JLIS.it

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-8587

(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: #8587. DOI: 10.4403/jlis.it-8587. 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, meta-languages, 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