Revista Cubana de Ciencias Informáticas
Vol. 13, No. 1, Enero-Marzo, 2019
ISSN: 2227-1899 | RNPS: 2301
Pág. 14-30
http://rcci.uci.cu

Tipo de artículo: Artículo original
Temática: Inteligencia artificial
Recibido: 13/07/2018 | Aceptado: 23/11/2018

A Linked Data-based Semantic Interoperability Framework for Digital Libraries

Un marco de interoperabilidad semántica basado en datos enlazados para bibliotecas digitales

Yusniel Hidalgo-Delgado[1*], Bin Xu[2], Alejandro Jesús Mariño-Molerio[1], Juan Pedro Febles-Rodríguez[3], Amed Abel Leiva-Mederos[4]

[1]Department of Programming Techniques, University of Informatics Sciences, Cuba
[2]Department of Computer Science and Technology, Tsinghua University, China
[3]Department of Postgraduate Education, University of Informatics Sciences, Cuba
[4]Department of Information Sciences, Central University of Las Villas, Cuba

[*]Autor para correspondencia: yhdelgado@uci.cu

Abstract
The growing popularity of the adoption of the linked data is increasing the semantic interoperability in the digital libraries era. The linked data principles provide an efficient way to interlink resources across diverse datasets. Several digital libraries around the world are publishing their legacy data from catalogs and authority files following the linked data principles. In this paper, we propose a linked data-based semantic interoperability framework for digital libraries. The proposed framework is based on three layers, supporting the data acquisition, linked data publication process and the building of value-added services for the digital libraries users. In order to evaluate the feasibility of the framework proposed, we have built a prototype as a proof of concept. The prototype demonstrates the effective implementation of wrappers as a data integration method to deals with the heterogeneity of the diverse data sources. Moreover, illustrates the importance of dealing with the quality of the bibliographic metadata form early stages of development.

Keywords: Digital libraries, linked data, ontologies, semantic web, semantic interoperability

Resumen
La creciente popularidad de la adopción de los datos enlazados está aumentando la interoperabilidad semántica en la era de las bibliotecas digitales. Los principios de los datos enlazados proporcionan una manera eficiente de interconectar recursos a través de diversos conjuntos de datos. Varias bibliotecas digitales en todo el mundo están publicando sus datos heredados de catálogos y archivos de autoridad siguiendo los principios de los datos enlazados. En este artículo, se propone un marco de interoperabilidad semántica basado en los datos enlazados para bibliotecas digitales. El marco propuesto se basa en tres capas, que soportan la adquisición de datos, el proceso de publicación de datos enlazados y la creación de servicios de valor agregado para los usuarios de las bibliotecas digitales. Para evaluar la viabilidad del marco propuesto, hemos creado un prototipo como prueba de concepto. El prototipo demuestra la implementación efectiva de envoltorios como un método de integración de datos para abordar la heterogeneidad de las diversas fuentes de datos. Además, ilustra la importancia de tratar con la calidad de los metadatos bibliográficos desde las primeras etapas del desarrollo.

Revista Cubana de Ciencias Informáticas
Vol. 13, No. 1, Enero-Marzo, 2019
ISSN: 2227-1899 | RNPS: 2301
Pág. 14-30
http://rcci.uci.cu

Palabras claves: *Bibliotecas digitales, datos enlazados, interoperabilidad semántica, ontologías, web semántica*

Introduction

The goal of the libraries is to manage the scientific knowledge and cultural heritage of the society. The libraries are categorized into three broad categories: Conventional Libraries, Digital Libraries and Hybrid Libraries Tanuj Singh and Alka Sharma (2015). Conventional libraries manage printed collections whereas digital libraries collect, store and disseminate the information in digital or electronic form. The combination of both conventional and digital libraries are hybrid libraries, where information is stored and communicated both in print and digital formats. According to IFLA and UNESCO (2011), a digital library is an online collection of digital objects, of assured quality, that are created or collected and managed according to internationally accepted principles for collection development and made accessible in a coherent and sustainable manner, supported by services necessary to allow users to retrieve and exploit the resources.

The interoperability and sustainability are keys to the vision of digital libraries able to communicate with each other. Digital libraries that conform to commonly agreed open standards and protocols improve world-wide knowledge dissemination and access IFLA and UNESCO (2011). The interoperability is the ability of a system or a product to work with other systems or products without special effort on the part of the customer. Interoperability is made possible by the implementation of standards Institute of Electrical and Electronics Engineers (1990). There are three types of interoperability: technical, syntactic and semantic. According to Commission (2010) the syntactic interoperability is about describing the exact format of the information to be exchanged in terms of grammar, format and schemes, while the semantic interoperability is about the meaning of data elements and the relationship between them. It includes developing vocabularies to describe data exchanges and ensures that the data elements are understood in the same way by communicating parties.

There are several international projects for developing digital libraries with big document collections. Europeana Haslhofer and Isaac (2011) integrates digital resources from multiple countries of the European Union. In addition, the National Library of Spain Vila-Suero et al. (2013) is one of the most biggest digital libraries of Europe and manages both, printed and digital collections. In North America, the Library of Congress of the United States store and manages documents and multimedia collections produced inside and outside of the United States. These digital libraries use semantic web technologies to increase the semantic interoperability of bibliographic metadata published.

This paper introduces a novel linked data-based semantic interoperability framework for digital libraries.

The proposed framework is based on three layers, supporting the data acquisition, linked data publication process and the building of value-added services for the digital libraries users. It includes a wrapper-based data integration method for integrating bibliographic data from heterogeneous and distributed data sources. Moreover, the framework provides a standardized linked data publishing process using tools and standards adopted by the linked data community.

This paper is structured as follows: Section Related Work presents some related work about the adoption of the linked data principles in digital libraries. Section Problem Statement presents a comprehensive understanding of the research questions to be addressed in the investigation. Section Proposed approach presents an overview of the semantic interoperability framework proposed. Section Prototyping describes the development process of a prototype. Finally, in Section Conclusions and future work we present our final remarks and future research lines in the near future.

Related Work

The adoption of the linked data principles by the digital library community is growing up in recent years, according to the number of research projects and published papers dealing with basic and applied problems regard to this topic. In Hallo et al. (2016) have been studied the current state of linked data in digital libraries. The study focuses on selected vocabularies and ontologies, benefits and problems encountered in implementing Linked Data in digital libraries. Also, the study provides a set of challenges, such as the need to develop tools for the Linked Data transformation and the quality control of the datasets published, among others.

Recent efforts are addressing the publishing process of linked data from legacy data sources, such as authority files and bibliographic catalogs in MARC and MARC21 Vila-Suero et al. (2013); Hallo et al. (2014); Crowe and Clair (2015); Chen (2017); Candela et al. (2018). These approaches are very useful for the library community, increasing the semantic interoperability between the digital library systems and creating useful links across related resources on the web.

Publishing bibliographic data following the linked data principles increase the semantic interoperability in the digital library ecosystem. In Binding and Tudhope (2016), the authors discuss various aspects of vocabulary mapping using linked data for improving the semantic interoperability across several digital libraries systems. Also, the potential use of such vocabulary mappings to assist cross-search over archaeological datasets from diverse countries was illustrated in a pilot experiment. In Nisheva-Pavlova et al. (2015); Binding and Tudhope (2016) an ontology-based semantic digital library is proposed. The authors introduce DjDL, a semantic digital library with Bulgarian folk songs. DjDl implements a search engine, providing access to a variety of resources. It provides two main types of search: keywords-based and semantic (ontology-based) search.

Another issue addressed is the quality of library linked data published Radulovic et al. (2018). In Talleras (2017) the author examines the quality of bibliographic linked data published by four national libraries from Europe. The examination was mainly based on a statistical study of the vocabulary usage and interlinking practices in the published datasets. The study finds that the national libraries successfully adapt established linked data principles, but issues at the data level can limit the fitness of use. Similar results were obtained in a systematic literature survey published in Zaveri et al. (2016). In addition, this survey provides a set of metrics to measure several quality dimensions of the linked data published.

In general, we have identified a set of common issues in the publication process of the library data as linked data, such as:

- Lack of approaches for publishing library linked data from scratch, most of the authors transform legacy metadata described in MARC and MARC21.
- Lack of effective tools for improving the quality of metadata at the data level, before the RDF-ization process. Most of the existing approaches are focused on the evaluation of the quality of the published library linked data.
- Lack of the implementations of value-added services based on linked data consumption for digital libraries users.

Problem Statement

In digital libraries the semantic interoperability represent a key factor for improving the access to library data from both, human and computers. However, in real-world sceneries, data and metadata are stored on heterogeneous and distributed data sources. Therefore, we formulate the following general problem:

How to increase the semantic interoperability of library data from heterogeneous and distributed data sources?

Based on the general problem statement, three subproblems are identified. They will be discussed in the following sections, thus, leading to the definition of corresponding research questions.

Integrating Heterogeneous and Distributed Data Sources

The first challenge, which can be deduced from the overarching problem statement is the integration of heterogeneous data sources. In this context, library data mean metadata fields and their digital objects obtained from conference proceedings, academic journals, digital repositories and even the web. Therefore, we formu-

late: Which existing Data Integration methods can be used to model and accordingly process library data to overcome heterogeneity and allowing integration, reusability and discoverability? **(RQ1)**.

Data sources are distributed and heterogeneous. For instance, data and metadata from online journals and digital repositories could be accessed over the web although Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) in a distributed way. However, the conference proceedings distributed on Compact Disk could be processed using metadata extraction tools. In both cases, we need to use data integration methods to overcome issues related to heterogeneity and distribution of the data sources.

Publishing Library Linked Data

The goal of library linked data is to help increase global interoperability of library data on the Web. There are several approaches for publishing library linked data in the literature, however, there is no consensus in the academic community and practitioners about what methodological guidelines should be used for publishing library linked data. The linked data publishing process is highly dependent from nature of data sources, the expertise of the data publishers and the maturity of the tools existing for doing this complex process. Therefore, we formulate: How to publish library data following the linked data principles by integrating standards, tools and emerging technologies? **(RQ2)**.

Semantic Digital Libraries Services

From the point of view of the developers of the semantic digital libraries, the publishing process of library data as linked data involves the use of standards, such as Resource Description Framework (RDF) and SPARQL Protocol and RDF Query Language (SPARQL). However, the non-experts users do not know about these standards and technologies. In this sense, we need to provide to users with value-added services to improve the access and usage of the library linked data published. Therefore, we formulate: How to implement semantic digital libraries services on top of library linked data published for improving the access by non-expert users and computers? **(RQ3)**.

Semantic Digital Libraries Services should not be enabled only for human consumption, the data silos should be useful for others digital libraries systems making possible the cooperative work between them. Also, data and metadata used by others systems should be automatically understood by using ontologies as knowledge representation form.

By combining the three challenges inferred from the general problem statement, this work will allow both, humans and computers, to access and exchange library data from heterogeneous and distributed data sources.

The aim of the work is to develop a linked data-based semantic interoperability framework for library data from heterogeneous and distributed data sources.

Proposed approach

To address the semantic interoperability issues in the digital libraries, we propose a linked data-based semantic interoperability framework. This framework consist of three layers. The first layer is the data acquisition layer, the second one is the linked data layer and the last one is the semantic digital library services layer. In figure 1, we show the layered-based framework from the bottom-up perspective.

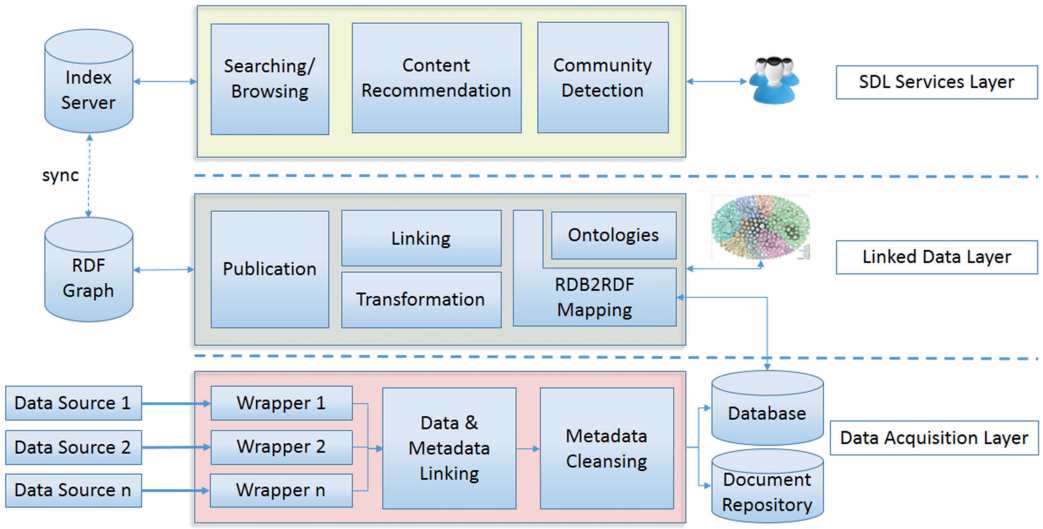

Figure 1. Semantic interoperability framework for digital libraries.

Data acquisition layer

The goal of this layer is to extract and store library data from heterogeneous data sources. At this stage, it is necessary to analyze several issues from each data source, such as data interoperability, quality of data, data schemas, update frequency and sustainability over time. The input of this layer is one or more data sources and the output is an intermediate database with the extracted bibliographic metadata and the repository of documents. This layer has three main components (1) wrapper component (2) data and metadata linking, and (3) data cleansing component.

In digital libraries domain there are data sources, such as open access journals, conference proceedings and even the web. These data sources are characterized by their heterogeneous and distributed manner. Wrappers

are the components of a data integration system that communicate with the data sources AnHai Doan et al. (2012). Wrappers components are needed to achieve the integration and interoperability issues across these systems.

Given a data source S, a wrapper W extracts structured data from S. Formally, W is a tuple (Tw, Ew), where Tw is a target schema, and Ew is an extraction program that uses the format Fs to extract a data instance from each data source S.

After executed each wrapper, the digital objects (documents) and bibliographic metadata are stored in the documents repository and a database respectively. At this stage, there are no links between digital objects and their corresponding bibliographic metadata. For that, the data and metadata linking component in our framework solve these issues. Given a set of documents D and a set of bibliographic metadata records M, we have to calculate the similarity S between the document Di with each bibliographic metadata record Mi. If the similarity S between Di and Mi is higher than a threshold T, we establish the link between them.

The quality of library data remains a crucial point that significantly affects the visibility and discovery of resources described in a Linked Data context. To address the quality issues, we propose a data cleansing component. The goal of this component is to clean up and normalize some fields of metadata, improving considerably their quality. This component includes data transformation for normalizing some metadata fields, such as dates, volumes, and numbers of journals. In these cases, a regular expression based approach could be enough.

In this component, a common issue is the Entity Resolution (ER). It refers to the issue of identifying and linking or grouping different manifestations of the same real-world object Getoor and Machanavajjhala (2012); Gal (2014). In the particular case of the library data, we identify two entities that could be identified and grouped, for example, the author's names and their affiliations. The ambiguity in the representations of authors names is a common issue in data sources, such as papers published in open access journals and conference proceedings distributed on compact disk. These issues arise due to lack of authority control systems. A similar issue arises with the affiliations of the authors.

The output of this layer is a database and the document repository. The database contains the bibliographic records about published papers in journals and conference proceedings. The document repository store the digital object collection (most of them in PDF format). At this point, should exists the links between digital objects and their corresponding bibliographic record.

Table 1. Alignment between the proposed framework and methodological guidelines proposed in Hidalgo-Delgado et al. (2017a).

Framework layers	**Methodological guidelines**	**Tasks**
Data acquisition layer	Data extraction	Wrapper development Data and metadata linking
	Data preprocessing	Metadata cleansing Data disambiguation Duplicate records detection
Linked data layer	Data modeling	Ontology development and reuse RDB2RDF mapping
	Data publishing	Transformation Linking Publication
Semantic digital library services layer	Data exploitation	Searching/Browsing Content recommendation Community detection

Linked data layer

The goal of this layer is to produce Library Linked Data from the bibliographic records stored in the previous layer. The output is one or several RDF graph. Producing library linked data involves making complex technical and methodological decisions. In the last years, several methodological guidelines and tools with specific goals have been developed by Library Linked Data practitioners.

At this stage, we propose the use of a set of methodological guidelines described in Hidalgo-Delgado et al. (2017a). It consists of five activities: (1) data extraction, (2) data preprocessing, (3) data modeling, (4) data publishing, and (5) data exploitation and follows an iterative and incremental approach. The Table 1 shown the alignment between the layers of the proposed framework, their corresponding activities in the methodological guidelines and the implemented tasks in the prototype (see Section Prototyping).

The RDB2RDF mapping component involves the use of a mapping language between a relational database and the RDF data model. We suggest the use of R2RML[1], a W3C standard for this task. Also, we have to reuse or in some cases, to develop new domain ontologies for modeling the source data, in this case, bibliographic metadata records.

The transformation, linking and publication components are very closely related to each other. Firstly, the bibliographic metadata is transformed into RDF triples through an RDF-ization process. Then, RDF triples

[1] https://www.w3.org/TR/r2rml/

previously generated are linked to similar triples existing in others RDF graph published on the Linked Open Data Cloud and finally, the resulting RDF graph is made accessible over the web space. At this point, we have stored the RDF graph into a triplestore, a special type of graph-based database highly optimized for storing RDF triples. Most of the available triplestores today provides an SPARQL Endpoint for querying and using the semantic data.

Semantic digital library services layer

The goal of this layer is to provide a set of value-added services to both, humans and computers. The input of this layer is an SPARQL endpoint and an indexing server. The SPARQL endpoint provides access to RDF triples stored in the triplestore defined in the linked data layer. Also, the indexing server provides a technical solution for improving the queries response time over the triplestore in real environments.

In this layer, we have to develop three main services. The first one is a web-based information retrieval system using textual and faceted search. This information retrieval system provides searching and browsing components over the linked data published in the indexing server, according to the queries formulated by the final users. In a previous research works, we stated that browsing and searching over an SPARQL Endpoint in production environments are time-consuming due to the queries complexity in large RDF graph. In this sense, the indexing server gets better queries response time instead of the native RDF triplestore.

The second one is a content recommendation component. This component allows generating a ranked list of items to recommend to the users according to recommendation criteria. In the linked data context, there are several content recommendations approaches Chicaiza et al. (2017); Hu and Medapati (2017); Vagliano et al. (2017). In this case, the content recommendation systems aim to provide additional items related to any bibliographic record on the RDF graph, such as authors working on the same scientific discipline or similar research papers.

The last one is the community detection component. This component allows identifying communities or partitions of nodes that share common properties in a network. The co-authorship networks are considered complex networks, where the nodes of the network are the authors and the edges between nodes provide the co-authorship relationships in one or more publications. In this component, we provide a value-added service to final users with the aim of detect communities of authors working on the same scientific discipline using the co-authorship network from the RDF graph.

The three-layer framework is generic and extensible, by adding new inputs to the data acquisition layer and the semantic digital library services layer. It means that we can develop any additional wrapper for any other

data source, and we can develop new services on the top layer using the RDF graph generated.

Also, the proposed framework has two others important features. The framework follows an iterative and incremental approach. It means that any can develop new semantic digital libraries by developing new components in successive iterations and by increasing the functionalities implemented in previous iterations. Finally, the proposed framework follows a pipeline approach, it means that the output of the previous component is the input of the next component.

Prototyping

In this section, we describe the prototype implementation process as proof of concept. The aim of the prototype is to evaluate the feasibility of the proposed framework in a real environment. The prototype instantiates the proposed framework by developing new software tools and reusing other ones. In the figure 2, we show the general architecture of the prototype implemented. The green line components were implemented by the authors and the red line components were developed by third-parties and reused in the prototype.

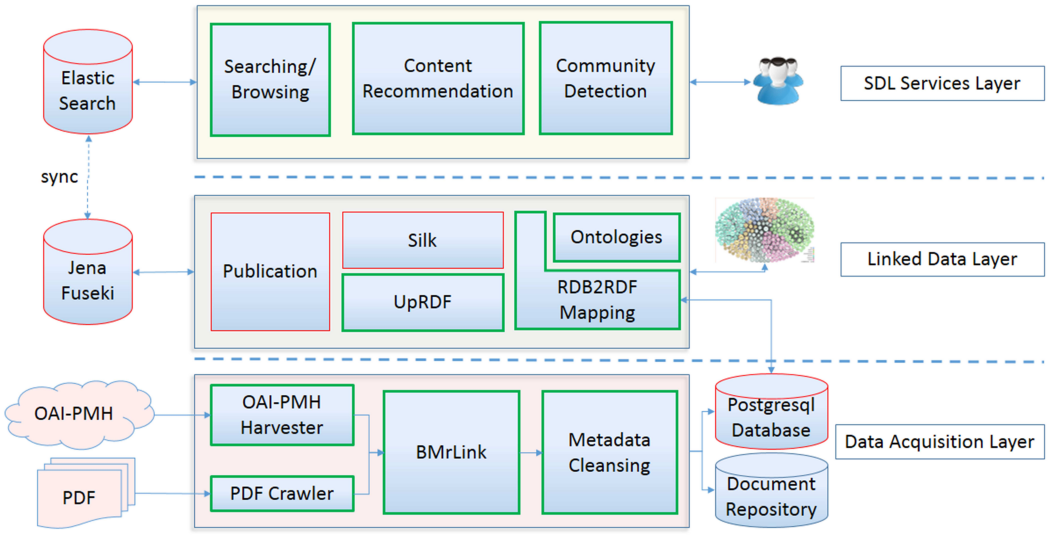

Figure 2. Instantiation of the semantic interoperability framework.

Data acquisition layer

In this layer were implemented four main components: an OAI-PMH Harvester (wrapper), a PDF crawler (wrapper), the BMrLink tool and the metadata cleansing component. The OAI-PMH harvester wrapper is

to harvest bibliographic metadata from OAI-PMH data providers, such as Open Access Journals. It supports the metadata harvesting from multiple OAI-PMH data providers using the direct connection or through a proxy server. Also, the tool support metadata extraction about records, authors, journals, collections, and organizations (see Figure 3). The PDF crawler supports the PDF extraction from open access journals or personal websites. The aim of this focused crawler is to retrieve the research papers from academic journals in PDF format. The PDF collection is stored in a remote server or a local directory.

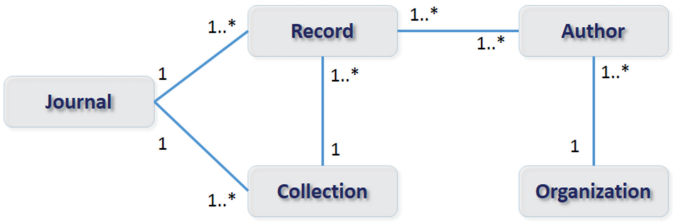

Figure 3. The entity relationship model of the main entities modeled from data sources.

After extracting from the same journal both, metadata records using the OAI-PMH harvester and the research papers in PDF format, it is necessary to create links between them. For this purpose, we implement the BMrLink tool Hidalgo-Delgado et al. (2017b). This tool includes a linking algorithm based on partitioning and comparing the titles extracted from the PDF files and the metadata records. The PDF titles are extracted reusing the CERMINE tool Tkaczyk et al. (2015). Finally, we implement a component for metadata cleansing. This component overcomes some issues related to the metadata records, such as author name disambiguation, institution disambiguation, and duplicate records detection. In the case of author name disambiguation and institution disambiguation, we design and implement an algorithm based on the edit distance to compute the similarity values of author names, authorship, affiliation and publication place. Next, these similarity values are used to build a vector for each author in the collection of documents from the bibliographic database. This vector is used to build clusters using k-means algorithm.

The output of this layer is the bibliographic metadata records stored in a relational database and the PDF files collection in the document repository. The relational database used for the prototype is PostgresSQL, and the document repository is stored in a local directory. Both are the inputs for the Linked data layer.

Linked data layer

In this layer were implemented three main components: RDB2RDF mapping, Ontologies register, and the UpRDF tool. The RDB2RDF mapping component aims to generate a mapping document from the relational

database scheme (see Figure 4). This mapping document is generated reusing the R2RML parser tool Konstantinou et al. (2014) and manually customized using the Ontologies register component. This component is able to register the prefixes, a brief description, and the URI of all ontologies used for modeling the data from the relational database.

```
map:record
    rr:logicalTable [ rr:tableName '"record"'; ];
    rr:subjectMap [ rr:class fabio:JournalArticle;
    rr:template 'http://localhost/record/{"id"}'; ];
    rr:predicateObjectMap [
        rr:predicate dc:source;
        rr:objectMap [ rr:column '"source"'; ];
    ];
    rr:predicateObjectMap [
        rr:predicate fabio:hasPublicationYear;
        rr:objectMap [ rr:column '"year_pub"'; ];
    ];
    rr:predicateObjectMap [
        rr:predicate dc:title;
        rr:objectMap [ rr:column '"title"'; ];
    ];
    rr:predicateObjectMap [
        rr:predicate bibo:uri;
        rr:objectMap [ rr:column '"url_identifier"'; ];
    ];
    rr:predicateObjectMap [
        rr:predicate fabio:abstract;
        rr:objectMap [ rr:column '"description"'; ];
    ];
    rr:predicateObjectMap [
        rr:predicate dc:date;
        rr:objectMap [ rr:column '"date"'; ];
    ];
```

Figure 4. R2RML mapping for the entity record.

After mapped and customized the R2RML document with existing ontologies, the triples RDF are generated. For that, we implement the UpRDF tool. This process is known as the RDF-ization process. During this process, we implement an incremental update approach. It means that the RDF graph generated is updated regularly after generated in the first execution of the RDF-ization process.

After the RDF-ization process, we have to discover and create valuable links between the RDF graph generated and others RDF graph existing in the web of data. For that, we reuse the Silk framework Volz et al. (2009), a similarity metrics based tool for discovering links between resources in the RDF graphs. We discover $owl : sameAs$ links between authors existing in our RDF graph and others RDF graphs published in the web of data.

Finally, the RDF graph generated is stored in the triplestore Jena Fuseki. The publication component provides

a web-based interface to configure a Linked Data Browser based on the Pubby tool. The aim of the publication component is to provides several ways of RDF publication for both, humans and computers. Also, this component provides an SPARQL endpoint for querying the triples by others their-parties applications and the components implemented in the semantic digital library services layer.

Semantic digital library services layer

The aim of this layer is to provide a set of value-added services to the users of the library linked data. More specifically, we implement three main services: an information retrieval service, a content recommendation service, and a community detection service. The information retrieval service is based on searching and browsing paradigms. In the first implementation of this service, we got a slow query response time. The queries were written and executed directly over the SPARQL endpoint provided by the Jena Fuseki triplestore. In a second iteration of the implementation, we introduce a novel method for indexing RDF graph into an indexing server, in our prototype, we use ElasticSearch server Mariño Molerio et al. (2018). The synchronization process between the Jena Fuseki and the ElasticSearch server is done automatically. Experimental results showed an important improvement in the query response time over the ElasticSearch server without losing the advantages of RDF as a data model.

Also, we provide a basic service for content recommendations based on similarity distance between the titles of the research papers in the RDF graph. The algorithm take advantage from the external $owl : sameAs$ links between resources, for example, it is able to recommend further papers written by the same author.

Finally, we implemented a communities detection service based on the method proposed by Ortiz-Muñoz and Hidalgo-Delgado (2016). The community detection refers to the problem to identify communities or partitions of nodes that shares common properties in a network. The approach considers the co-authorship relationships as an indicator to measure scientific collaboration and provide three steps to follow for detecting community taking into account the authors in the RDF graph published. The steps are co-authorship network modeling, communities detection, and communities visualization.

The prototype was implemented instantiating the proposed framework in Section Proposed approach. Each layer includes a set of tools implemented and reused in some cases. The prototype showed the feasibility of the proposed framework for improving the linked data based semantic interoperability of the digital libraries. The framework is independent of the tools and standards used in the construction process. It is able to integrate data from heterogeneous and distributed data sources and takes into account the quality of the bibliographic metadata before publishing them following the linked data principles. Also, the framework provides a set of value-added services for the digital libraries users.

The framework and the implemented prototype is based in three layers. Each layer intends to be aligned with the original research question defined in the Section Problem Statement. The RQ1 was focused on the data integration method used by the framework, in this case, we use the wrapper-based data integration method. It allows handling of several data sources at the same time, without change the subsequent components. The RQ2 was focused on the RDF-ization process and provides a set of best practices for publishing library linked data. Finally, the RQ3 was able to identify the most useful value-added services for library linked data consumption process.

Conclusions and future work

In this paper, we proposed a linked data-based semantic interoperability framework for digital libraries. With the publication of bibliographic metadata, following the linked data principles, we provide a mechanism of semantic interoperability that enables the discovery and reuse of the bibliographic metadata by other computer systems. The framework includes an innovative solution for integrating bibliographic data from heterogeneous and distributed data sources and for improving the quality of the bibliographic metadata form early stages of the development. The framework is extensible, by adding new components, metadata fields, ontologies, and value-added services. The prototype implemented show the feasibility of the proposed framework and was implemented using up-to-date standards and tools.

As future work, we will carry out experiments to measure some metrics and characteristics of the implemented prototype, such as the quality of the linked data published and comparing them with the quality of others related linked data in the web of data. Moreover, we are going to design several use cases to evaluate the framework and the implemented tools from the user perspective in real environments.

Acknowledgements

This work was partially supported by the China Scholarship Council (CSC) and the University of Informatics Sciences, Cuba, under the research project “Semantic Interoperability in Digital Libraries”.

References

AnHai Doan, Alon Halevy, and Zachary Ives. *Principles of Data Integration*. Elsevier, USA, 2012. ISBN 978-0-12-416044-6.

Ceri Binding and Douglas Tudhope. Improving interoperability using vocabulary linked data. *International Journal on Digital Libraries*, 17(1):5–21, March 2016. ISSN 1432-5012, 1432-1300. doi: 10.1007/

s00799-015-0166-y. URL https://link.springer.com/article/10.1007/s00799-015-0166-y.

Gustavo Candela, Pilar Escobar, Rafael C. Carrasco, and Manuel Marco-Such. Migration of a library catalogue into RDA linked open data. *Semantic Web*, 9(4):481–491, January 2018. ISSN 1570-0844. doi: 10.3233/SW-170274. URL https://content.iospress.com/articles/semantic-web/sw274.

Ya-Ning Chen. A Review of Practices for Transforming Library Legacy Records into Linked Open Data. In *Metadata and Semantic Research*, Communications in Computer and Information Science, pages 123–133. Springer, Cham, November 2017. ISBN 978-3-319-70862-1 978-3-319-70863-8. URL https://link.springer.com/chapter/10.1007/978-3-319-70863-8_12.

Janneth Chicaiza, Nelson Piedra, Jorge Lopez-Vargas, and Edmundo Tovar-Caro. Recommendation of open educational resources. An approach based on linked open data. In *Global Engineering Education Conference (EDUCON), 2017 IEEE*, pages 1316–1321. IEEE, 2017.

European Commission. European Interoperability Framework for European Public Services. Technical Report 744, Bruxelles, 2010. URL http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf.

Katherine Crowe and Kevin Clair. Developing a Tool for Publishing Linked Local Authority Data. *Journal of Library Metadata*, 15(3-4):227–240, October 2015. ISSN 1938-6389. doi: 10.1080/19386389.2015.1099993. URL https://doi.org/10.1080/19386389.2015.1099993.

Avigdor Gal. Uncertain Entity Resolution: Re-evaluating Entity Resolution in the Big Data Era: Tutorial. *Proc. VLDB Endow.*, 7(13):1711–1712, August 2014. ISSN 2150-8097. URL http://dl.acm.org/citation.cfm?id=2733004.2733068.

Lise Getoor and Ashwin Machanavajjhala. Entity Resolution: Theory, Practice & Open Challenges. *Proceedings of the VLDB Endowment*, 5(12):2018–2019, 2012. URL http://dl.acm.org/citation.cfm?id=2367564.

M. Hallo, S. Luján-Mora, and J. Trujillo. Transforming Library Catalogs into Linked Data. *Proceedings of the 7th International Conference of Education, Research and Innovation*, (ICERI2014):1845–1853, 2014. ISSN 2340-1095. URL https://library.iated.org/view/HALLO2014TRA.

María Hallo, Sergio Luján-Mora, Alejandro Maté, and Juan Trujillo. Current state of Linked Data in digital libraries , Current state of Linked Data in digital libraries. *Journal of Information Science*, 42(2):117–127, April 2016. ISSN 0165-5515. doi: 10.1177/0165551515594729. URL https://doi.org/10.1177/0165551515594729.

Bernhard Haslhofer and Antoine Isaac. data.europeana.eu: The Europeana Linked Open Data Pilot. *International Conference on Dublin Core and Metadata Applications*, 0:94–104, September 2011. ISSN 1939-1366. URL http://dcpapers.dublincore.org/pubs/article/view/3625.

Yusniel Hidalgo-Delgado, Reina Estrada-Nelson, Bin Xu, Boris Villazon-Terrazas, Amed Leiva-Mederos, and Andrés Tello. Methodological Guidelines for Publishing Library Data as Linked Data. In *Information Systems and Computer Science (INCISCOS), 2017 International Conference on*, pages 241–246, Ecuador, 2017a. IEEE. ISBN 978-1-5386-2644-3. doi: 10.1109/INCISCOS.2017.17. URL https://ieeexplore.ieee.org/abstract/document/8328114/.

Yusniel Hidalgo-Delgado, Ernesto Ortiz-Munoz, and Juan Pedro Febles-Rodriguez. A method for integrating bibliographic data from oai-pmh data providers. *IEEE Latin America Transactions*, 15(9):1695–1699, 2017b.

Si Ying Diana Hu and Suri B. Medapati. Systems and methods of using a knowledge graph to provide a media content recommendation, January 2017. URL https://patents.google.com/patent/US9547823B2/en.

IFLA and UNESCO. IFLA/UNESCO Manifesto for Digital Libraries. Technical report, USA, 2011. URL http://www.ifla.org/files/digital-libraries/documents/ifla-unesco-digital-libraries-manifesto.pdf.

Institute of Electrical and Electronics Engineers. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. New York, 1990.

Nikolaos Konstantinou, Dimitris Kouis, and Nikolas Mitrou. Incremental Export of Relational Database Contents into RDF Graphs. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, WIMS '14, Thessaloniki, Greece, 2014. ACM. doi: 10.1145/2611040.2611082.

Alejandro Jesús Mariño Molerio, Juan Carlos Moreira de Lara, Leduan Flores-Riera, and Yusniel Hidalgo-Delgado. Método para la indexación de grafos RDF desde un SPARQL Endpoint. In *Proceedings of the 3rd International Workshop on Semantic Web*, volume 2096 of *CEUR Workshop Proceedings*, pages 98–109, Havana, Cuba, 2018. URL http://ceur-ws.org/Vol-2096/paper9.pdf.

Maria Nisheva-Pavlova, Dicho Shukerov, and Pavel Pavlov. Design and implementation of a social semantic digital library. *Information Services & Use*, 35(4):273–284, January 2015. ISSN 0167-5265. doi: 10.3233/ISU-150784. URL https://content.iospress.com/articles/information-services-and-use/isu784.

Ernesto Ortiz-Muñoz and Yusniel Hidalgo-Delgado. Detección de comunidades a partir de redes de coautoría en grafos RDF. *Revista Cubana de Información en Ciencias de la Salud*, 27(1):90–99, 2016. ISSN 2307-2113. URL http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2307-21132016000100007.

Filip Radulovic, Nandana Mihindukulasooriya, Raúl García-Castro, and Asunción Gómez-Pérez. A comprehensive quality model for Linked Data. *Semantic Web*, 9(1):3–24, January 2018. ISSN 1570-0844. doi: 10.3233/SW-170267. URL https://content.iospress.com/articles/semantic-web/sw267.

Kim Talleras. Quality of Linked Bibliographic Data: The Models, Vocabularies, and Links of Data Sets Published by Four National Libraries. *Journal of Library Metadata*, 17(2):126–155, April 2017. ISSN 1938-6389. doi: 10.1080/19386389.2017.1355166. URL https://doi.org/10.1080/19386389.2017.1355166.

Tanuj Singh and Alka Sharma. Research Work and Changing Dimensions of Digital Library. In *Emerging Trends and Technologies in Libraries and Information Services*, pages 39–42. IEEE, 2015. ISBN 978-1-4799-5532-9.

Dominika Tkaczyk, PaweÅ, Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Lukasz Bolikowski. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(4):317–335, December 2015. ISSN 1433-2833, 1433-2825. doi: 10.1007/s10032-015-0249-8. URL https://link.springer.com/article/10.1007/s10032-015-0249-8.

Iacopo Vagliano, Diego Monti, Ansgar Scherp, and Maurizio Morisio. Content Recommendation through Semantic Annotation of User Reviews and Linked Data-An Extended Technical Report. *arXiv preprint arXiv:1709.09973*, 2017.

Daniel Vila-Suero, Boris Villazón-Terrazas, and Asunción Gómez-Pérez. datos.bne.es: a Library Linked Data Dataset. *Semantic Web Journal*, 4(3):307–313, 2013. URL http://www.semantic-web-journal.net/content/datosbnees-library-linked-data-dataset.

Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Silk-A Link Discovery Framework for the Web of Data. In *Proceedings of the 2nd Linked Data on the Web Workshop*, volume 538, 2009. URL http://vsr-mobile.informatik.tu-chemnitz.de/svnproxy/download/publications/doc/2009/06.pdf.

Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and SÃ¶ren Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016. URL http://content.iospress.com/articles/semantic-web/sw175.