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.

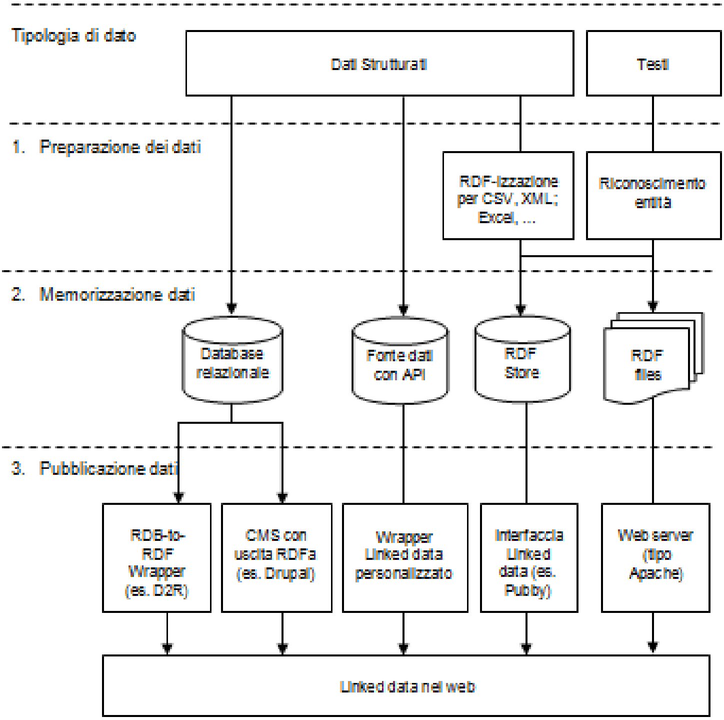


Figure 1: Workflow of the publication of heterogeneous data in linked data.

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 preexistent 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

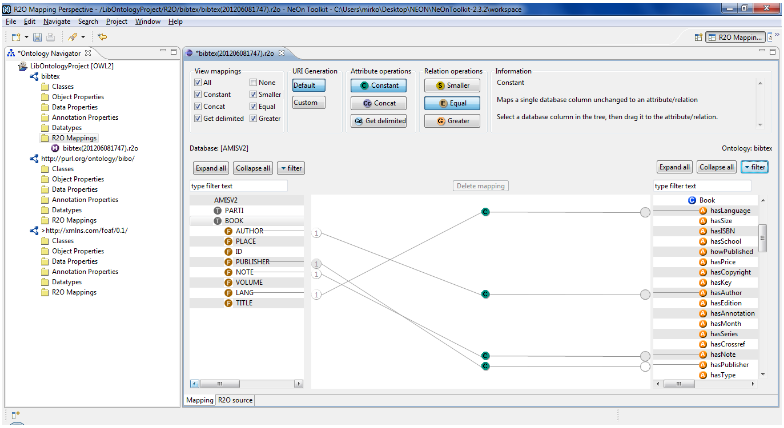
The relational database is Oracle, which contains bibliographic data and is structured, in short, in two different views, created in order to map two different ontologies: BOOK (mapped on the bibtex ontology) and PARTS (mapped on the bibo antology). Other tools used for the study include:

- 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.



Constant

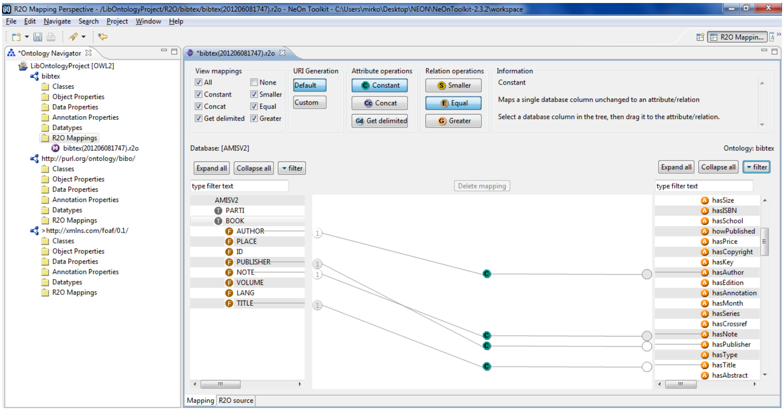
Maps a single database column unchanged to an attribute/relation

Select a database column in the tree, then drag it to the attribute/relation.

Figure 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.

Database: [AMISV2]

Ontology: bibtex

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>
```

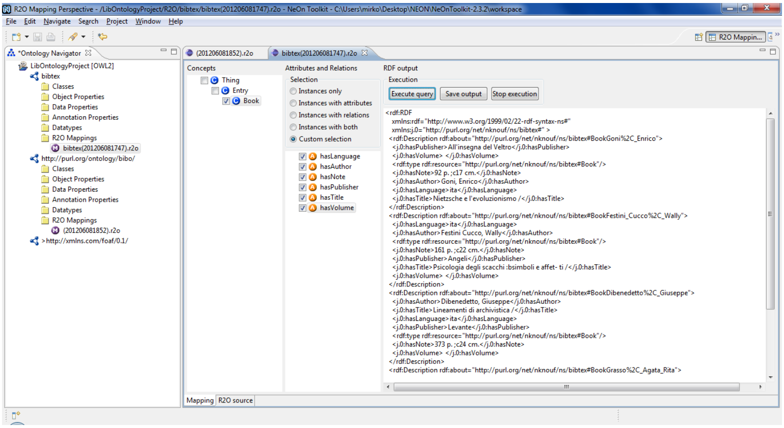
```
         <described-by>
           <attributemap-def name="http://purl.org/net/nknouf/ns/
               bibtex#hasLanguage">
             <selector>
               <aftertransform>
                 <operation oper-id="constant">
                   <arg-restriction on-param="const-val">
                     <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/
```



```
<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

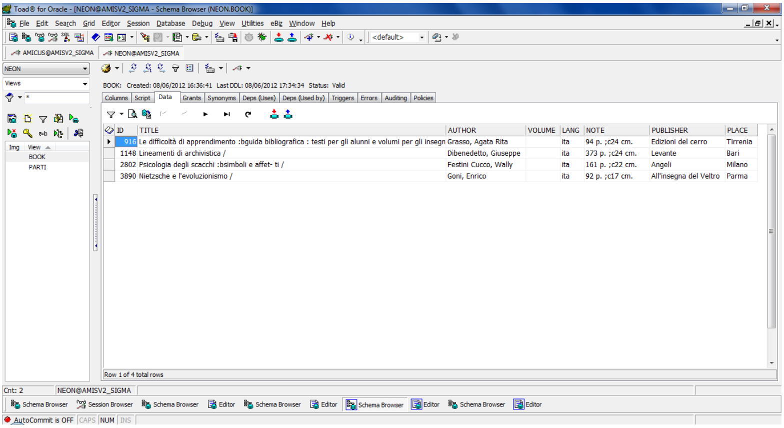
```
        bibtex\#Book"/>
    <j.0:hasLanguage>ita</j.0:hasLanguage>
    <j.0:hasAuthor>Goni, Enrico</j.0:hasAuthor>
    <j.0:hasNote>92 p. ; c17hcm.</j.0:hasNote>
    <j.0:hasTitle>Nietzsche e l'evoluzionismo /</j.0:hasTitle>
  </rdf:Description>
  <rdf:Description rdf:about="http://purl.org/net/nknouf/ns/
      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
<default>
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
NEON@AMISV2_SIGMA
Schema Browser Session Browser Schema Browser Editor Schema Browser Editor Schema Browser Editor Schema Browser Editor
AutoCommit is OFF CAPS NUM INS

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

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).

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013)

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