JLIS.it

Development of a metadata schema describing Institutional Repository content objects enhanced by "LODE-BD" strategies

Iryna Solodovnik

The move toward Linked Data will be the most significant change in library data in these two centuries
(J. Zaino, *The future of libraries*)[1]

1 Introduction

Issues like handling metadata, cross-referencing them consistently with authority control and semantic vocabularies, licensing activities valorizing scope and usage of digital resources within Institutional Repository (IR) infrastructures will become certainly increasingly challenging in the future years. It is due to emerging models and actualities like: Repositories of research data and Data Management in research infrastructures; interoperability among Repositories, with CRIS (Current Research Information Systems), and external services and applications; capturing research context in connection to re-

[1] http://semanticweb.com/the-future-of-libraries-linked-data-and-schema-org-extensions_b35315 .

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8792

search output by Service Providers; application trends of Semantic Web technologies service-oriented frameworks for bibliographic data; metadata management across disciplines with wide re-use of Repository data and services, as well as necessity of reliable value-added services over Trusted digital Repositories.

With this in mind, due attention must be paid for the development of qualitative and updated – according to current standards, guidelines and best practices – metadata application profiles supported by standard and "good practices" compilation and encoding strategies.

To provide more visible and sharable data on the web, different communities are aligning their digital contents according to current best practices for publishing and consuming data on the web, formalized within Linked Data (LD, Web of Data, Web 3.0) paradigm, the first practical expression of the Semantic Web – declared useful, feasible and applicable to all forms of data. Digital contents published as LD sets are presented graphically within Linking Open Data (LOD) Cloud,[2] namely a visual historic landscape with the evidence of many different L(O)D packages covering actually more than an estimated 50 billion facts[3] from different knowledge domains. These facts are of varying quality and most of them (published under Open Licenses) can also be re-used (consumed and enriched) by different agents.

In the last years, also different bibliographic datasets - including digital collections, metadata, semantic and authority files (mono e multi-lingual vocabularies, classifications, thesauri) – have been published and re-used according to "Tim's 5 star deployment scheme"[4]

[2]CKAN: Linking Open Data Cloud, http://datahub.io/group/lodcloud.

[3]Linked Heritage Project "Best practice report on cultural heritage", http://www.linkedheritage.eu/getFile.php?id=229.

[4]Tim Berners-Lee, Up to Design Issues, 2006, http://www.w3.org/DesignIssues/LinkedData.html; "Tim's 5 star" Open Data plan with examples, http://5stardata.info; OCLC video: "Linked Data for Libraries": short introduction to the concepts

principles. Library Linked Data (LLD) Report and CKAN Registry section for LLD,[5] Linked Open Vocabularies (LOV) Service,[6] the "Global interoperability and Linked Data in libraries" international "know-how" exchanging meeting (*Global interoperability and Linked Data in libraries. Special issue of "JLIS.it"*) can be cited within the first most important "witnesses" reporting and describing proliferating of bibliographical LD activities at the global scale. The landscape of bibliographical information – treated according to LD methodologies – is already enough widespread. Just to mention some connected experiences:

1. German National Library (DNB) LD Service for authority bibliographical data linking;[7]

2. Library of Congress LD authority files;[8]

3. LD collections from "The Open Library", "The European Library", "Europeana" and "WorldCat.org"[9] web services;

4. Hungarian National Library OPAC and Digital Library published according to LD and SKOS (Simple Knowledge Organi-

and technology behind Linked Data, how it works, and some benefits it brings to libraries, http://www.youtube.com/watch?v=fWfEYcnk8Z8.

[5]Library Linked Data Incubator Group: Datasets Value Vocabularies, and Metadata Element Sets W3C Incubator Group Report 25 October 2011, http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025; CKAN, Library Linked Data: http://datahub.io/group/lld.

[6]http://labs.mondeca.com/dataset/lov/index.html.

[7]http://openbiblio.net/2012/01/26/german-national-library-goes-lod-publishes-national-bibliography; http://files.d-nb.de/pdf/linked_data.pdf.

[8]http://authorities.loc.gov.

[9]http://openlibrary.org/; http://www.theeuropeanlibrary.org/tel4; http://pro.europeana.eu/linked-open-data;http://dataliberate.com/2012/06/oclc-worldcat-linked-data-release-significant-in-many-ways;https://www.oclc.org/data.en.html.

zation Systems)[10] formalisms;

5. British National Library bibliographical LD datasets connected to different LOD sets such as VIAF, LCSH, Lexvo, GeoNames, MARC country, "Dewey.info", RDF Book Mashup;[11]

6. LODUM, LOD service improving access to scientific and educational data at the University of Münster;[12]

7. "Burckhardtsource.org" and VOA3R digital infrastructures allowing enrichment, cross-relating and searching of cultural and scientific digital contents with LD technology support;[13]

8. "Data.bnf.fr" LD Project of the Bibliothèque nationale de France;[14]

9. LD at the Biblioteca Nacional de España;[15]

10. Public Library of Veroia in Web 3.0.[16]

An overview of consuming LD applications (faceted browsers,[17] LD browsing, LD search engine, On-the-fly *mashups* etc.) was recently good described in "Consuming Linked Data" document (Sequeda). To translate the initial success of Linked (Open) Data into a stable world-scale reality within bibliographical universe, encompassing

[10] http://iskouk.blogspot.com/2010/05/hungarian-national-library-opac-and.html.

[11] http://talis-linkeddata-libraries.s3.amazonaws.com/Linked%20Data%20Prototyping.pdf.

[12] http://code.google.com/p/lodum.

[13] http://burckhardtsource.org;http://voa3r.cc.uah.es.

[14] http://data.bnf.fr/docs/databnf-presentation-en.pdf.

[15] http://openbiblio.net/2012/02/02/linked-data-at-the-biblioteca-nacional-de-espana.

[16] http://gr.okfn.org/2012/10/libver/?lang=en.

[17] FAST (Faceted Application of Subject Terminology): http://fast.oclc.org, an Experimental OCLS Services for Controlled Vocabularies: http://tspilot.oclc.org/resources.

the Web 2.0 and commercial data alike, there are still several challenges to be addressed:

- "LD literacy" about benefits of publishing, re-using and integration of bibliographical resources as LD still needs to be widely promoted, directly (through standards) and indirectly (through "good practices");
- different requirements "to express metadata design patterns, both as templates for Linked-Data-compatible data formats and as reference points for creating and consuming coherent metadata within communities of discourse and practice"[18] according to a common *Resource Description Framework* (RDF, an international data exchange standard) should be re-evaluated;
- available strategies, e.g. "LODE-BD Recommendations" (Subirats and Zeng) regarding LD-enabling metadata encoding should be widely welcomed and implemented (De Robbio and Giacomazzi);
- processes for automatic alignment of metadata terms with LD-enabling sets should be better explored, formalized and shared as common models among different communities of practice;
- trust and common sense of LD are all still necessary: only trustworthy data patterns should be published as LD;[19]
- available *scientific data publication models* on top of LD (Bechhofer et al.) should be broadly transferred between research communities and exploited more deeply.

[18] DC-2013 "Linking to the Future" initiative, http://dublincore.org.

[19] It is the goal of LOD2 Project (FP7 Information and Communication Technologies Work Programme) to develop adaptive tools for searching, browsing, and testing authoring of LD, http://lod2.eu/Welcome.html.

Institutional Repositories (IRs) - as digital information systems promoting knowledge visibility on institutional digital research resources[20] - can be both publishers of their value datasets (e.g. metadata, vocabularies, collections) as well as consumers of available L(O)D sets.
For example, at Oregon State University ScholarsArchive@OSU both *Linked Dataset* covering University's *theses* and *dissertations* as well as *links* from this Dataset to external LD sets have been developed (Johnson and Boock). This activity has been started from converting MARC and Qualified Dublin Core metadata - describing the respective theses and dissertations - into LD through a RDF data model formalizing the expression of key data points for these resources. Afterwards, different relationships among IR's resources (with *handle* identifiers) have been described in a simple way (e.g. *rdfs:seeAlso*), as well as through complex semantics: mappings supported by internally and externally maintained LD datasets and controlled vocabularies for "Title", "Responsible Body", "Subject" entities. The querying of the entire *Linked Dataset* is possible via a *SPARQL* (*Protocol and RDF Query Language*) endpoint provided by the *Triple Store* that sits on top of the created knowledge LD base. Considering the importance of the above presented issues, this article is aiming at:

1. making a short overview of LD origins and its benefits for digital contents;

2. describing a role of controlled and semantic vocabularies in improving creation, access and retrieval of digital contents. A list of some important authority and semantic LD-enabling datasets will be provided;

[20] In the IR context the term "resource" can denote an article, monograph, thesis, conference paper, research report, presentation material, thesis, learning object, research data etc.

3. overviewing some approaches, documents and principles for creating metadata elements describing IR objects, focusing on the "Guidelines for metadata creation and management in Institutional Repositories" strategies (Conference of Rectors of Italian Universities (CRUI), Open Access Group, Italy);

4. presenting benefits of "LODE-BD" Recommendations (Subirats and Zeng), whose encoding *Decision Tree* strategies are devoted to support Repository metadata to become LD-enabled. Aside "literal" values for qualifying metadata properties, "LODE-BD" strategies are paying particular attention to assigning "non-literal" Uniform Resource Identifier (URI)[21] values. LD-enabling is also possible through mappings between Dublin Core (DC) metadata and more specific ontology-oriented metadata;

5. contributing with an extension to "Intellectual Property Rights" LODE-BD's *Decision Tree*, providing decision steps to licence choice. A list of some important licences - LD-enabling (identified by URIs) will be presented;

6. discussing briefly "Design-time" and "Run-time" LODE-BD implementation strategies and reporting thereupon some practice examples.

2 Linked (Open) Data: a brief reminder of its origins and benefits

In recent years, the concept *Linked Data* - referring to a set of best practices for publishing and connecting structured data on the Web - has

[21] The URI standard definition, RFC 2396: http://tools.ietf.org/html/rfc2396.

been already evolved as a high promising candidate into addressing one of the biggest challenges in the area of intelligent information management: the use of the Web as a platform for data and information integration in addition to document search. The term *Linked Data* (LD) was coined by Tim Berners-Lee in 2006 and formalized within already mentioned "Tim's 5 star deployment scheme", whose principles are being summarized as follows:

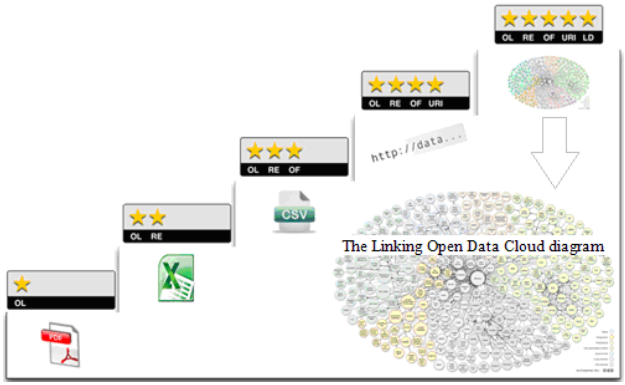


Figure 1: Tim's 5 star deployment scheme.
☆ Make datasets (contents) whatever format available on the Web under an *Open License*
☆☆ Make them available as structured data in RDF
☆☆☆Use *non-proprietary formats* (e.g. CSV instead of Excel)
☆☆☆☆ Use URIs to denote things, so that other agents can point at your datasets
☆☆☆☆☆*Link/combine* the data safely with other data in URIs global scheme to provide context

> [LD] isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of

> data. With Linked Data, when you have some of it, you can find other, related, data. (Berners-Lee)
>
> With Web advances to an era of *Open* and *Linked Data*, the traditional approach of sharing data within silos seems to have reached its end. From governments and international organizations to local cities and institutions, there is a widespread effort of *Opening* up and *Interlinking* their data. (Subirats and Zeng)
>
> Linked Data *does not of course in general have to be Open* - there is a lot of important use of Linked Data internally, and for personal and group-wide data. You can have *5-star Linked Data without it being Open*. However, if it claims to be Linked Open Data then it *does have to be Open*. (Berners-Lee)

Linked Open Data paradigm is a Linked Data strategy for global identity (Glaser and Halpin) of Open Data[22] (datasets published under *Open Licenses*) allowing re-use of LD datasets, freeing and enriching shared data between human and software agents. Below there are some benefits[23] that can derive from publishing and/or alignment of digital Repository content objects (resources, metadata,[24] research data) according to LD-enabling strategies:

1. possibility of linking, sharing, and querying (meta)data from different sources and formats. LD leads to organize a "silo" environment of disconnected resources from different Repositories in one space of structured connected data;

[22]The Open Data Handbook. Open Knowledge Foundation, 2010-2012, http://opendatahandbook.org/en.

[23]Benefits of the Linked Data Approach, http://www.w3.org/2005/Incubator/lld/wiki/Benefits; EC FP7 Support Action LOD-Around-The-Clock (LATC),http://5stardata.info.

[24]User Guide/ Publishing Metadata: "How to use DCMI Metadata as Linked Data. Publishing and Consuming Linked Data with RDFa", http://wiki.dublincore.org/index.php/User_Guide/Publishing_Metadata.

2. avoiding data redundancy (duplication) and keeping it updated;
3. cross-referencing to L(O)D authority and semantic files;
4. bookmarking of global encyclopedic cross-domain information (e.g. *DBpedia, Open Library data mirror in the Talis Platform, The Open Library, Freebase, GeoNames Semantic Web*) avalaible as LOD and reusing its parts;
5. directly processing data without being confined by the capabilities of any particular software, to perform data aggregation, calculations, visualisation, access, exporting, fine-granular control over the data items (e.g. load balancing, caching);
6. better measuring of data contributions in specific research disciplines. LD strategies can bring together the datasets living in disparate Repositories around the world that vary significantly by (or within) disciplines or even type of study;
7. new audience attracted by rich digital content developed on Repository LOD sets by means of APIs (Application Programming Interfaces) and web mash-ups often combining ”general“ APIs (Jarrar and Dikaiakos)
8. better user experience based on connected contextually relevant datasets. Users be more likely to visit again this or that Repository or Portal enhanced by LD-enabling strategies.

3 The importance of controlled vocabularies and semantic schemes

Controlled vocabularies and semantic schemes - such as lists of authority control including standard name identification, classification

systems, thesauri, topic maps, ontologies - are known generically as Knowledge Organization Systems (KOSs). KOSs provide a systematic way to better organize, access and retrieve knowledge inherent to information resources, through the mandate use of predefined, authorized and semantically expanded terms, with indication of different variations, spellings and misspellings, uppercase versus lowercase variants (Guerrini, Tillett, and Sardo). Without using KOSs in describing digital resources, both users and machines are stymied in their efforts to better access and aggregate them (Salo). Controlled vocabularies are maintained by an Authority (e.g. NACO Authority of the Library of Congress) ensuring that all terms are defined consistently and have well-defined relationships. In theory, any piece of information is amenable to authority control such as *personal* and *corporate names*, *uniform titles*, *series*, and *subjects*, trying to bring "structure and order" (to collocate materials that logically belong together but which present themselves differently) to the task of helping users to find information.

Assigning, for example, to an *author*, *subject*, *license* etc. a particular unique heading (term expressed by string or web address identifier), which is then used consistently, uniquely, and unambiguously to describe all references to that "piece" - can be combined into a database and called an *Authority File*. This files should be maintained and updated as well as "logical linkages" to other connected files/records should be provided by metadata practitioners and other information professionals. Different controlled and semantic vocabularies have been already published according to SKOS (Simple Knowledge Organization Systems)[25] RDF/S formalisms and released as L(O)D sets, in order to be comprehensible, shared and re-used among different actors on the web. Use of these KOSs to qualify certain metadata values could also facilitate LD-enabled linking in IRs, as

[25] http://www.w3.org/TR/2009/REC-skos-reference-20090818.

it was already demonstrated in the already mentioned *VOA3R* and *ScholarsArchive@OSU* Repositories.

Despite the availability of different SKOS/LD KOSs,

> Future work on Linked Data [in Institutional Repositories] should address gaps [...] Since most Thesis authors and many other Repository submitters do not appear in major Library Name Systems, these [controlled vocabularies published as LD] solutions are of limited help. What is needed is a Locally maintained Name database [...] for internal Name Authority based on the Simple Knowledge Organization System (SKOS) vocabulary (Johnson and Boock)

and possibly published as LD under Open licenses, which would allow for derivatives to be created (e.g. multilingual versions, connection with other LOD authority and semantic files). Normalized and semantically enriched (through URIs values) metadata terms could present a qualitative basis for high-tech navigation interface modules (e.g. faceted search[26]) to refine and expand search and retrieval results:

> applying Standard Subject Vocabularies and Classification Schemes is more expensive than assigning a few uncontrolled keywords [...] expenditures in development often result in greater efficiency and effectiveness for the end user. Use of a Standardized Subject Thesaurus or other Controlled Vocabulary, for example, can provide greater precision and recall in searching, and can enable future functionality, such as faceted subject browsing and dynamic searching of subject matter. (NISO Framework Working Group 58-59)

[26]EIFL, Knowledge without boundaries, http://www.eifl.net/faceted-search.

4 Some approaches, documents and principles for creating qualitative and extensible metadata elements describing IR objects

At various stages of an *information object's life cycle*,

> creators of digital objects should be encouraged to embed as much metadata as possible within the object before it is shared or distributed [...] Institutions should be aware that, depending upon the nature of their collections, a single Metadata Schema may not suffice for all their needs. Thus a judicious combination of metadata schemas may be the best solution for some materials.[27]

The metadata schema from CRUI Guidelines offers an extend use of 15 Unqualified (simple) DC metadata with additional refinements and elements. DC simple presents basic metadata elements to describe IR content objects, in order to support minimum interoperability among OAI-compliant Repositories by means of OAI-PMH protocol. Preferences to use DC metadata can be explained by its simplicity ("almost anyone can use it, or at least parts of it: hence, it is the metadata of choice for Institutional Repositories, where users upload their own works and create their own metadata"[28]), as well as by its high integrating capability (e.g. DC-Library Application Profile, Scholarly Works Application Profile, VOA3R AgRes AP Metadata Terms).
In order qualified metadata from Data Providers are not be flattened and depleted by harvesting OAI-PMH mechanisms, both Data and

[27] http://framework.niso.org/node/24.
[28] http://framework.niso.org/node/24.

Service Providers should support common and widely shared standards and protocols as well as qualitatively developed cross-walking schemes (mappings among schemas), limiting loss of data or their specificity.

It is a good practice when metadata elements motivated choice, along with their consistent compilation and encoding design approaches and requirements are declared in the appropriate IR Policies. These last are also important for the development of a widely-spread new trend for Repositories such as *Data Management Plans* (DMPs)[29] aiming to qualitatively support entire life cycle both of metadata and research data[30] complementing the context of deposited content objects.

With qualitatively programmed, encoded and widely cross -referenced metadata, "*Institutional Repositorie*s will be ultimately to form an *International Network* of indexed Repositories searchable from a single interface",[31] deploying a single virtual entry-point for exchanging and augmenting open bibliographic data improving the dissemination of research results in via Open Access. During the selection and development of metadata elements it would be appropriate to make the continuous confrontation with six NISO's (National Information Standards Organization)[32] principles for "good metadata". "*Good metadata*":

1. conforms to community Standards in a way that is appropriate

[29]Data Management Plans. Digital Curation Center, http://www.dcc.ac.uk/resources/data-management-plans.

[30]"Research Data", University of Bath, http://www.bath.ac.uk/research/data.

[31]Statement from the University of Oregon Libraries, http://library.uoregon.edu/diglib/irg/SB_Role.html.

[32]http://framework.niso.org/node/24. On February 2013 NISO launched a new initiative to develop Standard for "Open Access Metadata and Indicators" (standardized bibliographic metadata and visual indicators to describe the accessibility of Journal articles with respect to how "open" they are): http://www.niso.org/publications/newsline/2013/newslinefeb2013.html#report2.

to the materials in the collection, users of the collection, and current and potential future uses of the collection;

2. supports interoperability;

3. uses Authority Control and Content Standards to describe objects and collocate related objects;

4. includes a clear statement of the conditions and terms of use for the digital object;

5. supports the long-term curation and preservation of objects in collections;

6. are objects themselves and therefore should have the qualities of good objects, including authority, authenticity, archivability, persistence, and unique identification.

"Good" (qualitative) metadata requires an understanding of both data that is going to be described and standard/s by which such a description would be possible. The section "Metadata validations" (Conference of Rectors of Italian Universities (CRUI), Open Access Group, Italy 11-12) of CRUI Guidelines underlines that metadata quality, in turn, determines the quality of functions performed and services offered both by Repositories (Data Providers) and their aggregators (Service Providers), considering the context of interoperability within the OAI model. In creating "good metadata" elements, it is also worth referring to such an authoritative document as "User Guide/Creating Metadata" developed within DCMI Community.[33]

[33] http://wiki.dublincore.org/index.php/User_Guide/Creating_Metadata#Guidelines_for_the_creation_of_medium_content.

4.1 CRUI Guidelines: requirements for creation of qualitative IR metadata

To ensure metadata accuracy and their qualitative compilation during the self-archiving process of digital materials in the IR, CRUI Guidelines recommends to:

1. Assist users during self-archiving (based on the metadata insertion process) of their content objects. It may be possible through the establishment of facilities such as *metadata editors* with dynamic lists for auto-completion and capture/import of metadata values from different authoritative sources (e.g. internal and external authoritative files to control values of ”Responsible Body”, ”Subject”, ”Place” metadata).

2. Validate metadata inserted prior to its exposure to the final users and Service Providers. Effectiveness and efficiency of the metadata import/export are closely related to the use of Unique Identifiers (e.g. URI). It is a good practice when Unique Identifiers are assigned automatically within IRs platforms to research products and authors. Using Unique Identifiers as ”non-literal” data values describing metadata properties should reassure the stability of metadata elements they addressing to, as well as their interoperability among different systems. Moreover, the duplication of metadata values will be easily avoided, effective filters for the discovery of related resources (e.g. created by the same author), as well as efficient navigation tools can be developed. Unique Identifiers could be also of great importance in creating qualitative connections between research content and its evaluation processes (e.g. IRs as technical infrastructures for research management and assessment[34]). Effectively exploiting within networks the po-

[34]Institutional Repositories for Research Management and Assessment, on the

tential of Unique Identifiers assigned by IRs, alongside with CERIF (Common European Research Information Format) and other research data metadata standards as well as with applying of widely-accepted scientific disciplinary sector classifications, greater integration between Open Access Repositories (OAR) and Current Research Information Systems (CRIS e Euro-CRIS)[35] can be achieved. Currently,

> Many different research information systems (RIS) implement CERIF data model [which] has concepts of base relations and link relations (with role and temporal duration) [...] Several RIS providers had also published Web APIs using SOAP or REST technologies to support web applications and mash-ups with data from other systems. These APIs varied and were proprietary [...] Bringing "data islands" to a global, interconnected data space leveraging RDF, SPARQL, and OWL ontologies. In that context, reuse of well-established ontologies beyond FOAF, Dublin Core, and BIBO should be explored. (Jeffery and Corson-Rikert)

3. Provide each Repository with professional metadata support. Considering that the validation of metadata quality is an organizational management issue rather than a procedural one, it is a good practice to establish within each IR a support unit directed by metadata professionals.

In the near future

"Open Access scholarly Information Sourcebook" portal, http://openoasis.org/index.php?option=com_content&view=article&id=165&Itemid=335.

[35]The World Confederation of Open Access Repositories (COAR) and euroCRIS recently announced a strategic partnership. Specific attention will be paid to the domain of interoperability between different OA Repositories and CRIS to ensure appropriate management of research results, http://www.coar-repositories.org/news/eurocris-and-coar-join-forces-building-up-a-mutual-partnership-2.

> it is very likely that [all] local Repositories will be forced to employ a *quality metadata* content description and metadata harvesting system [as] Most leading citation databases consider metadata, or as the case may be metadata harvesting systems, conditional for integrating or monitoring the Repository [Moreover] In order to fulfil their mission and maintain a high quality Standard, these local Repositories have to seek and implement innovations in compliance with the latest technologies and information resources development so that their content can be unequivocally identified and meta-described with a view to content distribution. (Šimek 88)

The "metadata quality" concept recalls the concept of "trusted environment", which is being actively promoted within the frame of (certificated) *Trusted Digital Repositories*[36] developed in respect with the requirements of widely-accepted Standards, trusted recommendations and guidelines.

The core metadata elements presented by CRUI Guidelines are aiming to cover a basic description of the following types of digital content research objects: *Article, Patent, Book and Part of the book, Conference object, Paper of conference, Poster of conference, Annotation, Review, Doctoral Thesis, Master Thesis, Bachelor Thesis, Working Paper*. The Metadata schema that will be presented in the penultimate paragraph of this article will state:

- restructured and well defined metadata elements from CRUI Guidelines according to "LODE-BD" metadata groups of common properties;
- an extended number of metadata elements from "Guidelines" according to proposed "LODE-BD" mappings;

[36] Interesting contributions on this theme were released during International Conference 2012 "Cultural Heritage online – Trusted Digital Repositories", Florence, http://www.rinascimento-digitale.it/conference2012-culturalheritageonline-materials.phtml.

- choices for encoding of metadata elements from "Guidelines" according to "LODE-BD" strategies.

4.2 "LODE-BD" Recommendations

"LODE-BD" Recommendations are encompassing important components that a Data Provider may encounter when decides to produce sharable LOD-ready structured data describing bibliographic resources such as *Articles*, *Monographs*, *Theses*, *Conference Papers*, *Presentation Material*, *Research Reports*, *Learning objects*, etc. (Subirats and Zeng 4). "LODE-BD" aims at addressing two questions:

1. how data - hosted by diverse Open Repositories - can be better exchanged across Data Providers;

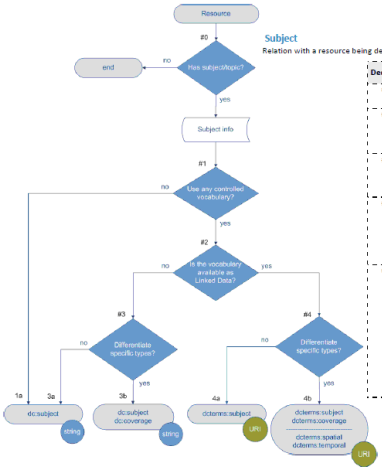
2. how to encode this data within LOD-enabled metadata.

"LODE-BD" provides a selected number of widely used metadata standards and the emerging LOD-enabled vocabularies. Metadata terms from the DCMES (dc:) and DCMI Metadata Terms (dcterms:) are the fundamentals, while metadata terms from other namespaces are supplemented when additional Repository needs should be met. These supplemented metadata are including the namespaces from BIBO Ontology, AGLS Metadata Standard of the Australian Government Locator Service, eprint (UKOLN Eprints Terms, SWAP), and MARCrel (MARC List for Relators). All metadata terms are presented in a crosswalk table. Based on different cross-referred metadata namespaces and controlled vocabularies, the descriptive metadata would of course benefit in terms of their consistency, extensibility, semantic and authority richness.
Referring to the development stage of metadata terms according to "LODE-BD", Repository managers should address the following issues:

- What kinds of entities and relations there should be involved in describing and accessing bibliographical resources?
- What properties should be considered for publishing meaningful/useful LOD-ready bibliographic data?
- What metadata terms are appropriate in any given property when producing LOD-ready bibliographical data from a local database? (Subirats and Zeng 5)

4.2.1 "LODE-BD" *Decisions Trees*. Between "literal" and "non-literal" metadata values

The real strength of "LODE-DB" development stages are Decision Trees (Figure 2) designed to facilitate the selection of the appropriate strategies adjustable to Data Providers according to their local needs, while all moving towards the goal of metadata exchange and re-use f their values on the Web of Data.



Subject

Relation with a resource being described: Resource has subject/topic.

Decision	Question	Answer	Action	Value Type	Metadata Term	Examples Value
#0	Has subject/topic?	No	End			
		Yes	Continue to #1			
#1	Use any controlled vocabulary?	No	1a	String	dc:subject	paddy Pacific Islands & Oceania 19th century
		Yes	Continue to #2			
#2	Is the vocabulary available as linked data?	No	Continue to #3			
		Yes	Continue to #4			
#3	Differentiate types of subjects?	No	3a	String	dc:subject	Rice Pacific Islands Nineteenth century
		Yes	3b	String	dc:subject dc:coverage	Rice Pacific Islands Nineteenth century
#4	Differentiate types of subjects?	No	4a	URI	dcterms:subject	http://aims.fao.org/aos/agrovoc/c_6599 [2] http://aims.fao.org/aos/agrovoc/c_5487 [3]
		Yes	4b	URI	dcterms:subject dcterms:coverage dcterms:spatial dcterms:temporal	http://aims.fao.org/aos/agrovoc/c_6599 [2] http://aims.fao.org/aos/agrovoc/c_5487 [3] http://aims.fao.org/aos/agrovoc/c_5487 [2] http://id.loc.gov/authorities/sh85091984 [4]

Figure 2: LODE- BD *Decision Tree* and explanation table to describe and encode "Subject" information (Subirats and Zeng 31-32)

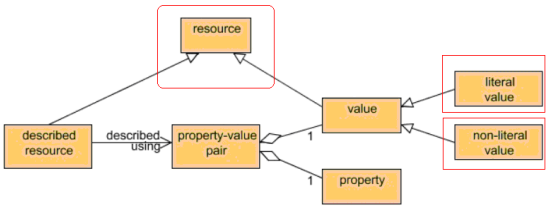
LODE-BD *Decision Trees* are developed to support description and encoding of the following metadata elements:

1."Title information"	Title/Alternative title
2. "Responsible Body"	Creator. Contributor. Publisher
3. "Physical Characteristics"	Date. Identifier. Language. Format/Medium. Edition/Version. Source
4. "Holding/Location information"	Location/Availability
5. "Subject Information"	Subject/Topic
5. "Description of Content"	Description/Abstract/Table of Contents. Type/Form/Genre
6. "Intellectual property rights"	Right Statements
7. "Usage"	Audience/literary indication/ education level
8. "Relation"	Relation between resources. Relation between agents

All *Decision Trees* are starting from the property describing a *Resource* instance and are delivered in flowcharts with various acting points, giving a "step-by-step" solutions for decisions to be made, further explained within text based tables, with notes, steps, and examples matching encoding suggestions, whenever essential. Within these explanations two types of metadata values - that can be chosen to qualify certain metadata properties - are provided:

1) **Literal value**. *This is typically a string of characters using a Unicode string as a lexical form, together with an optional language tag or data-type, to denote a "Resource".*

Examples of metadata namespace **dcterms:alternative** *"A Feast of Beans"* **dcterms:available** *"2006-07"*^^dcterms:W3CDTF ...



2) **Non literal value**. *This value presents physical, digital or conceptual entities indicated by Unique Identifiers.* LODE-BD "Decision Trees" help Data Provider to evaluate the existing gap between current use of literal values and their evolution to a LD approach (i.e. by using "non-literal" URI values from Controlled Vocabularies and other LD sets).	Examples of metadata namespaces: **dcterms:conformsTo** <http://www.w3.org/2001/XMLSchema> **dcterms:contributor gnd:135066719** gnd:135066719 *foaf:familyName "Elliott"; foaf:givenName "Missy" ; foaf:nick "Missy E"...*

Properties of some DC metadata namespaces ("dc:" and "dcterms:") – as it is demonstrated within LODE-BD explanatory tables and good described in the User Guide "*How to use DCMI Metadata as Linked Data*"[37] - may be qualified both by "literal" and "non-literal" values. However, to produce LD-enabled metadata that can be easily harvested by Service Providers on the web, the use of "dcterms:" namespace properties qualified by "non-literal" (URI) values is recommended.

The pragmatic relevance of LODE-BD *Decision Three*'s approach for producing LOD-enabled metadata is that each Data Provider can highlight within the concrete *Decision Tree* its own decision paths, marking the metadata terms to be used as well as choosing vocabularies and standards on their support. "LODE-BD" are not limited to subject-specific domains, thus being appropriate for use by any Data Provider accordingly to local needs. Nevertheless, "Decisions regarding what Standard(s) to adopt will directly impact the degree of LOD readiness of the bibliographic data" (Subirats and Zeng 1,3).

[37] http://wiki.dublincore.org/index.php/User_Guide/Publishing_Metadata#Properties_of_the_terms_namespace_used_only_with_non-literal_values.

4.2.2 "Intellectual Property Rights". Controlled vocabularies LD-enabling

Before a certain resource is published, it is important to decide under which License it will be presented to users. As it was already mentioned in connection with "Tim's 5 star deployment scheme", it is advisable to publish digital contents on the Web under an *Open License*, in order they can be freely: shared (copied, distributed and transmitted), remixed (adapted), used by any 3rd party (including commercial) to produce derivates, anyhow with attribution the work to the author or licensor. This should be applied even more to research resources produced in public domain (De Robbio). However, considering that some IR resources could be connected with issues of: "Embargoed access" (the resource is of Closed Access, until released for Open Access on a certain date), "Restricted access" (Open Access, but with restrictions) and "Closed access" (opposite of Open Access), aside "Open" also "Not open" licenses may be used to denote "dcterms:rightsHolder", "dcterms:licence" metadata properties. Authors can find useful informational support about *Intellectual Property Rights* and *Licenses* within good compiled services like SHERPA/Romeo "Publisher copyright policies & self-archiving"[38], "Diritto d'autore" (service offered by University of Padova Library System).[39]

Some decision steps to choose a particular License describing the use of resource are presented in Figure 3 on the next page, as an extension to "LODE-BD Decision Tree" referring to "Rights: Situations and best practices for encoding the data".

After a certain License is chosen, a value ("literal" and/or "non literal") identifying officially the License type should be encoded in the appropriate metadata property (ies), as according to "LODE-BD"

[38] http://www.sherpa.ac.uk/romeo.

[39] http://www.cab.unipd.it/servizi/diritto-dautore.

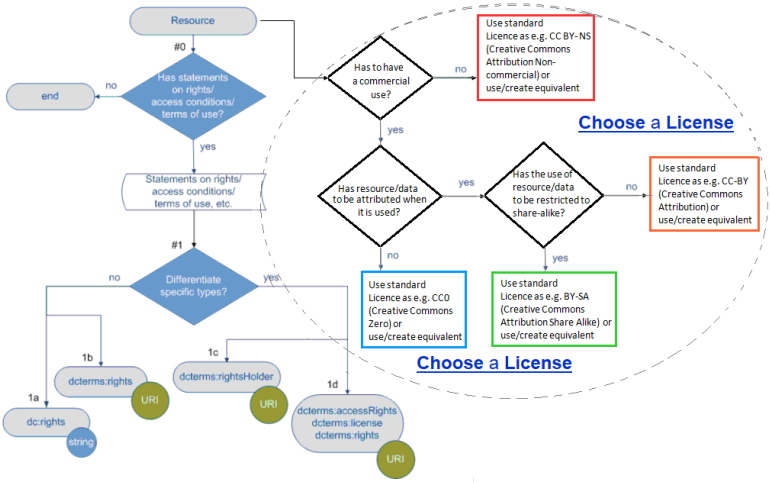
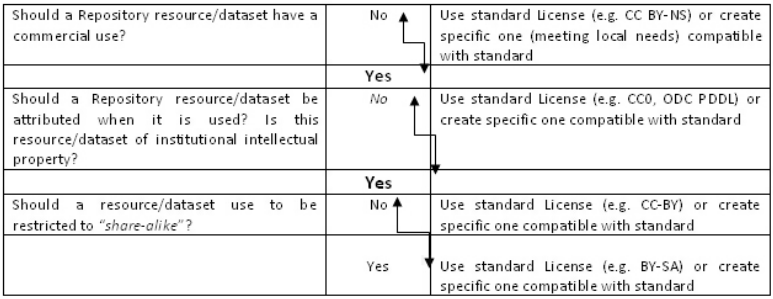


Figure 3: *Decision Tree*: a choice of a License for publishing Repository datasets

Figure 4: The extension provided to the LODE-BD Decision Tree in Figure 3 on the preceding page can be interpreted as follows:



encoding strategies. In Appendix … some wide-used (”Open” and ”Not-open”) Licenses are provided, together with their ”non literal” (URI) legal identifiers, which can help ”Intellectual Property Rights” metadata to become LD-enabled.

4.2.3 ”LODE-BD”: mapping of metadata with ”schema.org” mark-ups

In the ”LODE-BD” Appendix 4 cross-walks from certain metadata elements to ”schema.org” mark-ups are provided. ”Schema.org” mark-ups - natively relevant for webmasters[40] - (i.e., html tags used by webmasters to markup their pages in ways recognized by major search engines including *Bing*, *Google*, *Yahoo!* and *Yandex*) can be also used to improve representation and search of bibliographic information on the web. When you are exploring how data is inter-related on the web in order to learn more about patterns or things implicit in the data, is when it would be of benefit not only consider a RDF graph or LD view but also ”schema.org” mark-ups. This is

[40] http://schema.org; http://schema.org/docs/full.html.

particularly relevant to intelligence applications, scientific research and many other types of applications exposed on the web.

Different bibliographical data has already been supported by "schema.org" mark-ups in services such as, for example, *WorldCat.org*, *Data.bnf.fr*,[41] VOA3R Open Access Repository. The reason why "schema.org" is included in the "LODE-BD" is essentially through two reasons:

1. the benefit of creating micro-data by individual sources, e.g. webmasters or authors themselves when they publish data on the web, instead of going through a Repository and get exposures. It is another way to expose resources. It does not replace any metadata schema as, in case of "LODE-BD" proposed schema, it is to be complementary to DC metadata terms;

2. because it is multiple schemes, many of the properties used for bibliographic description also are used by other types of resources. Assuming there will be more resources use "schema.org", the chance of interoperability is high. Repositories also can harvest from those data which would have various benefits.

The "Schema Bib Extend Community Group"[42] within the "W3C Web Schemas Task Force" is preparing different proposals for extending "schema.org" vocabularies to improve representation and search of bibliographic data on the web.

[41] https://www.oclc.org/en-US/news/releases/2012/201238.html; http://data.bnf.fr/docs/databnf-presentation-en.pdf.

[42] http://www.w3.org/community/schemabibex.

5 Metadata schema for description of IR digital content objects

In the Metadata Schema as according to CRUI Guidelines, the mapping to OAI_DC metadata will be provided.
The metadata elements from CRUI Guidelines consider the Unicode encoding standard, important for the consistent representation and handling of text expressed in most of the world's digital writing systems, using XML schema as the primary medium based on "mix and match" method combining elements and sub-elements, related attributes, and controlled attribute values throughout the element sets. "LODE-BD" promotes the encoding of metadata elements within RDF/XML schemas to support their semantic consistency required today in most digital environments. Both CRUI Guidelines and "LODE-BD" assume that the metadata they provide could be more complex and structured, first of all in view of creating a more balanced framework that may allow to accommodate better different metadata models according to different Repository local needs for representation and management of their digital content objects.
The aim of alignment metadata elements from CRUI Guidelines according to "LODE-BD" is to show how metadata terms selected for the description of IR digital objects can be enhanced by encoding "LODE-BD" strategies. Summarily, such an aligning will lead to:

1. radically-improved metadata workflows. Data integration and reusability will save time for the development of new metadata indexes;

2. better IR resource description and discovery (searching and browsing) on the Web of Data. IRs will be able redirect their users straight from the Repository discovery interfaces to the connected knowledge DMSs (Data Management Systems)

Data Hubs provided by different related datasets in the LOD Cloud. The Repository contents will increase tremendously in their visibility and integration on the Web;

3. better data exchange through collectively shared data, based on common LD values;

4. the development of common search interface like "Institutional Repository WorldShare Platform" (see experience of "WorldCat Local"[43]) for search of IRs digital contents interconnected through LD-enabled metadata values worldwide;

5. creation, sharing and use of new applications enhancing the dissemination channels and accessibility of L)O)D sets through IR services, contributing qualitatively to Open Research Commons space[44] (White).

The metadata schema presented in Appendix can be considered an a tentative to create an application profile (AP) for IRs objects based on DC metadata (presented by CRUI Guidelines) and modeled according to "LODE-BD" (structuring metadata in categories; encoding strategies based on motivated use of "literal" and "non-literal" values; choice of cross-walking to more specific metadata terms and to "schema.org"). The aim of this presentation is also to show the usefulness of "LODE-BD" Recommendations to enhance expressive quality of IRs DC descriptive metadata.
The concept of AP was emerged within the DCMI as a way to declare which elements from which namespaces would be better to use in a particular application or project. Metadata elements can be

[43]Single-search access to 1.071+ billion items from your library and the world's library collections, http://www.oclc.org/worldcatlocal/default.htm.
[44]http://aims.fao.org/community/open-access/blogs/building-institutional-repositories-global-research-commons.

combined together by implementers in different ways, optimizing descriptive and system local needs.
The presented metadata profile can be considered as a part of "Design-time" implementation strategy defined by "LODE-BD". Both "Design-time" and "Run-time" LODE-BD strategies will be discussed in the next conclusive paragraph.

6 LODE-BD "Design-time" and "Run-time" implementation strategies

To align and implement descriptive metadata according to "LODE-BD" strategies, Data Provider may follow next two options (Subirats and Zeng 44) (Figure 5 on the following page):

1. "Design-time", i.e. changing current metadata model, replacing it with "LODE BD" proposals for selection and modeling of descriptive metadata. The choice of this strategy means also some changes to a current metadata database and services accessing it.

2. "Run-time" (on the fly)[45] option means that - while keeping the current metadata model and database structure unchanged - Data Provider should add a *conversion service* mapping and translating chosen metadata values from "literal" to "non-literal", following to "LODE-BD" Decision Tree's encoding strategy.

In Figure 5 on the next page due attention is given to the description of the "Run-time" strategy, pointing on conversion of "Subject"

[45]Example: those using OAI protocol, such as National Science Digital Library (NSDL): http://nsdl.org/contribute, here the wiki (for Contributors and Developers) can be followed to find the documents.

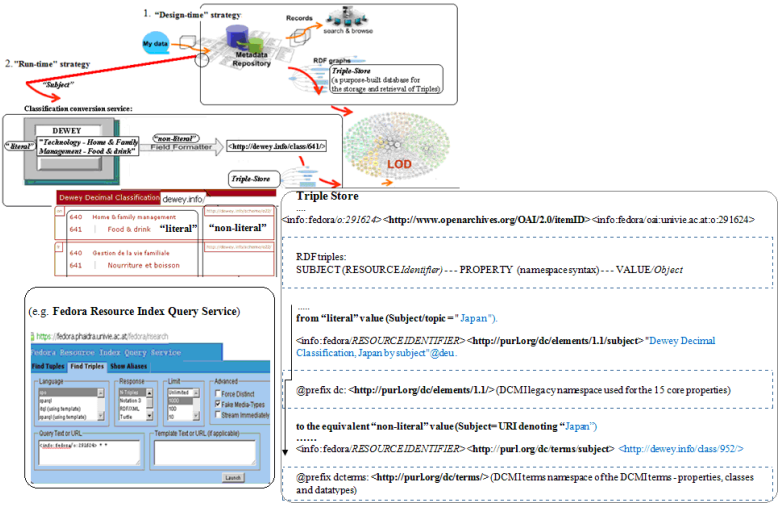


Figure 5: Metadata modelling according to LODE-BD ”Design-time” and ”Run-time” strategies.

metadata value from "literal" to "non-literal" value language. As is it shown, the "Subject" information can be described both by "literal" ("Japan" from Dewey Classification language @deu) and "non-literal" values (URI to equivalent concept from "Dewey.info" LD service). Both "literal" and "non-literal" values can be traced as graphs (Figure 6) in *Triple Store*.

Figure 6: Exemplification of RDF graphs registered in *Triple Store*.

Triple store is a purpose-built database optimized for the storage and retrieval of triples (RDF), representing data entities composed of Subject (Resource[46]) - Property(Predicate) – Object (Value). Triple stores can be seen like the advantage for performance of Data Providers, also because all the information traced in a Triple store can be retrieved via a query language (e.g. a query language of Fedora Resource Index Query Service; Figure 5 on the preceding page). In addition to queries, triples can usually be imported/exported using RDF LD-enabling and other formats. A "non-literal" URI value denoting the "Subject" information in relation to a certain Repository resource – can be imported by other Providers (implementing Triple stores over querying of graph-based RDF models) using the same or related scheme(s) to qualify the "Subject" information:

[46] "To benefit from and increase the value of the World Wide Web, agents should provide URIs as identifiers for Resources", http://www.w3.org/TR/webarch/#uri-benefits

> if another party might reasonably want to create a hypertext link to it, make or refute assertions about it, retrieve or cache a representation of it, include all or part of it by reference into another representation, annotate it, or perform other operations on it.[47]

This assertion corresponds to the fourth and fifth stars of the mentioned "Tim's 5 star deployment scheme": (4) "use URIs to denote things, so that other agents can point at your datasets", (5) "*combine* the data safely with other data in URIs global scheme to provide context", thus contributing to the richness of content and context exchange within the global Linked Open Data space and, therefore, on the Web of Data. Anyhow, "triplifying" data by automatic script should be avoided as it is not the same as developing well-structured triples suitable for Repository applications. Proper data modeling is an essential first step in any implementation. Attempts to automatically generate billions of RDF "triples" and publish them on the Web is not the same as producing high quality data sets of properly modeled data, according to Standards, Recommendations and Guidelines.

> Simply transforming database schemas into RDF does not create Linked Data [...] To create automatic links between RDF triple stores on the web should be possible, otherwise there is a risk of creating RDF silos. The easiest way to facilitate the establishing of automatic linking between datasets is the use of Standard Vocabularies, including Standard Vocabularies for describing data/metadata elements and Standard Vocabularies for indicating values.[48]

The way how information content and context exchange can be obtained in an information service "on the fly", is good demon-

[47] http://www.w3.org/TR/webarch/#uri-benefits.
[48] http://aims.fao.org/linked-data/getting-started.

strated within the AGRIS searching platform[49] (Figure 7). Through "Open_AGRIS Beta" application, AGRIS platform supports the search and retrieval of resources described in its system by AGROVOC Thesaurus LOD set, included in AGRIS application through web service. Through AGROVOC "non-literal" (URIs) values, "Subject" metadata terms can be connected with different resources on the web, whose topic description is based on values of AGROVOC and related online datasets.

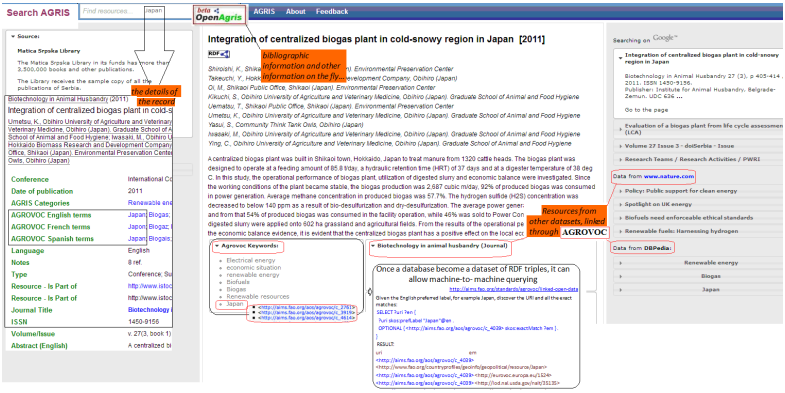


Figure 7: Search and retrieval of AGRIS resources and related datasets described by AGROVOC Thesaurus.

The search functionalities proposed by "AGRIS" can be considered one of the best examples presenting all components working together through "Subject" metadata values, supported by the controlled AGROVOC LD-enabled vocabulary using URIs to identify concepts and mapping concepts.

[49] International Information System for the Agricultural Sciences and Technology, http://agris.fao.org.

LODE-BD Decision Trees' metadata encoding strategies are based on the concept "usefulness to others". In the context of IR, their usefulness can be interpreted in terms of developing a rich IR LD-enabled metadata schema that can be re-used by different web actors, contributing to enhance visibility and semantic interoperability of IR digital content objects on the global scale.

References

Bechhofer, Sean, et al. "Why linked data is not enough for scientists". *Future Generation Computer Systems* 29.2 (2013): 599 –611. (Cit. on p. 113). Web. <http://www.sciencedirect.com/science/article/pii/S0167739X11001439>.

Berners-Lee, Tim. "Linked Data. Designed Issues". (2006). (Cit. on p. 117). Web. <http://www.w3.org/DesignIssues/LinkedData.html>.

Conference of Rectors of Italian Universities (CRUI), Open Access Group, Italy. "Linee guida per la creazione e la gestione di metadati nei Repository Istituzionali". (2012). (Cit. on pp. 115, 123). Web. <http://www.crui.it/HomePage.aspx?ref=2066>.

De Robbio, Antonella. "Forme e gradi di apertura dei dati. I nuovi alfabeti dell'Open Biblio tra scienza e società". *Biblioteche oggi* 30.6 (2012). (Cit. on p. 131). Web. <http://www.bibliotecheoggi.it/content/201200601101.pdf>.

De Robbio, Antonella and Silvia Giacomazzi. "Dati aperti con LODe". *Bibliotime* 10.2 (2011). (Cit. on p. 113). Web. <http://eprints.rclis.org/16440>.

Glaser, Hugh and Harry Halpin. "The linked data strategy for global identity". *IEEE Internet Computing* 16.2 (2012): 68–71. (Cit. on p. 117). Web. <http://eprints.soton.ac.uk/333924/>.

Global interoperability and Linked Data in libraries. Special issue of "JLIS.it". 2013. (Cit. on p. 111). Web. <http://leo.cilea.it/index.php/jlis/issue/view/536>.

Guerrini, Mauro, Barbara Tillett, and Lucia Sardo. *Authority control. Definizioni ed esperienze internazionali. Atti del convegno internazionale, Firenze, 10-12 febbraio 2003.* Firenze: Firenze University Press, Associazione Italiana Biblioteche, 2003. (Cit. on p. 119). Web. <http://www.fupress.com/Archivio/pdf/4383.pdf>.

Šimek, Pavel. "Using Metadata Description for Agriculture and Aquaculture Papers". *Agris on-line Papers in Economics and Informatics* 4.4 (2012). (Cit. on p. 126). Web. <http://online.agris.cz/files/2012/agris_on-line_2012_4_simek_vanek_ocenasek_stoces_vogeltanzova.pdf>.

Jeffery, Keith G. and Jon Corson-Rikert. "euroCRIS and VIVO. Part II Cooperation as Strategic Partners". (2012). (Cit. on p. 125). Web. <http://www.vivoweb.org/files/presentations/12Fri/euroCRIS_LOD_and_\%20VIVO.pdf>.

Johnson, Thomas and Michael Boock. "Linked Data Services for Theses and Dissertations". *Proceedings of the 15th International Symposium on Electronic Theses and Dissertations*. Lima. 2012. (Cit. on pp. 114, 120). Web. <http://hdl.handle.net/1957/32977>.

NISO Framework Working Group. *A Framework of Guidance for Building Good Digital Collections*. 3rd ed. Baltimore: NISO, 2007. (Cit. on p. 120). Web. <http://www.niso.org/publications/rp/framework3.pdf>.

Salo, Dorothea. "Name Authority Control in Institutional Repositories". *Cataloging and Classification Quarterly* 47.3/4 (2009). (Cit. on p. 119). Web. <http://minds.wisconsin.edu/handle/1793/31735>.

Sequeda, Juan F. "Consuming Linked Data". Proc. of Semantic Technology Conference, 2011. (Cit. on p. 112). Web. <http://www.slideshare.net/juansequeda/consuming-linked-data>.

Subirats, Imma and Marcia L. Zeng. "LODE-BD Recommendations 2.0 : How to select appropriate encoding strategies for producing Linked Open Data (LOD)-enabled bibliographic data". (2012). (Cit. on pp. 113, 115, 117, 127, 128, 130, 137). Web. <http://aims.fao.org/lode/bd>.

White, Wendy. "Institutional repositories: contributing to institutional knowledge management and the global research commons". *4th International Open Repositories Conference*. 2009. (Cit. on p. 136). Web. <http://eprints.soton.ac.uk/48552/>.

IRYNA SOLODOVNIK, Scuola Dottorale Internazionale degli Studi Umanistici (SDISU), Università della Calabria.
iryna.solodovnik@gmail.com

Solodovnik, I. "Development of a metadata schema describing Institutional Repository content objects enhanced by "LODE-BD" strategies". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8792. DOI: 10.4403/jlis.it-8792. Web.

ABSTRACT: Based on "Guidelines for metadata creation and management in the Institutional Repositories" (CRUI, Italy, 2012) and "LODE-BD Recommendations" (AIMS, 2012), and exploring other principles and strategies for qualitative development of metadata representing contents and properties of digital contents, this article presents the specific metadata profile for description of Institutional Repository information resources. This profile is allocated within a metadata schema provided by well-defined metadata terms, compilation specifications, and alignment (mapping) strategies to more specific metadata terms and value properties enabling basic metadata to become more efficient in authority control context, richer in their semantic profiles and more accessible and usable on the web by means of Linked Data sets. Developing and implementing metadata schemas aligned completely or partially with Linked Data paradigm will provide metadata exchange among different Linked Data-enabling repositories, potentiate semantic relationship browsing and querying of their contents, enable their participation in the Linked Open Data cloud and contribution to an open research commons space.

KEYWORDS: Authority file; Institutional repositories; Knowledge management; Linked data; Metadata; Persistent identifiers.

ACKNOWLEDGMENT: The author would like to thank for some information support from AIMS members: Imma Subirats and Marcia Lei Zeng.

Submitted: 2013-02-23
Accepted: 2013-04-25
Published: 2013-07-01