JLIS.it

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

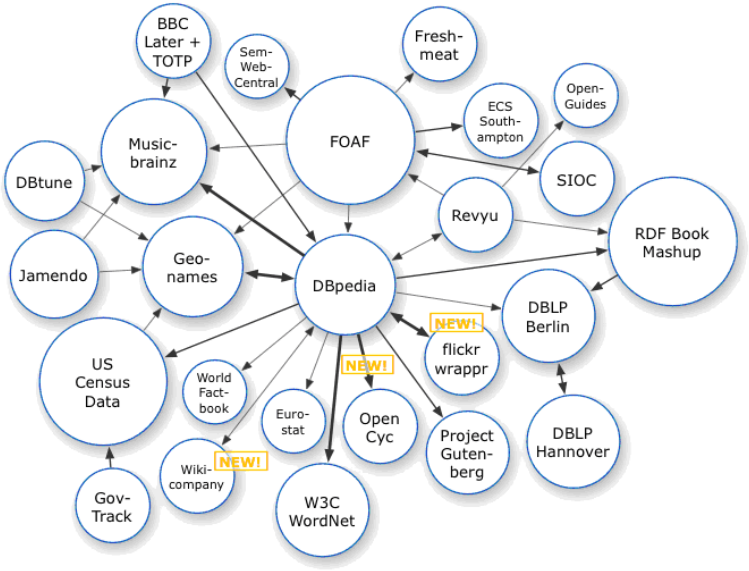

Figure 1: Diagram of the linked open data cloud (LOD) in 2007.

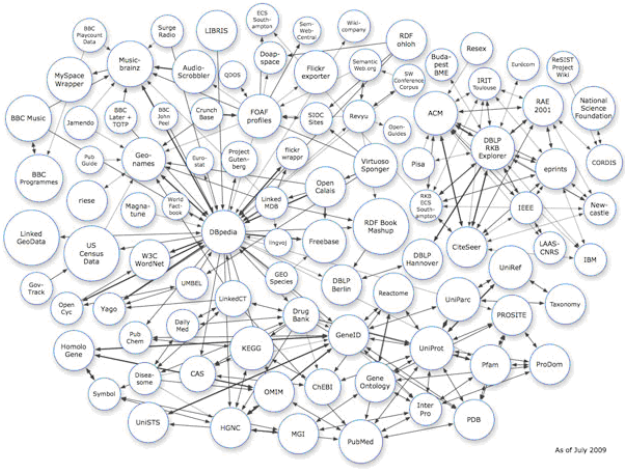

Figure 2: Diagram of the linked open data cloud (LOD) in 2009.

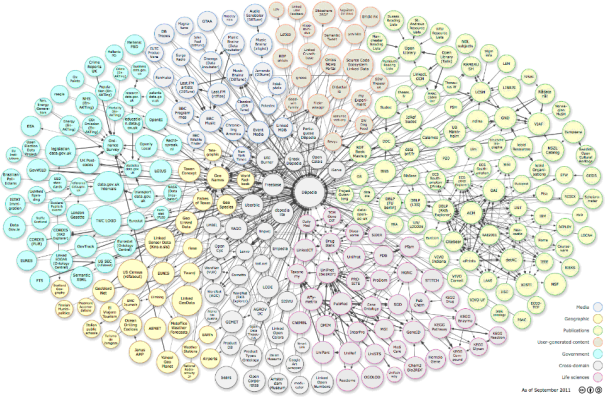

Figure 3: Diagram of the linked open data cloud (LOD) in 2011.

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.

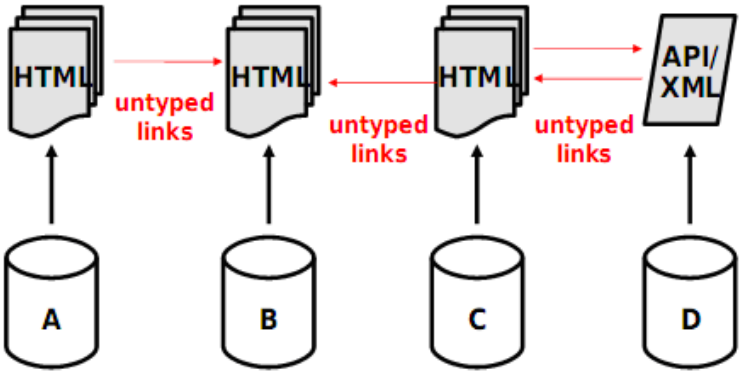

Figure 4: Representation of the web of documents, 17th International World Wide Web Conference W3C Track @ WWW2008, Beijing, China 23-24 April 2008 - Linked data: principles and state of the art.

- 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.

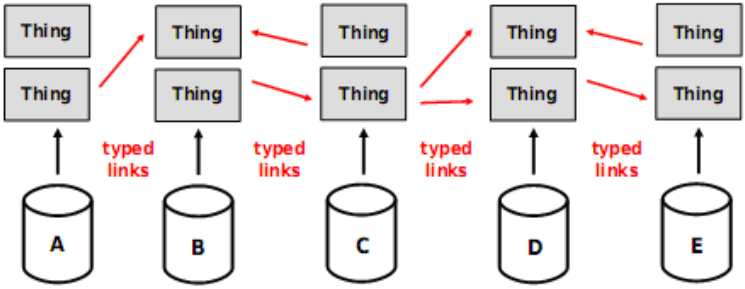

Figure 5: Representation of the web of data 17th International World Wide Web Conference W3C Track @ WWW2008, Beijing, China 23-24 April 2008 - Linked data: principles and state of the art.

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.

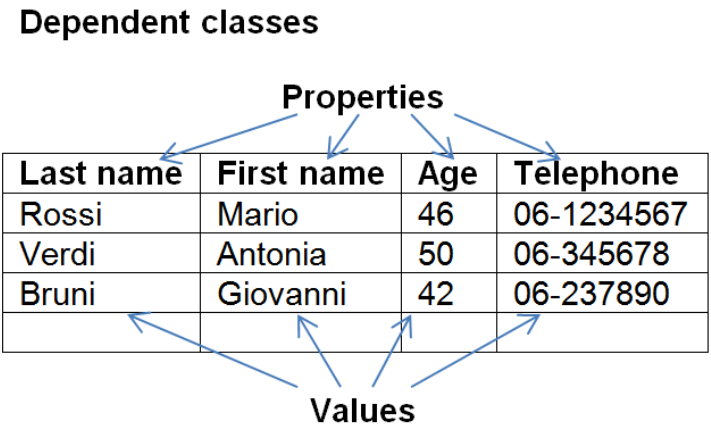

Figure 6: Representation of a relational database.

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:

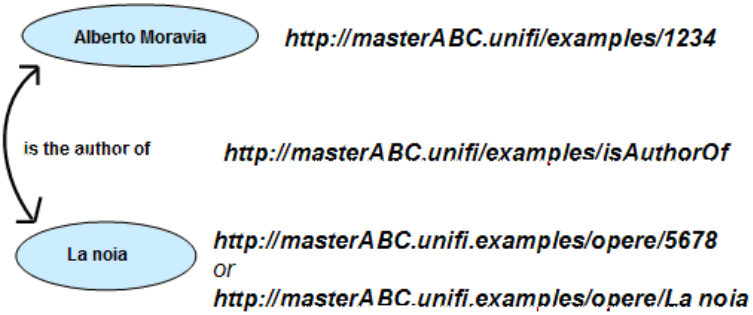

Figure 7: Representation of a triple (nodes and arcs) in RDF.

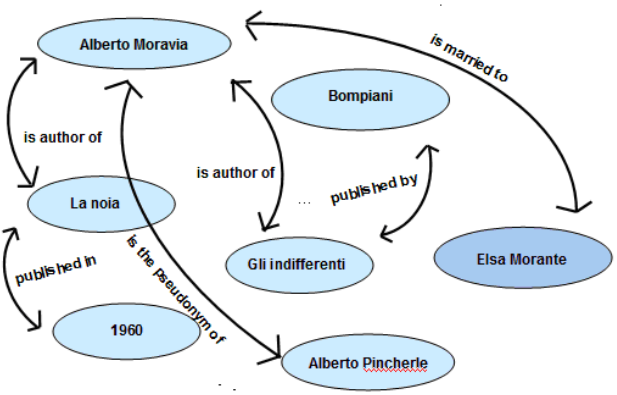

Figure 8: Representation of a network of assertions or triples.

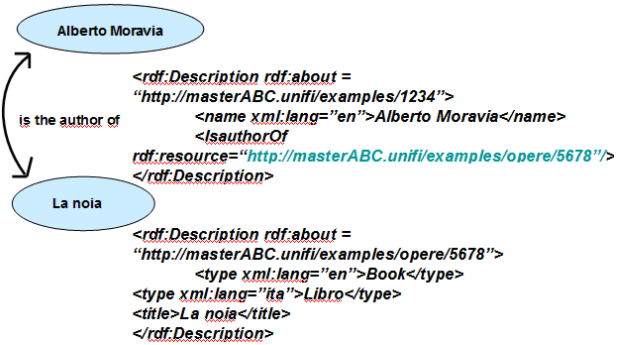

Figure 9: Representation of a triple in RDF/XML.

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 princi-

ples 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