Guidelines to producing structured interoperable data from Open Access Repositories

An example of integration of digital repositories of higher educational institutions LatAm

Nelson Piedra, Janneth Chicaiza and Jorge Lopez-Vargas
Universidad Técnica particular de Loja
Departamento de Ciencias de la Computación
Loja, Ecuador
{nopiedra, jachicaiza, jalopez2}@utpl.edu.ec

Edmundo Tovar Caro
Universidad Politécnica de Madrid
Dpto. Lenguajes y Sistemas Informáticos e Ingeniería Software, Madrid, Spain
edmundo.tovar@upm.es

Abstract— One of the fundamental concepts of Open Educational Resources (OER) is "the ability to freely adapt and reuse existing pieces of knowledge." The application of Semantic Web approach and Linked Data technologies to Open Education seeks to turn data and metadata from open educational repositories into actionable interoperability for the improvement of discovering, using and reusing of OER. Interoperability is not an end in itself. Instead, optimizing the level of interoperability has societal and educational value as a means to others purposes. Interoperability can have a positive impact on open innovation, user choice, ease to reuse and adapt educational materials, global discovery of open and diverse content, among other things. This paper reports on the implementation of Linked Open Data for open access repositories in a new interoperable and global open educational ecosystem. The goal is to improve the metadata interoperability between various collections of open material, so as to facilitate the discoverability and subsequent combining, remixing, or adapting OER; that is, OER data should be easily accessible to any user: human being or a machine agent. This work addressed two challenges in the OER ecosystem: providing evidence of globally discoverability and reusability academic resources. Although there is much further potential for teaching and learning to realize, linked open data is a critical enabler of global and interoperable OER ecosystem.

Keywords— Digital Library, Linked Data, Semantic Web, RDF, OAI-MPH, Ontology

I. INTRODUCTION

The Internet has created a global scenario in which the conditions of interoperability [10] guarantee the discovery, distribution, and re-usage of digital resources, namely those are not restricted by issues of local management, technological models or tools.

The Open Archive Initiative (OAI) was created in order to achieve inter-operability between various digital library repositories. This initiative thus proposes a protocol for harvesting metadata called OAI-PMH [14]. OAI-PMH is based on open standards that guarantee the automatic interoperability between emitters and receptors of digital resources, i.e. by incorporating the independence of the software and incentivizing technological neutrality and innovation. Although OAI-PMH facilitates the exchange of metadata on the Web, problems may still persist when integrating data that is extracted from various repositories. Within an open environment such as the Web, however, it is not always possible to standardize processes for the description and publication of metadata. Therefore, each institution individually manages various formats or schemas of metadata or vocabulary.

What's more, in order to reduce technological barriers and to integrate information related to digital libraries, that is, within a heterogeneous environment, we aimed to focus on Web Semantic technologies, namely Linked Data (LD). Within this context of using the Web as a global repository of linked data, it was noted that significant advances had been made in extracting and recovering useful information for users, as well as in the processing of meaning (semantics of information), and the recovery of knowledge that forms part of web pages, including that which concerns the interpretation of specific searches of sentences within a given context, or related to the specific needs of the users.

From the point of view of access to metadata in distributed repositories, the Open Archives Initiative (OAI) proposes a protocol for the interchange and harvesting of metadata called OAI-PMH. This protocol provides a low degree of inter-operability. However, by using an approach based on semantic technologies, the inter-operability of data can achieve a higher level. Thus, each digital library can maintain its specific local qualities. In addition, it will not be necessary to reassign them in order to standardize the exchange, re-usage or harvesting of digital resources.

We present a general framework for the publication of OER that were extracted from digital repositories such as OAI-PMH, that is, by following the principles of Linked Data. The objective of this study is to improve the integration and the inter-operability of resources stored on Digital Libraries. In the following section, we describe the technology and the current proposals for improving the inter-operability of metadata. Section 3 describes the proposed framework for the gathering of metadata and subsequent publications of LD. The proposal facilitates a diversity of methods and standards that are incorporated into the processes of each digital resources provider. In addition, section 4 provides details from a pilot study, which was applied to the analysis of open digital repositories. We present a process for the extraction of metadata and for the creation and publishing of LD. Section 5 describes specific scenarios concerning the usage of the newly-created data. Finally, section 6 covers the respective conclusions and future areas of research.

II. Background and Related Work

A. Enablers for collaboration between community of Open Educational Resources

Interoperability plays a crucial role in integration of information systems, and content sharing in the education environment. There are many formal standards bodies and specifications (in areas such as metadata, content, learner information, and platforms) that have become involved in educational technology standardization, including IEEE, the International Standards Organization (ISO), and IMS Global Learning Consortium. The development and implementation of specifications and standards is not a simple and straightforward process. The stakeholder's interests may conflict at different stages of the standards development and implementation projects.

There is a need for a consistent approach to allow different groups of stakeholders, experts and users to communicate and collaborate effectively in the creation and development of open education initiatives. However, several issues and barriers which prevent the integration and interoperability, hinder technology innovation in open education initiatives, particularly: complex and inflexible standardization processes; insufficiency of inclusiveness multi-stakeholder collaboration and participation, in the production and adoption of specifications and standards; lack of capability to adapt and/or create derivative works from specific standards (adaptation ability is desirable because communities and applications differ and one size will not fit all); restricted agility, reuse and lead to unnecessary costs.

In order to support educational technology innovations there is a need for the development and adoption of interoperability approach that are more open, less governed and aligned more closely with agile development models. This approach should be more open and transparent from developers, users and implementers in the community of open education that have grown up with the Web. In consequence, an interoperability ecosystem must deliver pluggability as a core aspect of its approach to the global datasets. This work describes the opportunity for a semantically interoperable ecosystem and explores how openness enables rapid innovation in academic context.

B. The OAI-PMH Protocol

From a technological point of view, digital libraries are repositories that store digital objects and that use OAI-PMH to expose their metadata. OAI-PMH is a protocol that was proposed by the Open Archives Initiative and which facilitates the extraction of metadata (which is described according to specific formats or schemas of metadata), i.e. from digital repositories. In order to obtain metadata, data servers were utilized to perform requests known as verbs. OAI-PMH supports 6 verbs that facilitate the access of information related to digital repositories, the metadata formats, the collection of resources, and the detailed description of each resource.

OAI-PMH is compatible with many tools that aid the creation of institutional repositories such as Eprints (http://www.eprints.org/), DSPACE (http://www.dspace.org/), Fedora (http://fedorarepository.org/), among others. OAI-PMH divides this phenomenon into providers of data and providers of services. The former are repositories that expose their metadata by means of OAI-PMH. The latter, which are also called "harvesters or collectors", develop services of aggregate value based on metadata, which are obtained by providers. Each repository independently stores its digital objects for the OAI-PMH.

C. Dublin Core (DC) Metadata Initiative

The Dublin Core or the Dublin Core Metadata Initiative (DCMI, http://www.dublincore.org) is the most widely used meta-information initiative at the global scale.

DCMI is used to describe the metadata of digital resources. To maximize the potential of interoperability with other data collections, it was necessary to use DC as the basis of these metadata schemas. More recently, however, the DC series has become an operational infrastructure for developing Web Semantics. Among the DC metadata that are used to describe web content, we can mention: Title, Subject, Description, Source, Language, Creator, Publisher and Rights. Despite the basic elements (none of which are obligatory), there are other mechanisms that can be used to adapt the DC to concrete information needs. This enables this model of metadata to be applicable to any digital information service or system.

OAI-PMH has come into widespread use by digital repositories around the world, but without the possibility of unification and interoperability with other datasets, the victory could be short-lived. There's a pressing need for an open, transparent, and semantically interoperable ecosystem. The hard dimension of interoperability is semantics, no tech. This approach or standard creates a "one size fits all" manner of doing "something" when there are many innovative and competing ways to do that "something", without limit the goals of open education.

D. Integration Approachs

Integration facilitates the combination of existing information resources in various contexts. It provides users with a unified view of such resources and can also serve as a source of data for various applications [11]. For Example, the LINDAT/CLARIN Centre for Language Research Infrastructure provides technical background and assistance to institutions or researchers who wants to share, create and modernize their tools and data used for research in linguistics or related research fields. The project also provides an open digital repository[1] and archive open to all academics who want their work to be preserved, promoted and made widely available.

---

[1] See https://lindat.mff.cuni.cz/en/about-lindat-clarin

Various communities can benefit from adopting a integration focus and using semantic interoperability within the following contexts: (a) the movement of open educational resources [16]; (b) where users in general provide access to a richer series of information from various repositories (federated searches); (c) where librarians and archivists can have access to shared data to describe their resources and to reduce redundancy; (e) where Web Developers face reduced problems of heterogeneity by using various formats or data semantics [2]; (f) where researchers can extract knowledge from open repositories, for example, by using search experts in a pre-determined field, in detecting of new research themes, and for the analysis of scientific networks [9], among others.

To achieve the above goal, first it is necessary to resolve problems of heterogeneity at the level of: (a) repositories; (b) data formats; (c) schemas of metadata, and (d) vocabularies or data dictionaries. To approach each of these dimensions, various proposals have been developed. The lowest level of inter-operability (i.e. at the level of repositories) is achieved with protocols such as OAI-PMH. This level facilitates the extraction and exchange of metadata, although its usage does not ensure the re-usage and integration of information. To guarantee the semantic inter-operability between various collections or schemes [6,7] propose the creation of links between equivalent objects. Within this context, the usage of open vocabularies and the creation of open data in legible formats for machines are essential to integrating elements from various repositories, namely those that do not focus on prior standards and agreements among providers.

The potential uses of LD into Open Access Repositories to include the ability to enhance discovery, search and reuse of OER. By encoding OER in a machine-readable format, Web brings these advantages into an interoperable ecosystem and can continue to improve discovery systems on the Web. OER community can build OER apps that reflect the global linked nature of OER and relationships between the entities in them.

Recent cases in the creation of LD have been presented in works such [1, 12, 15, 18, 19]. Within the same line of thinking, that is, by means of their current work, the authors aim to contribute to improve the exchange, the re-usage, sharing, and enrichment of data, as well as the institutional and academic collaboration of digital libraries

III. TOWARDS THE INTEGRATION OF REPOSITORIES VIA LINKED DATA

Semantic Web technologies and, more precisely, Linked Data are changing the way information is stored, published connected, and exploited. The term "Linked Data" refers to a set of best practices for publishing and connecting structured data on the Web.

Web Semantics as envisaged by Berners-Lee [3] adds meaning to the Web of Documents, that is, with the aim of providing an environment where it is possible to access the data contained on websites, as well as automatically processing information more precisely and completely. In this work, the authors focus on the concept of Web Semantics from the perspective of processing vast volumes of LD. This vision implies that the data are stored in a globally distributed database [8].

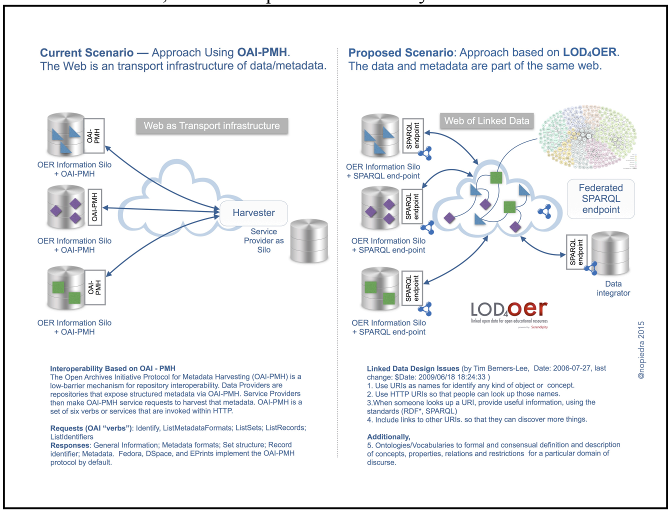


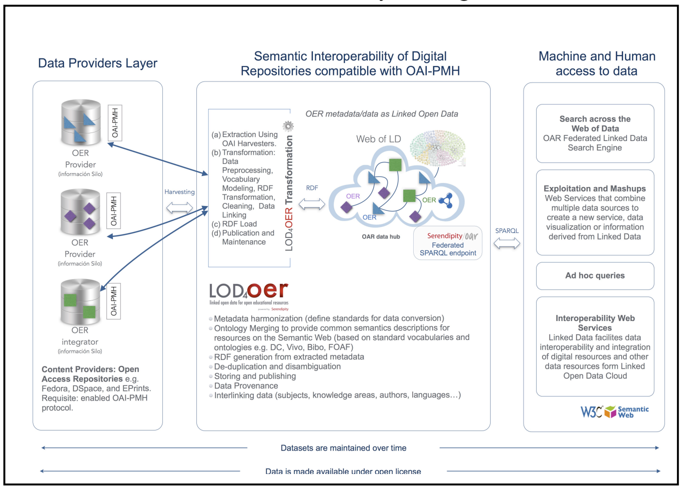
Fig. 1. General Scenarios for Interoperability between open digital repositories. From OAI-PMH to Linked Open Data for OER approach (LOD_4OER). LOD_4OER focus possesses advantages in as far as context is concerned; Current repositories are only based on OAI-PMH where the Web is used as the main infrastructure for the transportation of data and metadata. Whereas, for LD the data and its semantics are part of the same Web. These are identified by means of URIs and are described in a language that facilitates the reading and automatic processing of data by means of machine agents. Digital resources can be discovered, reused and adapted to new educational needs. Adapted from [19]

LD is mainly about publishing structured data in RDF using URIs rather than focusing on the ontological level or inference. OER provided with Linked Data (Linked Open Educational Resources Data) supports the process of discovery, reuse, integration and interoperability of open educational materials. The publication of linked data is founded upon four basic design principles by Tim Berners-Lee [4]: (1) The usage of Uniform Resource Identifiers (URIs) to identify Web resources; (2) The usage of HTTP URIs so that the users can localize and consult these resources (that enable resources to become more easily accessible); (3) The provision of useful information regarding the resource, e.g. when an URI has been consulted; and utilizing RDF to describe resources and SPARQL for query-based consultation. (4) The incorporation of links to other URIs related to the data contained in the resource so as to maximize the discovery of information on the Web. The W3C's Semantic Web provides a common framework namely Resource Description Framework (RDF) for describing resources on the Web. With RDF, automated software can store, exchange, and use machine-readable information distributed throughout the Web, in turn enabling users to deal with the information with greater efficiency and certainty; also, RDF data can be shared and reused through application, enterprise, and community boundaries.

Figure 1 shows a general roadmap for transformation of OER data and metadata as Linked Open Data. An Interoperable and open OER Information System, require to allowing a user access the data of entities that represent an OER, and retrieve linked-referential information from content/knowledge resources.

IV. PROPOSED FRAMEWORK

This section describes the life-cycle of generation of LD for



a series of selected repositories. This consists of a series of components and inter-related activities. Figure 2 shows the framework proposed for the extraction, transformation, publication and exploitation of OAI-PMH metadata as LD.

Fig. 2. General Overview of the Framework. Linked Data provides to data consumers an opportunity to merge Web-resources distributed across different datasets. Linked Data Standards are defined by W3C, which enables implementations of interoperable datasets that are adaptable, repeatable, scalable and sustainable at substantially lower cost than traditional methods. The data cycle is based on [13] and [19].

A. *Selection of data sources*

The collection of digital libraries corresponds to the member universities of the Network of Repositories belonging to the Ecuadorian Consortium for the Development of the Advanced Internet (CEDIA, http://www.cedia.org.ec). CEDIA is comprised of more than 30 higher education institutions from Ecuador, and has multiple international alliances with other networks. The material included in the selected repositories consists of: atlas, CDs, DVDs, e-books, encyclopedias, brochures, games, books, digital recordings, journals, and theses. This represents an unparalleled opportunity to take advantage of the data that are contained in them.

B. Harvesting of metadata from OAI repositories

The OAI-PMH protocol is the fundamental basis of the model for harvesting metadata related to academic resources, which is defined according to the Dublin Core schemas. The Harvester2 application was used to harvest metadata by means of OAI-PMH verbs. The extracted metadata are stored in a database of raw relational data in the form of triplets, for future conversion to Linked Open Data.

C. *Data Cleaning*

The cleaning of data is carried out with the objective of detecting and correcting corrupt or erroneous data. The process consists of analyzing inconsistent patterns in the data and in executing cleaning schemas. Among the cases that have been detected is variation in the format of certain metadata. A more concrete case was detected, however, in the language of the resources. Abbreviations such as "en", "en_US", "eng" and "English" were used to refer to the English language. In addition, variations were detected in the description of authors. The pattern was considered valid when names were followed by the sequence: <Surname, Name>.

Finally, problems of ambiguity were eliminated for the topics (dc: subject) and were added to each digital resource. A semi-automatic correction scheme could be implemented based on the cleaning of detected patterns.

D. *Vocabulary modelling*

Ontologies and open vocabularies constitute a basic scheme to describe Web resources and entities. In this study, vocabularies were examined to represent application profiles of digital repositories, digital resources, to classify topics, and to describe organizations, authors, data catalogs and repositories. The activities for modeling the ontology are the following: (a) data analysis and abstraction; (b) concept mapping, (c) searching of reusable ontological and non-ontological resources; (d) selection and reuse of controlled vocabularies; and (e) designing persistent URIs.

TABLE I. MAPPING BETWEEN EXTRACTED METADATA AND LINKED OPEN DATA VOCABULARIES (PARTIAL DESCRIPTION)

Term	***Metadata***	***Description***
	Application Profile	
Agent (class)	foaf:Agent	An entity that is associated with Catalogues and/or Datasets. If the Agent is an organization, the use of the Organization Ontology is recommended.
Catalogue (class)	dcat:Catalog	A catalogue or repository that hosts the Datasets being described.
Dataset (class)	dcat:Dataset	A conceptual entity that represents the information published.
Literal (class)	rdfs:Literal	A literal value such as a string, integer or date; Literal may be typed, e.g. as a data according to xsd:date.
Resource (class)	rdfs:resource	Anything described by RDF
dataset (property)	dcat:dataset	This property links the Catalogue with a Dataset that is part of the Catalogue.
description (property)	dct:description	This property contains a free-text account of the Catalogue. This property can be repeated for parallel language versions of the description.
publisher (property)	dct:publisher	This property refers to an entity (organization) responsible for making the Catalogue available.
title (property)	dct:title	This property contains a name given to the Catalogue. This property can be repeated for parallel language versions of the name.
license (property)	dct:license	This property refers to the license under which the Catalogue can be used or reused.
	Content	
Title	dcterms:title	Title of the resource
Topic	dcterms:subject	Phrases that describe the title or content of the resource
Description	dcterms:description	Summary or description of the content
Source	dcterms:source	Sequence of characters used to identify a work from the actual resource
Type	dcterms:type	Category of the resource
Relation	dcterms:relation	The identifier of a second resource and its relation with the current resource
Coverage	dcterms:coverage	Characteristics of the spatial or temporary coverage of the intellectual content of the resource
	Intellectual Property	
Author or creator	dcterms:creator	Person responsible for the creation of the intellectual content of the resource.
Editor	dcterms:publisher	Entity responsible for doing the resource that is found on the network in its current form.
Other collaborator	dcterms:contributor	Person or organization that has had a significant intellectual contribution, but which is secondary.
Rights	dcterms:rights	Note on copyright
	Instantiation	
Date	dcterms:date	Date in which the resource was placed at the disposition of the user

Term	Metadata	Description
Application Profile		
Format	dcterms:format	This is the data format of a resource
Identifier	dcterms:identifier	Sequence of characters to solely identify a resource, for example: URN, URL, ISBN, etc.
Language	dcterms:language	Language of the intellectual content of a resource

1) Concept Mapping: This activity is performed in order to map the main terms related to the dominium (field) of institutional digital repositories. Among those concepts to be identified are: (a) the Bibliographical Resource, i.e. bibliographical entity; (b) the Person, i.e. the creator or contributor of the bibliographical entity; (c) the corporation, i.e. the organization that created the publication of the bibliographical entity; and (d), the topics or themes, i.e. the topics related to the bibliographical entity.

With the context of this work, the metadata of a digital work were classified into three groups. Each group indicates the class or the field of information that is recorded, namely those elements related to the content of the resource, or the elements related to the resource when it is intellectual property, or the elements that are related to the instantiation of the resource (see partial description on Table I). In terms of the **thematic classification** of the bibliographical materials, the concepts were mapped via the vocabulary recommended by W3C in order to organize knowledge, or SKOS. SKOS models concepts and schemas related to concepts, lexical labeling, semantic relations, documentation, groupings of concepts, mapping properties and annotations.

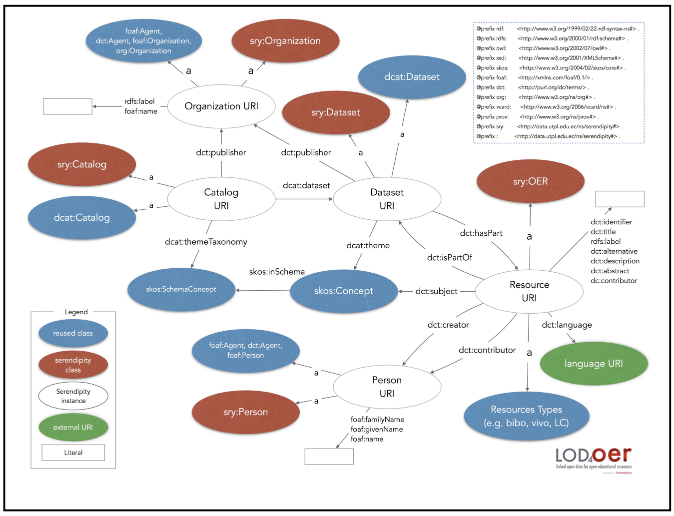


Fig. 3. Excerpt of vocabulary modeling to digital repositories

2) Searching of ontological and non-ontological resources for re-usage: The reuse of ontological and non-ontological resources reduces the time of development and the related costs during this phase, which also contributes to the quality of the ontology [17]. Given that there are various vocabularies to model bibliographical resources, we aimed to select those that facilitated the integration and interoperability of data on the Web: (a) RDF Schema (W3C recommendation http://www.w3.org/TR/rdf-schema/) and OWL (W3C recommendation: http://www.w3.org/TR/owl2-overview/) are used to describe concepts of the vocabulary. (b) Simple Knowledge Organization System (SKOS) to establish a model for the organization of knowledge, taxonomies, and other thematic hierarchies. The SKOS vocabulary was used to represent concepts or themes that were dealt with by the resource. (c) DCAT (developed under the responsibility of the Government Linked Data Working Group at W3C see http://www.w3.org/TR/vocab-dcat) was designed to facilitate inter-operability between catalogs of data published on the Web. (d) FOAF (see http://xmlns.com/foaf/spec/) was developed to represent persons and organizations, as well as their attributes and relationships with other concepts. (e) Dublin Core (DC) provides a vocabulary of "base" characteristics, which are capable of providing descriptive information about any resource. (f) Dublin Core Metadata Initiative (DCMI Metadata Terms see: http://dublincore.org/documents/2012/06/14/dcmi-terms/), is used to represent documents, as well as its attributes such as the title, creator, and relations with other entities. (g) Bibliographical Ontology Specification (BIBO, http://purl.org/ontology/bibo/) provides the main concepts and properties for describing citations and bibliographical references. (h) VIVO (see: http://vivoweb.org/ontology/core), helps describe various concepts in the academic and scientific fields. (i) Open Provenance Model Vocabulary (http://open-biomed.sourceforge.net/opmv/ns.html#Process), is used to describe the origin of data on the Web. (j) Schema.org (see http://schema.org), helps describe concepts such as materials, data properties and relations between classes.

The **application profile** specified for each data portal2 is based on the specification of the Data Catalog Vocabulary. DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogues published on the Web. Additional classes and properties from other well-known semantic vocabularies are re-used where necessary. The Application Profile is intended to facilitate data exchange between digital repositories, and therefore the classes and properties defined in this scope are only relevant for the data to be exchanged; there are no requirements for communicating systems to implement specific technical environments. The only requirement is that the systems can sharing and interoperate data in RDF in conformance with this Application Profile. During the modeling and specification, we have considered the following criteria: (a) Ensure compliance with DCAT; (b) Improve the discovery of datasets[3]. Priority to those aspects that can be used in searching and navigating; (c) Ensure simplicity: keep a narrow scope; (d) Ensure general nature of application profile; and, (e) Ensure application domain neutrality.

The above-mentioned vocabularies semantically describe related concepts. The attributes, properties and entities that are

[2] A data portal is a Web-based system that contains a data catalogue with descriptions of datasets and provides services enabling discovery and re-use of the datasets.

[3] A dataset is a collection of data, published or curated by a single source, and available for access or download in one or more formats.

not considered in these vocabularies have been designed as part of an open vocabulary (see Fig. 4)

3) Selection and reuse of Controlled vocabularies

In this work, controlled vocabularies may be introduced to improve interoperability, technical communication and knowledge management. The use of controlled vocabulary ensures that everyone is using the same word to mean the same thing. Controlled vocabularies are used to describe (concepts and relationships -terms-) a field of interest or area of concern. For instance, to declare a person in a machine-readable format, a vocabulary is needed that has the formal definition of "Organization", such as the Friend of a Friend (FOAF) vocabulary, which has a Organization class that defines typical properties of a person, or the Organization Ontology - W3C.

Controlled vocabularies provide a way to organize knowledge for subsequent retrieval. Web searching could be dramatically improved by the development of a controlled vocabulary for describing Web content; the use of such a vocabulary could culminate in a Semantic Web, in which the content is described using a machine-readable metadata scheme. They are used in subject indexing schemes, subject headings, thesauri, taxonomies and other forms of knowledge organization systems (e.g. SKOS systems). Controlled vocabulary schemes mandate the use of predefined, authorized terms/concepts that have been preselected by the designers of the schemes, in contrast to natural language vocabularies, which have no such restriction. Its use is key, though; the use of controlled vocabularies can be costly compared to free text searches because human experts or expensive automated systems are necessary to index each entry. Furthermore, the user has to be familiar with the controlled vocabulary scheme to make best use of an ecosystem. Controlled vocabularies, such as the Library of Congress Subject Headings, are an essential component of bibliography.

4) Strategy for identifying resources by means of persistent URIs

Two types of URIs were designed for this: one to identify the components of the vocabulary (classes, properties and relations) and another one to describe resources (digital repositories and bibliographical material). To describe the materials, we used HTTP URIs – thereby taking into account the principles for the publication of data proposed by Tim Berners-Lee, which was done according to the following patterns:

Prefix: oar-utpl
URI base: http://data.utpl.edu.ec/serendipity/oar/
Schema: http://data.utpl.edu.ec/serendipity/oar/schema#
Resources: http://data.utpl.edu.ec/serendipity/oar/resource/
Properties: http://data.utpl.edu.ec/serendipity/oar/property/
Categories: http://data.utpl.edu.ec/serendipity/oar/category/
Graph: http://data.utpl.edu.ec/serendipity/oar
SPARQL endpoint: http://data.utpl.edu.ec/serendipity/oar/sparql

TABLE II. MAPPING BETWEEN TYPE OF OER AND RELATED TERMS

Type of resource	***URI of equivalent concept***
Article	schema:Article; bibo:AcademicArticle
Book	schema:Book bibo:Book
Other	Bibo:Document
Presentation	bibo:Slideshow vivo: Presentation
Technical Report	bibo:Report bibtex:Techreport
Thesis	bibo:ThesisDegree bibo:Thesis

E. Generation and Publication of Linked Data

In order to create RDF from the harvested data, an individual generator was developed based on Jena. An important stage in the process of generating RDF data was the assignation of URIs to the extracted texts. In this sense, the metadata of the bibliographical sources were mapped with URIs from the most appropriate terms. This enabled the resources to be inter-operable and to be integrated with other data series. Table II shows an example of text mapping for the text corresponding to the type of bibliographical material and its corresponding URIs. Once the RDF data have been created and stored, we can select web interfaces or applications to display the data in a legible format. There are some tools, however, that can be connected to the RDF repository and can present recovered data such as web pages, or as graph schemas. One of the most popular tools to display RDF data in tabular form is Pubby, (available at: http://wifo5-03.informatik.uni-mannheim.de/pubby/) which is a java application that is used by recognized repositories of triplets. A number of SPARQL were likewise implemented as part of the validation process.

Query: Topics related to "Artificial Intelligence". Each digital resource is associated with one or various topics. The association between a resource and a topic is established by means of the metadata *dc:subject*. From any topic of interest (e.g. "Artificial Intelligence"), other concepts can be retrieved in order to create tag clouds or taxonomies of related topics. The facility to find and exploit relationships between concepts enables the creation of different tools support oriented to end-user. The Fig. 4 shows a sub-graph of concepts related to a particular discipline of interest. The interactive view has been built with results generated by Query 1.

```
PREFIX dct: <http://purl.org/dc/terms/>
SELECT DISTINCT ?relatedSubject
count(?relatedSubject) AS ?frequency
WHERE {
   ?resource a dct:BibliographicResource.
   ?resource dct:subject ?resSubject ; dct:subject
?relSubject.
   OPTIONAL{
      ?resSubject rdfs:label ?subject .
      ?relSubject rdfs:label ?relatedSubject .
   }
   FILTER REGEX(?subject, "Artificial
Intelligence", "i")
} GROUP BY ?relatedSubject
ORDER BY DESC(?frequency)
LIMIT 10
```

F. Enrichment and reconciliation of data

In order to improve the discovery of resources, it was necessary to create RDF linkages with external resources. These were published in the cloud for open linked data and were connected with external repositories. The topics and key words were linked with each digital resource, which consisted of controlled vocabulary content and classification schemas. More specifically, external links were established with UNESCO nomenclature and English/Spanish/Latin-America DBpedia datasets.

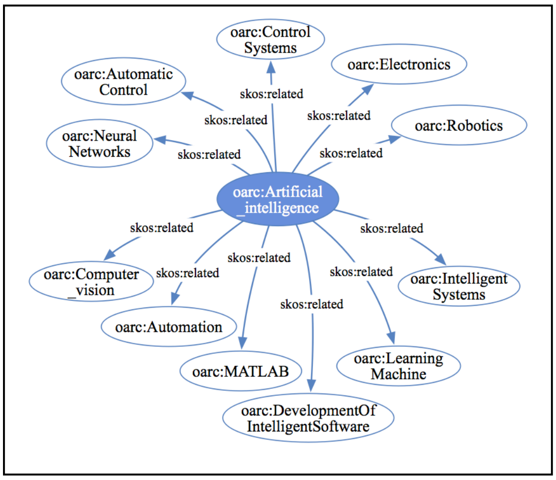


Fig. 4. Concepts related to "Artificial Intelligence" discipline

V. Guidelines for LD Publication

The goal of this section is to compile the most relevant data management best practices for the publication and use of high quality data published by OER providers around the world as Linked Open Data.

A. Understand the business requierements

The firs step in an interoperability project is to understand the goal: what are the requirements/priorities/expectative trying to achieve? This should be documented both at a high level in terms of desired strategic objectives and priorities, and at a detailed level in the form of business requirements. The identification of key and common information sharing and integration requirements can be used as the basis for identifying and engaging with their local stakeholders on what interoperability is needed to meet strategic drivers without diving straight into very detailed requirements discussion.

Key considerations: (a) Create stimulus for change. It's clear that the OER community, educational policymakers, OER developers and most educators want large improvements in access and quality in education. (b) Manage roadmaps and expectations – understand what each OAR provider wants to achieve, prioritize as a repository network and be interoperable on what will be delivered when. (c) Ensure you engage local sponsor to establish a good layout when presenting OER information. (d) Bear in mind that evolving understanding of business requirements is inevitable – plan for this by adopting a linked data agile generation and publication approach whenever possible.

B. Life Cycle of Linked Data generation and Publication

This section sets out a series of best practices designed to facilitate development and delivery of OER and Open Access Repositories as Linked Open Data. Using LD Principles, users can describe and query Linked Data from multiple sources at once and combine it without the need for a single common schema that all data shares. Prior to international data exchange standards for data on the Web, it was time consuming and difficult to build applications using traditional data management techniques. As more open linked data is published on the Web, best practices are evolving too.

Key considerations:

Selection of data sources refers to digital libraries that are of interest to a pre-determined project. The Federation of Digital Libraries defines libraries as organizations that provide resources consisting of specialized personnel that are used to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and guarantee persistence over a period of time the collections of digital work in such a way that they are of easy access and economically available for their usage by a community or by a series of communities [5]. Harvesting of metadata from repositories.

Modeling of vocabulary or ontologies. In this stage, mapping relations are established with other vocabularies. The usage of ontological and non-ontological resources is essential to increasing the level of inter-operability within the Linked Open Data context. **Selection and reuse of controlled vocabularies**. Requirements: (i) Be published under an open license; (ii) Be operated and/or maintained by a recognized standards organization or another trusted organization; (iii) Be properly documented (labels, descriptions, notes, examples, etc.); (iv) Have labels in multiple language; (v) Contain a relatively small number of concepts/terms that are general enough to enable a wide range of resources to be classified; (vi) Have terms that are identified by URIs with each URI resolving to documentation about the term; (vii) Have associated provenance, data license, persistence and versioning policies.

Design of "Good URIs". The Web makes use of the URI as a single global identification system, which promotes large-scale "network effects". Therefore, in order to benefit from the value of LD, publishers need to identify their resources using URIs. This implies: (a) Provide HTTP URIs as identifiers for their informational resources. There are many benefits to participating in the existing global network of URIs, including linking, caching, and indexing by semantic agents. (b) Use HTTP URIs to enable people to "look-up" or "dereference" a URI in order to access a representation of the resource identified by that URI. (c) Provide at least one machine-readable representation of the resource identified by the HTTP URI (e.g. a human-readable HTML representation or a machine-readable RDF/XML or JSON-LD). (d) URI Persistence: URIs should be stable and reliable in order to maximize the possibilities of reuse and persistence that Linked Data brings to users. (e) Make URIs readable and keeping them more stable by removing descriptive information that will likely change. (f) URI Opacity: URIs should be constructed to ensure ease of use during development and proper consideration. Web clients accessing such URIs should not parse or otherwise read into the meaning of URIs. (g) Defining and documenting a persistent URI strategy and implementation plan. Persistent URIs are used to retain addresses to information resources over the long term.

Conversion of data to RDF formats: Refers to: (a) Converted data extracted from standard, open and interoperable formats can be done in such a way that it facilitates access and re-usage, and resolves the problem of digital resources restricted to storage devices. (b) Cleaning of generated data. These are activities that aim to reduce ambiguity and to purge the information that is extracted and generated during the process of conversion.

Linkage of data by means of semantic relations (countries, bodies, concepts) with existing sources. In this way, we can establish linkages between a series of open data and contribute to its integration on a global scale and create the network effect. **Publication and exploitation of data**. The publication of data on the Web by means of standard technologies proposed by W3C improves accessibility, availability and integration of these resources to others.

Concerning license vocabularies, implementers are encouraged to use widely recognized licenses such as Creative Commons licenses, (see http://creativecommons.org/licenses/), and in particular the CC Zero Public Domain Dedication (CC0 1.0http://creativecommons.org/publicdomain/zero/1.0/), the Open Data Commons Public Domain Dedication and License (PDDL, http://opendatacommons.org/licenses/pddl/), the ISA Open Metadata License, (ISA Open Metadata License v1.1, http://joinup.ec.europa.eu/category/licence/isa-open-metadata-licence-v11), the European Union Public License (EUPL, http://joinup.ec.europa.eu/software/page/eupl) or an open government license such as the UK Open Government License (see http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/), the Open Data Rights Statement Vocabulary (See http://schema.theodi.org/odrs/) and Open Digital Rights Language (ODRL) Initiative (See http://www.w3.org/community/odrl/).

C. Celebrate small successes.

Goals and priorities setting, and success in achieving goals are arguably the root of the development of organizations and individuals involved. The actions are designed to help achieve the goal more efficiently and effectively. However, sometimes the organizations set fairly lofty goals with respect to interoperability and integration, and while we are taking action towards those goals it seems as though very little progress is being made. The interoperability and integration still has a long way to go, but it is certainly the way to reduce costs and improve productivity

Key considerations: (a) The initiatives aiming to increase interoperability, should follow a "breakthrough strategy," which consists of locating and starting at once with the benefit that can be achieved quickly and then using these first successes as stepping stones to increasingly ambitious integration and interoperability advantages. It is key, decries overemphasis on strategic planning. Instead, the initiative must focus on accomplishing an approach based on short-term result, a success. (b) It is important to select a Linked Data life cycle with an incremental and flexible approach. This must be designed to help achieve the strategic goals more efficiently and effectively; as well as support a strategy to achieve small and continued successes. (c) Change the rules for success along the path to accomplishing the goal settings. Make it easy for team to feel successful, and celebrate even the smallest success. (d) In general, the people always focused on what you haven't accomplished. Success is largely about having a certain state of mind, a success mind set. One success produces to another and the snowball effect goes to work. (e) By celebrating smaller successes, the teams are actually starting to communicate this attitude towards themselves and their partners. Success is a feeling to an extent, and by celebrating small successes the organization and your team make it easy for all to feel successful.

VI. Conclusions and Future Work

Generated data act as a knowledge base for relationships between digital resources, authors, organizations, topics, geo data and other entities. Each URI could be used in search and discovery OER systems for navigation, analytics and visualization of content. The management of vast collections of digital resources should aim to guarantee the successful localization of the most adequate contents for users. To achieve this, two services were presented that took advantage of the potential of linked data and Web semantic technologies, i.e. with the objective of improving the recovery of Web resources.

Within the context of open access, providers of open digital should aim to adopt a model that improves the integration of repositories. This should be done in such a way that it supports standards of both formal and open metadata for the description of resources, i.e. whose level of specificity, granularity and complexity are achievable. The model that is to be adopted should thus respect and guarantee conditions of local autonomy within a framework of global technological inter-operability. Placing a focus on Linked Data therefore maximizes the potential of inter-operability and the integration within a context such as the Web. In other words, there is a high heterogeneity within a distributed context. The paper describes the approaches used, and it outlines the lessons learnt. This lessons include: specify an appropriate open data license; the use of HTTP URIs to as naming strategy of objects; data persistence; data provenance; multilingual support; reuse standard vocabularies whenever possible; provide machine access toward OER Linked Open Data; using Web open standards; and linking OER data with other open datasets.

It is of particular interest to the authors of this paper to contribute to the improvement of both the visibility and the access to the digital materials that are used or generated in the academic and scientific field. For this reason, we will seek to continue designing services that explore the potential of linked data. More specifically, we are exploring other networks of federated repositories from Latin America with the aim of forming a central hub that provides integrated information based on the work and fields of research from each institution from the region. In addition, a semantic searcher based on facets is also being developed to facilitate the exploration of content by means of categories, namely those that are associated with resources. This type of navigation helps to identify different attributes of the documents. Moreover, in this way, Internet users may easily find relevant resources.

ACKNOWLEDGMENT

The work has been funded by the *Universidad Técnica Particular de Loja* (UTPL), and partially by *Consorcio Ecuatoriano para el Desarrollo de Internet Avanzado* - CEDIA and Regional Government of Madrid (eMadrid S2013/ICE-2715). The Scholarship partially funded by *Secretaría Nacional de Educación Superior, Ciencia y Tecnología e Innovación* of Ecuador (SENESCYT).

REFERENCES

[1] Anibaldi, S., Jaques, Y., Celli, F., Stellato, A., Keizer, J. (2015) Migrating bibliographic datasets to the Semantic Web: The AGRIS case. Semantic Web, 6(2), 113-120.

[2] Baker, T., Bermés, E., Coyle, k., Dunsire, G., Isaac, A., Murray, P., y otros (2011). Library Linked Data Incubator Group Final Report. Obtenido de http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/#Benefits_of_the_Linked_Data_Approach.

[3] Berners-Lee, Tim; Hendler, J.; Lassila, Ora (2001). The Semantic Web, Scientific American, 284(5), 34-43.

[4] Berners-Lee, Tim (2006). Linked Data-Design Issues. Obtenido de: http://www.w3.org/DesignIssues/LinkedData.html

[5] Chinwe, V. A., & Majesty, I. E. (2011). Digital library deployment in a university. Library Hi Tech, 29(2), 373-386. doi:http://dx.doi.org/10.1108/07378831111138233

[6] Gendt, M.; Isaac, A.; Meij, L. & Schlobach, S. (2006). Semantic Web Techniques for Multiple Views on Heterogeneous Collections: A Case Study. Research and Advanced Technology for Digital Libraries. Springer Berling Heidelberg, vol. 4172.

[7] Francesconi, E.; Faro, S. ; Marinai, E. & Perugi, G. (1008). A Methodological Framework for Thesaurus Semantic Interoperability. Proceeding of the Fifth European Semantic Web Conference, 76-87.

[8] Heath, Tom & Bizer, Christian (2011). Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, 1(1), 1-136. Morgan & Claypool.

[9] Rowlands, Ian and Bawden, David (1999), "Digital Libraries: a conceptual framework," Libri Journal, vol. 49, pp. 192-202.

[10] IEEE (1990). IEEE Standard Glossary of Software Engineering Terminology. IEEE Standards Board, New York.

[11] Lenzerini, M. (2002). Data Integration: A Theoretical Perspective. In Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, Madison, Wisconsin, USA. doi: 10.1145/543613.543644

[12] Malakhov, D., Serebryakov, V., Teymurazov, K., Shorin, O. (2014) Semantic integration of bibliographic records. CEUR Workshop.

[13] Piedra, N. Tovar, E. López, J. Chicaiza, J. (2014). Consuming and producing linked Open Data: The case of Opencourseware. Emerald EarlyCite. DOI: 10.1108/PROG-07-2012-0045

[14] OAI (2015). The Open Arvhives Initiative Protocolo for Metadata Harvesting, Implementation GuideLines. Obtenido de: http://www.openarchives.org/OAI/openarchivesprotocol.html.

[15] Sztyler, T., Huber, J., Noessner, J., Murdock, J., Allen, C., Niepert, M.(2014). LODE: Linking digital humanities content to the web of data. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries.

[16] Tovar, E. & Piedra, N. (2014). Open Educational Resources in Engineering Education: a perspective to improve Reusability of Resources and Data. IEEE Transactions on Education, 57(4). DOI: http://10.1109/TE.2014.2359257

[17] Villazón-Terrazas, B. (2011). A Method for Reusing and Re-engineering Non-ontological Resources for Building Ontologies.

[18] Zengenene, D., Casarosa, V., Meghini, C. (2014). Towards a Methodology for Publishing Library Linked Data. Communications in Computer and Information Science, 385 CCIS, 81-92.

[19] Piedra, N. Chicaiza, J. Quichimbo, P. Saquicela, V. Cadme, E. López, J. Espinoza, M. Tovar, E. (2015) Marco de Trabajo para la Integración de Recursos Digitales Basado en un Enfoque de Web Semántica. RISTI. doi: 10.17013/risti.e3.55-70. 2015.