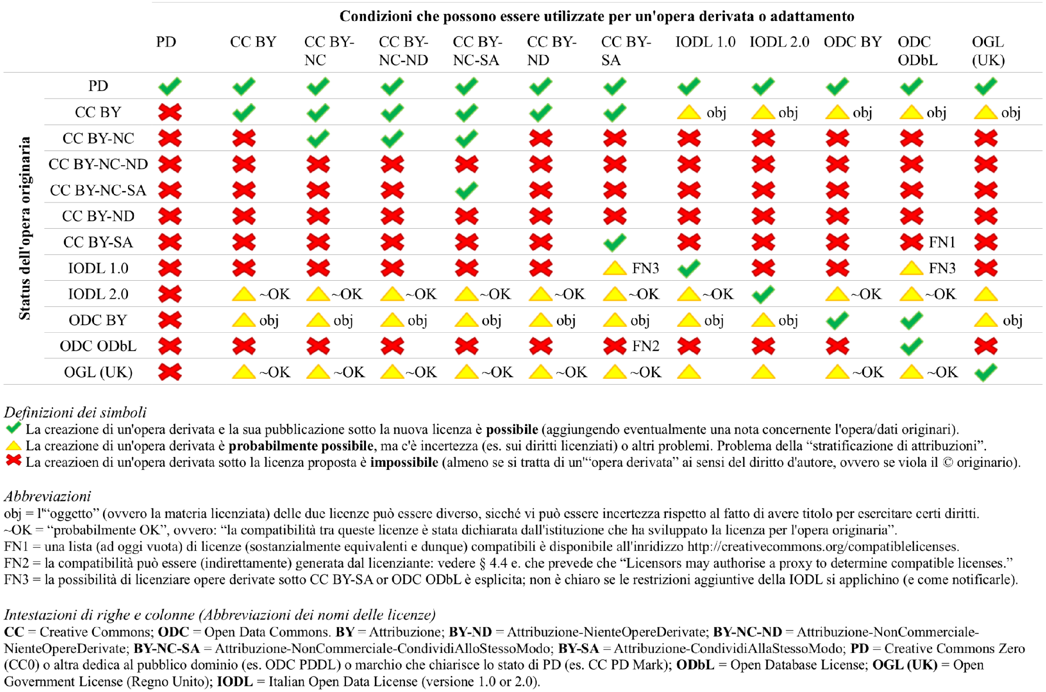
release of (open) Public Sector Information. One of the first countries to do so (also because of the choices of CC concerning the sui generis database right) was the United Kingdom, with its ”Click Use” license and its current non-transactional evolution, the Open Government License (OGL). The OGL is essentially equivalent to a CC or ODC Attribution license, but it includes some specific provisions concerning ”Crown copyright” and other clauses addressing standard public sector worries, such as forbidding the use of the released information in such a way that suggests any official status.[9] The OGL approach adopted in the UK was almost immediately and is still followed all over the world(Judge) (and in Europe in particular). For instance, France adopted its own License Ouverte, while Italy produced the Italian Open Data License (IODL), which was released in various versions, starting from a non-commercial beta version to arrive (going through a 1.0 copyleft version) to the current 2.0 version, which is a simple attribution license.

Legal Interoperability Is an (Open) Issue

From the previous section, it should be clear that the ”market” offers several different (open) licensing solutions, but are they somehow compatible from the point of view of a reuser of open (government) data?
Even if we remain within the CC licensing system, there are compatibility problems (as observed, amongst others, by Guibault (Guibault)). In fact, building on the table that Creative Commons drew to describe compatibility *w*ithin the CC license suite,[10] the License Inter-

[9] http://www.nationalarchives.gov.uk/doc/open-government-licence.

[10] http://wiki.creativecommons.org/Frequently_Asked_Questions#Can_I_combine_two_different_Creative_Commons_licensed_works.3F_Can_I_combine_a_Creative_Commons_licensed_work_with_another_non-CC_licensed_work.3F.



		Condizioni che possono essere utilizzate per un'opera derivata o adattamento											
		PD	CC BY	CC BY-NC	CC BY-NC-ND	CC BY-NC-SA	CC BY-ND	CC BY-SA	IODL 1.0	IODL 2.0	ODC BY	ODC ODbL	OGL (UK)
Status dell'opera originaria	PD	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	CC BY	✖	✓	✓	✓	✓	✓	✓	△ obj	△ obj	△ obj	△ obj	△ obj
	CC BY-NC	✖	✖	✓	✓	✓	✖	✖	✖	✖	✖	✖	✖
	CC BY-NC-ND	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖
	CC BY-NC-SA	✖	✖	✖	✖	✓	✖	✖	✖	✖	✖	✖	✖
	CC BY-ND	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖	✖
	CC BY-SA	✖	✖	✖	✖	✖	✖	✓	✖	✖	✖	✖ FN1	✖
	IODL 1.0	✖	✖	✖	✖	✖	✖	△ FN3	✓	✖	✖	△ FN3	✖
	IODL 2.0	✖	△ ~OK	△ ~OK	△ ~OK	△ ~OK	△ ~OK	△ ~OK	△ ~OK	✓	△ ~OK	△ ~OK	△
	ODC BY	✖	△ obj	△ obj	△ obj	△ obj	△ obj	△ obj	△ obj	△ obj	✓	✓	△ obj
	ODC ODbL	✖	✖	✖	✖	✖	✖	✖ FN2	✖	✖	✖	✓	✖
	OGL (UK)	✖	△ ~OK	△ ~OK	△ ~OK	△ ~OK	△ ~OK	△ ~OK	△	△	△ ~OK	△ ~OK	✓

Definizioni dei simboli

✓ La creazione di un'opera derivata e la sua pubblicazione sotto la nuova licenza è **possibile** (aggiungendo eventualmente una nota concernente l'opera/dati originari).

△ La creazione di un'opera derivata è **probabilmente possibile,** ma c'è incertezza (es. sui diritti licenziati) o altri problemi. Problema della "stratificazione di attribuzioni".

✖ La creazioen di un'opera derivata sotto la licenza proposta è **impossibile** (almeno se si tratta di un'"opera derivata" ai sensi del diritto d'autore, ovvero se viola il © originario).

Abbreviazioni

obj = l'"oggetto" (ovvero la materia licenziata) delle due licenze può essere diverso, sicché vi può essere incertezza rispetto al fatto di avere titolo per esercitare certi diritti.
~OK = "probabilmente OK", ovvero: "la compatibilità tra queste licenze è stata dichiarata dall'istituzione che ha sviluppato la licenza per l'opera originaria".
FN1 = una lista (ad oggi vuota) di licenze (sostanzialmente equivalenti e dunque) compatibili è disponibile all'inridizzo http://creativecommons.org/compatiblelicenses.
FN2 = la compatibilità può essere (indirettamente) generata dal licenziante: vedere § 4.4 e. che prevede che "Licensors may authorise a proxy to determine compatible licenses."
FN3 = la possibilità di licenziare opere derivate sotto CC BY-SA or ODC ODbL è esplicita; non è chiaro se le restrizioni aggiuntive della IODL si applichino (e come notificarle).

Intestazioni di righe e colonne (Abbreviazioni dei nomi delle licenze)

CC = Creative Commons; **ODC** = Open Data Commons. **BY** = Attribuzione; **BY-ND** = Attribuzione-NienteOpereDerivate; **BY-NC-ND** = Attribuzione-NonCommerciale-NienteOpereDerivate; **BY-NC-SA** = Attribuzione-NonCommerciale-CondividiAlloStessoModo; **BY-SA** = Attribuzione-CondividiAllaStessoModo; **PD** = Creative Commons Zero (CC0) o altra dedica al pubblico dominio (es. ODC PDDL) o marchio che chiarisce lo stato di PD (es. CC PD Mark); **ODbL** = Open Database License; **OGL (UK)** = Open Government License (Regno Unito); **IODL** = Italian Open Data License (versione 1.0 or 2.0).

Figure 1: License Interoperability Table

operability Table in Fig.1 attempts at offering a broader synoptic view on license interoperability complexity. Let us start from a given dataset (the ″original″ one), whose license is listed on the first column. For each cell on the same line, I try to answer the question ″can I use the license mentioned in the header of this column to release a new dataset incorporating a significant part of the original one?″. As you can imagine, a green symbol means ″yes″, a red one means ″no″, while a yellow one means ″maybe, but with some limitations or uncertainty″. Fujita and Tsukada (Fujita and TsukadaYasuyuki) describe a similar problem in the following way: ″Alice produces a content c and attaches a license l1 to it, and then posts it to a website s1. Bob likes c and wants to advertise its wonderfulness widely by posting it to a website s2. As for contents posted to s2, it is necessary to attach a license l2. The problem is whether Bob can post c to s2. For instance, if we assume that l1 only permits non-commercial use and l2 permits commercial use, Bob cannot post c to s2. This is because the commercial use contradicts the rule of l1.″ Using the definitions of this example, in the License Interoperability Table the first column lists some possible l1 licenses, while the first line lists the same licenses playing the role of l2. As in the example, if l1 includes the NC element, l2 must also include it: indeed, as you can check on the Table, the only green symbols are in correspondence of other NC licenses. (Incidentalmente, si noti un corollario di questo esempio, già evidenziato da Seneviratne, Kagal, and Berners-Lee; Ricolfi et al.) As a further illustration, if the original dataset is in the public domain (e.g. available under CC0 or the PDDL), then it is possible to achieve perfect interoperability (as observed by Hatcher Hatcher). Indeed, PD dedications/waivers are ″universal donors″, because they ensure one-way compatibility with any other licensing tool, as you see from the first line of green symbols. By ″one-way″ I mean that the derivative dataset of a public domain dataset could be licensed in any foreseeable way, but the opposite does not apply. Actually, as you can see in the first column of symbols, the only way to (legally) license a derivative dataset with CC0 or the PDDL is to start with a public domain dataset. Unfortunately, as soon as we depart from the ″original dataset in the public domain″ scenario, we are in trouble. Even simple liberal attribution licenses are clearly interoperable just with the other and more (or equally) restrictive licenses from the same licensing suite, but some legal uncertainty arises as soon as we consider the option of releasing, for instance, under ODC BY a derivative dataset of a CC BY original dataset. In fact, in this case, it may be unclear if a CC license (especially if we are dealing with a 2.5 version or earlier) grants the necessary permissions concerning the sui generis database right. It would be impossible to enter in further details in this short paper, but let me mention that some yellow symbols are also due to the fact that national open data licenses include clauses which are not related with copyright (e.g. they require the licensee to respect of the national Data Protection Act). Since these, possibly redundant, clauses are not included in standard public licenses, compatibility is uncertain. To conclude, notice that the License Interoperability Table is admittedly oversimplified. You may have different interpretations about virtually any cell and this is the best proof of a serious interoperability problem in the open data domain! In fact, it does not matter if some legal scholars could argue that mixing two datasets may be possible under certain conditions. The crucial issue is that reusers need to clearly understand what they can (or cannot) do, without asking their lawyers and, ideally, without reading too many licenses. Frankly speaking, I doubt that they can.[11]

[11] Moreover, all these problems just become more complex when license interoperability issues concerning the copyright domain are summed to issues concerning the terms of service of online services exposing data, as discussed, for instance, by (Palfrey and Gasser).

Conclusions

Both license stewards (i.e. the organizations drafting open licenses) and license users (i.e. the data holders/publishers) may play an important role in achieving license interoperability in the open data domain. License stewards should, first of all, beware of what I call "license vanity" and "push the egos of the lawyers off of center stage"(Lessig). In particular, they should recognize that the vast majority of the adopters of share-alike/copyleft licenses want their creations to be interoperable and, to achieve that, they would happily tolerate some minor legal/technical flaws in the licenses they adopt. Moreover, license steward - and national governments in particular - could facilitate license interoperability if they addressed non-copyright worries with other tools, such as privacy notices, disclaimers and any kind of soft-law or non-binding norm. For instance, they should never create a new attribution license which differs from the standard public licenses just by a clause about the respect of data protection law, since - by definition - data protection law is already binding. An informal reminder would perfectly serve the goal of the public sector body, while adding a license clause to the same end triggers the drafting of a new license and more legal uncertainty. Indeed, standard copyright licenses, together with the appropriate notices and disclaimers, could form a "licensing framework" clarifying all relevant issues, without breaking license interoperability. Actually, the Government of New Zealand already did that through its NZGOAL framework,[12] based on the use of CC BY. From the point of view of the end-user, to date, the only interoperability-proof solution is the dedication to the public domain (e.g. CC0 or the PDDL), but this approach neglects the existing demand for attribution/provenance requirements (which is especially widespread

[12] http://ict.govt.nz/guidance-and-resources/information-and-data/nzgoal.

amongst public sector bodies and frequently for good reasons, e.g. related with accountability) or share-alike clauses (which enable the typical self-defensive but inclusive approach adopted by online communities). Technically speaking (from a legal point of view) a combination of Open Data Commons licenses applied to databases and Creative Commons licenses applied (when appropriate) to their content could represent an ideal solution, but this approach is far too complex, so that - to my knowledge - just a few projects are actually adopting it. And, in any case, if different users are adopting different solutions, license interoperability will be endangered. As Linksvayer(Linksvayer) puts it, ”a single universal recipient license (i.e., a single widely used copyleft license, or the equivalent) for all non-software works, including databases, is crucial.” Hopefully, Creative Commons licenses in their 4.0 version will finally license all relevant rights (including the database sui generis right) in a simple and consistent way,[13]making CC BY-SA capable of playing this role of ”universal recipient license”. Waiting for the emergence of such a license, data holders are warned: there are no universal recipients and universal donors are the key to achieve interoperability, hence the dedication to the public domain is the only way to maximize the potential of their data.

References

Aliprandi, Simone. “Open Licensing e Banche Dati”. *Informatica e diritto*. (2011): 25–43. (Cit. on p. 442).

Bizer, Christian, Tom Heath, and Tom Berners-Lee. “Linked Data-The Story So Far”. *International Journal on Semantic Web and Information Systems*. (2009): 1–22. <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>. (Cit. on p. 442).

[13] As one may expect from the current draft open for public discussion. See http://wiki.creativecommons.org/4.0

Dodds, Leigh. “Rights statements on the Web of Data”. *Nodalities Magazine*. (2010): 13–14. (Cit. on p. 443).

Elkin-Koren, Niva. *Exploring Creative Commons: A Skeptical View of a Worthy Pursuit*. Ed. P. Bernt Hugenholtz and Lucie Guibault. Kluwer Law International, 2006. (Cit. on p. 443).

Fujita, Kunihiko and TsukadaYasuyuki. “An Analysis of Interoperability between Licenses”. ACM, 2010. (Cit. on p. 447).

Guibault, Lucie. “Creative Commons: Struggling to 'Keep it Simple'”. Ed. Jens Gaster, Erich Schweighofer, and Peter Sint. 2008. (Cit. on p. 445).

Hatcher, Jordan S. “Open data and the law”. *Nodalities Magazine*. (2010): 5–6. (Cit. on p. 447).

Judge, Elizabeth F. *Enabling Access and Reuse of Public Sector Information in Canada: Crown Commons Licenses, Copyright, and Public Sector Information*. Ed. Michael Geist. 2010. 598–642; 620–625. (Cit. on p. 445).

Krötzsch, Markus and Sebastian Speiser. “ShareAlike Your Data: Self-referential Usage Policies for the Semantic Web”. ISWC, 2011. (Cit. on p. 442).

Lessig, Lawrence. *CC in Review: Lawrence Lessig on Compatibility*. 30 november, 2005. (Cit. on p. 449).

Linksvayer, Mike. *Position Paper for the Share-PSI.eu Workshop: Removing the Roadblocks to a pan-European Market for Public Sector Information Re-use*. 15 april 2011. (Cit. on p. 450).

Miller, Paul, Rob Styles, and Tom Heath. “Open Data Commons, a License for Open Data”. LDOW, 2008. (Cit. on pp. 442, 444).

Palfrey, John and Urs Gasser. “Mashups Interoperability and eInnovation”. *Berkman Center Research Publication*. (2007). <http://dx.doi.org/10.2139/ssrn.1033232>. (Cit. on p. 448).

Ricolfi, Marco, et al. “The ”Licensing” of Public Sector Information”. *Informatica e diritto*. (2011): 129–146. (Cit. on p. 447).

Seneviratne, Oshani, Lalana Kagal, and Tim Berners-Lee. “Policy-Aware Content Reuse on the Web”. *The Semantic Web - ISWC 2009*. 2009. 553–568. (Cit. on p. 447).

FEDERICO MORANDO, Nexa Center for Internet & Society, Politecnico di Torino–DAUIN.
federico.morando@polito.it

Morando, F. ”Legal Interoperability: Making Open (Government) Data Compatible with Businesses and Communities”. *JLIS.it*. Vol.4, n.1 (Gennaio/January 2013): Art: #5461. DOI: 10.4403/jlis.it-5461. Web.

ABSTRACT: ”Legal interoperability” could be defined as the possibility of legally mixing data coming from different sources (including governmental data, data generated by online communities and data held by private parties). Legal interoperability is similar to technical interoperability, since it is a prerequisite for mixing data and create new knowledge or services. But it also has its own peculiarities, for instance because it could be achieved simply choosing the appropriate licensing scheme, but also because self-help mechanisms which could – at a certain price – guarantee technical interoperability to third parties cannot (lawfully) solve legal interoperability issues. In the mid/long run, legal interoperability could be achieved thorough the evolution of legal frameworks in order to harmonize the landscape of Government Data. In the short term, the shortcomings generated by diversified legal frameworks may be alleviated through the careful choice of copyright licenses. The presentation will focus on the latter aspects, discussing existing public licenses (such as the Creative Commons and Open Data Commons ones), representing a de facto standard in this domain, and the main open data licenses developed by European governments (e.g. the Open Government Licenses in the UK, the French License Ouverte or the Italian Open Data License).

KEYWORDS: Data mining; Governement data; Interoperability; Library linked data

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15

The Seminar was sponsored by

FLORENTINA STUDIORUM UNIVERSITAS

ICCU

casalinilibri

COMUNE DI FIRENZE

ittig

BNCF

Fondazione Rinascimento digitale

CRUI

AIB
ASSOCIAZIONE ITALIANA BIBLIOTECHE

ITALE

AIE
Associazione Italiana Editori

ExLibris
The bridge to knowledge

JLIS.it

LINKeD HeRITAGe

@cult

SPACE
Cultura e innovazione

JLIS.it

Dipartimento SAGAS, Storia, Archeologia, Geografia, Arte e Spettacolo

con il supporto di:

supported by:

Casalini libri

Le Lettere

La piattaforma ICT, lo sviluppo e la manutenzione dell'installazione di OJS che ospita JLIS.it sono forniti da:

ICT platform, developing and maintenance for the OJS installation hosting JLIS.it are provided by:

CINECA

Direttore Responsabile ai termini di legge: Nicola Cavalli

In attesa di iscrizione nel registro stampa del Tribunale di Milano.

Finito di stampare nel mese di febbraio 2013 da

Ledizioni
The Innovative LEDIpublishing Company

http://www.ledizioni.it