International Journal of Digital Content Management (IJDCM)
Vol. 3, No. 5, Summer & Fall 2022
dcm.atu.ac.ir
DOI: 10.22054/dcm.2022.67524.1083

ATU PRESS

Original Research

Semantically integrating Digital Libraries: Proposed Architecture

Mehdi Alipour Hafezi*

Assistant Professor, Knowledge and Information Science, Faculty of Psychology and Educational Sciences, Allameh Tabataba'i University, Tehran, Iran.
meh.hafezi@gmail.com

Receive Date: 16/04/2022 Revise Date: 07/05/2022 Accept Date: 26/05/ 2022 Publish Date: 10/06/2022

Abstract

Purpose: Library users need to have integrated access to digital library content and services. Preparing these services needs syntactic and semantic interoperability infrastructure. Nowadays most digital library applications syntactically have the ability to exchange with each other. Therefore the purpose of this paper is to propose semantic integration architecture for digital libraries to enable them to prepare new semantic services for their users.

Method: This study was carried out in the following steps: firstly research literature was collected and studied. Studying research literature helps the researcher to prepare a primary version of the checklist to collect data in this study. The next step aimed to survey the semantic interoperability ability of digital library management systems. In this regard, 26 Iranian digital library management systems were studied by the researcher-made checklist in the previous step. The studied digital library management systems have syntactic

* Corresponding Author: meh.hafezi@gmail.com
How to Cite: Alipour Hafezi, M. (2022). Semantically integrating Digital Libraries: Proposed Architecture, *International Journal of Digital Content Management (IJDCM)*, 3(5), 51-77.
DOI: 10.22054/dcm.2022.67524.1083

eISSN: 2717-3313
ISSN: 2717-3305

interoperability functionalities. Then this study's semantic integration architecture was proposed according to the analysis and results of the study. Finally, the "demonstration" method was used to assess the proposed architecture.

Findings: Findings showed that studied digital library management systems focused on textual materials and paid very little attention to other library information resource formats. Moreover, in many cases, the studied digital library systems did not fill the contents of semantically needed metadata fields. On average, they attempt to document metadata fields such as subject and keywords, by existing reference materials such as Library of Congress Subject Headings and subjective thesauruses that they needed.

Conclusion: According to the findings, the architecture of the semantic integration model based on the results of this study was proposed with three data, inference machine, and application layers.

Keywords: Semantic interoperability, Digital libraries, Integration.

1. Introduction

Information and Communication Technology (ICT) facilitate producing and publishing of information for everyone. Given the possibility of producing and publishing content on the web, managing them has become a major challenge. Based on this challenge, the general consequence of this facilitation of dissemination is the acceleration of the flow of information in the digital environment. A huge amount of information and information resources currently exist, and as Bornmann and Mutz reported the growth rate of scientific publication up to 2012 was 8 to 9 percent (Bornmann & Mutz, 2014). Consequently, identifying and making access to information became a major challenge for users.

The development of general and specialized search engines on the web is a pragmatic solution to overcome the mentioned challenge. Moreover, activities of systems, such as OCLC, to establish big

repositories or warehouses to manage digital information and facilitate their access are on this path. On the other hand, some organizations and companies, such as big publishers like IEEE, and Emerald. Etc., attempt to manage their publications in warehouses and some other enterprises try to collect and serve other organizations' publications (such as NDLTD) in one repository and make them accessible for users. The main issues in creating access to these reservoirs are the high cost of access and their dispersion. In this fairground, libraries are a part of the process of making access to information materials. Libraries try to make access to various types of information repositories for their users. The low cost or free use of libraries has made them an important source of access to information. But in terms of access to information, the diversity of repository systems from publishers to brokers and libraries remains a challenge. Consider a person with information needs, who require access to information or information materials. The first tool that most people try to use at first glance is public or less often specialized web search engines. But in the case of scientific resources, these tools cannot be useful enough because of their limitation in coverage and their limitation on public resources. Most scientific resources would not be retrieved by web search engines, because of their limited access to index commercial scientific databases. In these cases, people will not make access to their needed information materials; even in some cases, they will not be aware of their existence (Mai, 2003). Certainly parts of these users are researchers and scientists and this situation has a direct negative effect on their research results. In practice, this statute is more pronounced in developing countries. One of the results of this constraint may lead to duplicate research or starting new projects without referring to some related literature. Therefore, this limitation affects the results of future research.

In this regard, the main mission of libraries is to create quick, easy, and accurate access to worthwhile information resources (Wallace, 2004). With the advent of the Web, libraries attended quickly to the web environment and offer their services and information materials through it. Following these developments and applying ICT technologies, we saw new versions of libraries that digitally provide their content and services entitled digital libraries (DLs) (Lesk, 1997). The informal definition of DL is that they are libraries that store their content in digital formats and provide them to their users in a digital

environment and through information networks, along with library services. (Arms W., 2000).

Now DLs are in progress and as a result, we encounter with many scattered DLs all over the world. Factually researchers need to access the subjective DLs and related repositories to have a holistic view of their research domains. In such situations, the main issue is the existence of multiple DLs as separate islands without any connections between them. Nowadays making integrated access to information and information resources seems vital. So it is essential to establish connections among DLs and other scientific repositories. Enforcement strategy in this area attempts to reduce users' efforts to identify and retrieve their needed information. At the moment users do not have enough time and patience to separately refer to repositories and DLs to access their needed information. In most cases being aware of all of the existing information, and repositories are impossible. Therefore the practical solution is to integrate DLs and information repositories to present advanced and integrated services.

Nowadays users need to retrieve their needed information semantically from DL systems. We saw that the web path of development in web 3.0 is allocated to the semantic web. Therefore, users with no or limited ability of semantic knowledge refer to data repositories and request semantic retrieval. In this regard, DLs should prepare semantic retrieval services for their users. This service would be complicated when we try to add it to the integrated systems. In other words, preparing semantic integration by disparate DLs need a complicated method and architecture.

Interoperability factually is the opportunity for information exchange and integration in and among communities and applications (Arms, et al., 2002) (Chen H., 1999) (Moen, 2001) (Tennant, 2001)(Zeng & Chan, 2004). To prepare interoperability services, we need to connect repositories to work as a net. Information systems firstly should connect physically or conceptually. Afterward, we can offer value-added services on it. One such important service is preparing semantically connected repositories. In other words, interoperability in DLs includes syntactic and semantic interoperability (Shen, 2006). In practice, syntactic interoperability should be based on creating integration in DLs. After preparing syntactic interoperability, we can develop semantic interoperability to improve the quality of integrated services (Martínez-Costa,

Menárguez-Tortosa, & Tomás Fernández-Breis, 2013). Semantic interoperability is the ability of information systems to exchange information based on shared, pre-established, and negotiated meanings of terms and expressions (Veltman, 2001) (Loutas, Kamateri, & Tarabanis, 2011). According to the mentioned definition of the semantic level of interoperability, we need to make semantic relations between terms and expressions.

The main purpose of this study is to propose semantic integration architecture to enable distributed digital libraries in preparing new semantic services for their users. Therefore, this study surveys the methods of making semantic interoperability in DL systems. Then according to the current situation of DLs in the field of semantic interoperability, a suitable architecture was proposed in this paper. In this way, the following sections are devoted to studying the literature. The methodology of doing this study is the next section. Afterward, the proposed architecture was presented and finally, the architecture was assessed.

2. Literature Review

Interoperability is a general domain in integrating computer systems. Each administrative domain that needs to prepare integrated access to its content and services, requires interoperability functions. A study on this subject shows that diverse domains such as Banking, Health services, E-commerce, M-commerce, DLs, and so on, apply interoperability techniques to exchange data between disparate and distributed computer systems. Requirements to prepare integrated access to information resources and their digital contents in scattered DLs show the importance of this operation, and interoperability, for DLs. For instance, union catalogs such as Worldcat, provided by OCLC, were developed to achieve this goal at the metadata level. Moreover, a study in semanticist integration of information systems, shows that research projects and studies have been performed since 1985, and according to retrieved resources in this study, so far 29 research papers have been carried out in this research domain.

Some research projects allocated to applicable tools such as WordNet (Szyma´nski, 2011), Ontology or RDF (Chen, Finin, & Joshi, 2003) (Vetere & Lenzerini, 2005) (Hunter, 2003) (Sahay, Zimmermann, Fox, Polleres, & Hauswirth, 2013) (Chen Y.-N., 2015) (Agosti, Ferro, & Silvello, 2016), linked data technology (Bizer,

Heath, & Berners-Lee, 2009) (Bizer, 2009) (Moon & Han, 2016)(Hidalgo-Delgado, Xu, Jes´us Mari˜no-Molerio, Febles-Rodr´ıguez, & Abel Leiva-Mederos, 2019), and intelligent techniques (Martín, León, & López, 2015) that are used in making semantic integration. Some others were sub-collections of larger projects that were related to library information exchange and semantic integration. DELOS (DELOS, 2004) and Telematics for library programs (2000) are examples of these kinds of projects. Telematics includes 102 sub-projects in four main categories. One of the objectives of this macro project is related to networking libraries. Five cases of these projects were related to interoperability, but no one of them was related to semantic integration.

One of the sub-categories of the DELOS project was related to semantic interoperability. The related project was named "Semantic Interoperability in DL Systems" which was performed in 2005. In the mentioned project data structure, categorical data, and factual data were mentioned as levels of semantic interoperability in DLs. Also, the six following domains proposed for semantic interoperability: standards, core ontology, knowledge organization systems, semantic services such as metadata and registration terms, the role of architecture, and infrastructures such as syntactic coding, identifiers, protocols, and web service semantic descriptions (DELOS, 2005).

Another research is a project that covered semantic interoperability in the case of Electronic Health Records (EHR). The proposed architecture in this project includes five layers: Layer 1 is allocated to structured heterogeneous data; Layer 2 is for semantic mapping; Layer 3 covered semantic mediator; Layer 4 includes virtual homogeneous data, and finally Layer 5 is related to the Application layer. Moreover, this study describes the challenges of each layer in the demonstrated architectural model (Martinez-costa, Kalra, & Schulz, 2014).

Some other researchers were case studies such as Chan (2004), Warren and Alsmeyer (2005), Guha (2006), Issac, Schlobach, Matthezing, and Zinn (2008) (Mayer, Stumptner, Grossmann, & Jordan, 2013) (Nisheva-Pavlova, Shukerov, & Pavlov, 2015), (Ahmad Khan & Bhatti, 2018) (Fafalios, et al., 2021). For instance, Issac, Schlobach, Matthezing, and Zinn (2008) studied cultural heritage by presenting controlled vocabularies. Others were related to using specific techniques such as RDF (Han, 2006) or other related standards (Macedo & Isaías, 2013) some others were allocated to

introducing the specific models in semantic retrieval (Huang, Ke, & Yang, 2005) (Pasad & Madalli, 2008), and some others were related to integrating methods and reducing complexity in integration by using reduce the ambiguity of words method and reclassifying resources (Mayer, Mutschke, & Petras, 2008).

One of the important and highly relevant concepts to semantic interoperability in DLs is Linked data. Linked data is about using the Web to create typed links between data from different sources (Bizer, Heath, & Berners-Lee, 2009) (Hidalgo-Delgado, Xu, Jes´us Mari˜no-Molerio, Febles-Rodr´ıguez, & Abel Leiva-Mederos, 2019). It uses URI[1], HTTP[2]and RDF[3]technology to linked data on the Web (Klyne & Carroll, 2004) (Berners-Lee, 2009). One of the projects in this domain is the Linking Open Data (LOD) project (2013).

The main goal of this project is to bootstrap the Web of data by identifying existing data sets that are available under open licenses, converting these to RDF according to the Linked Data principles, and publishing them on the Web (Bizer, Heath, & Berners-Lee, Linked Data - The Story So Far, 2009).

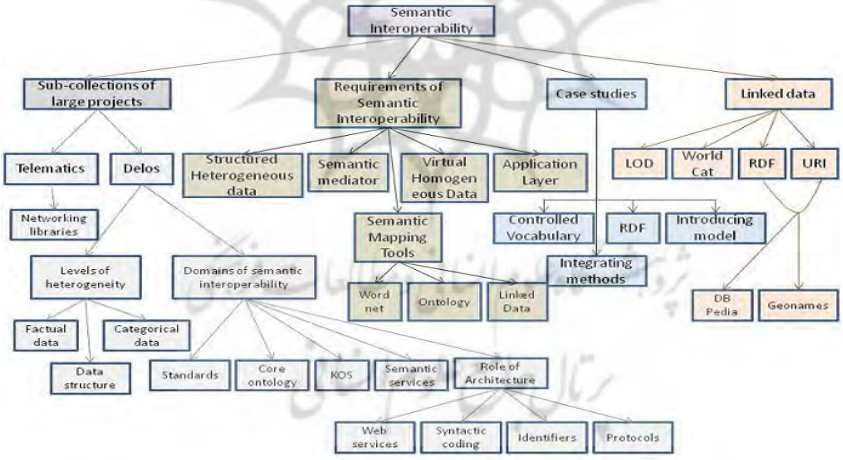

Figure 1. Status of research projects in semantic interoperability

---

[1] Universal Resource Identifier
[2] Hyper Text Transfer Protocol
[3] Resource Description Framework

Now many large databases such as DBpedia (DBpedia, 2014), Geonames (GeoNames Ontology, 2012), and so on present their data by using URI and RDF and take part in this project (SweoIG/TaskForces/CommunityProjects/LinkingOpenData, 2013).

Moreover, OCLC start to use linked data and has created work descriptions for bibliographic resources found in the WorldCat. They bring together multiple manifestations of work into one logical authoritative entity. As mentioned on the website this is the first step in what will be an evolutionary and revolutionary journey, to provide interconnected linked data views of the rich entities (works, places, concepts, people, organizations, and events) captured in the vast shared collection of bibliographic records that make up WorldCat (Data Strategy and Linked Data, 2014).

Figure 1 shows that the research can be clustered into 4 categories, Sub-collections of large projects, Requirements of semantic interoperability, Case studies, and Linked Data. Telematics and DELOS are two mega projects that in some cases address semantic interoperability. DELOS project refers to levels and domains of semantic interoperability. They pay just attention to the role of architecture in semantic interoperability. Projects in the requirements of semantic interoperability demonstrate the five data structures, semantic mediators and tools, and application layers. Case study research allocated more to the controlled vocabulary, integrating methods, RDF, and in some cases introducing models. Finally linked data projects try to show the application of linked data and needed tools such as DB Pedia and Geonames.

All, a study in the literature demonstrates that making integrated access to information services is extremely important, especially in the healthcare domain. Also even though studied researchers use a suitable model in preparing semantic interoperability, no one of them offered a model or even resulted to propose architecture. Also, studied projects use their business model and do not publish their applied architecture. Moreover, the LOD project is the only project that is used to connect data and documents to link the Web published data. Also, OCLC newly started to link their bibliographic data, and metadata, on the Woldcat. Therefore this study tried to propose semantic interoperability architecture for DLs in a specific area based on a study of the current status of DLs. So according to the literature, this study tries to answer the following questions

1. How is the status of semantic metadata fields in studied DLs?
2. How is the status of documenting metadata fields in studied DLs?
3. How would be the architecture of semantic integration in studied DLs?

3. Methodology

To achieve the main purpose of the study, related databases, journals, and operational and research papers were surfed and studied primarily. Also, their references were checked to find more related research resources. In this step, the main goal was to identify the status of research and projects in DLs semantic integration. Therefore, related concepts to semantic interoperability architecture are marked and deeply studied to extract the main needed idea. Also, this step helps us to create the initial version of the checklist that is completed and used in the next step to collect data.

The next step was related to studying the current status of DLs in the semantic integration domain. The suitable research method that can be used in this step was a descriptive survey. Hence this method was applied to recognize the circumstance of DLs from the semantic integration perspective. Studying the current status needs the researcher to refer to DL's website and collect data. In this way, a researcher-made checklist is prepared and used to collect data. The checklist as mentioned was based on the literature study. The main issue that was taken into consideration was the ways of filling metadata fields by DLs. The checklist was examined by specialists according to research objectives. Therefore, the validity of the checklist was obtained by experts in the field of meaning and semantics in information systems. In this regard, six experts in the field of linguistics, library and information science, and computer science identified and justified the purposes of the study. Some modifications were made to the checklist based on the expert's suggestions. Consequently, the final version of the checklist was prepared.

The current status of Iranian DLs was studied by the checklist as a sample case. These cases were selected because of their being accessible to researchers and having the ability of syntactic interoperability based on the prior study (Alipour-Hafezi, Horri, Shiri,

& Ghaebi, 2010). The other cause was their similarity to DLs in developing countries. As a result, 26 DLs were detected in the research population. These DLs were studied by the researcher-made checklist to identify their situation and readiness for preparing semantic interoperability. The researcher visited DLs and collected data, based on the preparation checklist. Also in some cases, researchers are compelled to have interviews with DLs' authorities to collect actual data or get approved for the pre-collected data.

The next step was allocated to design semantic integration architecture. In this step, modeling methodology was used to design and showed the proposed architecture. To prepare the architecture, the used models in integration, not limited to DLs, and also the Iranian DLs status and readiness in semantic interoperability were studied, and finally, a suitable model was designed and proposed.

Evaluating the proposed architecture was the last step in this study. There are different ways of evaluating and validating models. In this study, firstly, the problem was accurately described and then the proposed conceptual model was constructed to overcome the mentioned issue. In this situation, the suitable method for evaluating the model can be the "demonstration" method. In this method, primarily, the issue should be described and then the solution should be constructed. The construction of the solution will show that the solution is realizable (Vaishnavi & Jr., 2008). Therefore the constructed architecture is validated by information technology specialists, library and information science professionals, and system analysts. In this way firstly the aims and major findings of the research were presented to the mentioned specialists and then the designed architecture was presented to them. Following this, their recommendations were received and minor improvements were done based on their point of view. Consequently, the final version of the architecture was obtained.

4. **Findings**

Identifying and analyzing the research literature, direct us to study the status of Iranian DLs from the semantic integration perspective. In this way, two main fields are examined. The first one was studying the contents of bibliographic fields that covered semantic data (questions 1 and 2). It means that the fields are considered and their contents can be used for semantic data links. The next one was related to

identifying semantic integration architecture that is suitable to implement on Iranian DLs (question 3). Therefore, first of all, the fields of interest in metadata fields are identified. Title, author, co-author, description, keyword, subject, publisher, and an identifier such as ISBN, ISMN, DOI, etc. are the related metadata fields. So the 26 Iranian DLs were examined in terms of how to complete the mentioned bibliographic fields.

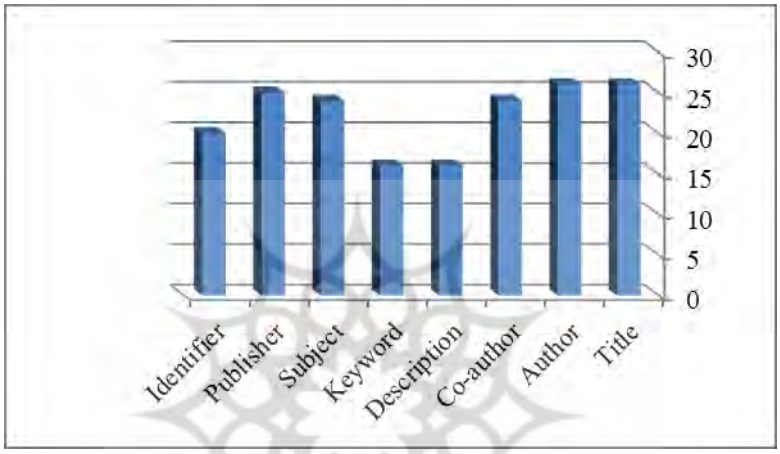

Figure 2. The status of Iranian DLs in filling the meaningful metadata fields

Findings showed that Title and Author (100%), Co-author (92%), Subject (62%), and Publisher (96%) fields completed by most of the studied DLs as showed in figure 2. The next question was the documenting content of the Author, Co-author and publisher fields.

Table 1. Documenting fields by studied DLs

Author documentation	No.	Percent	Publisher documentation	No.	Percent
Authors' names do not documented	3	12	Publishers' names do not documented	9	35
Using legal page	6	23	Using legal page	7	27
Document listing of celebrities and authors' names	16	62	Document listing of institutions and government agencies' names	5	19

Reference databases (such as LOC)	20	77	Publishers' list	4	15
Exist document list in DL software	4	15	Publisher website	5	19

According to table 1, most of the studied DLs (about 62%) predicate the authors' names, especially with reference databases. This could help us to be faced with homogeneous data in this field in making integrated access to data. Also, some of the used DL applications (about 15%) have the ability in smoothing out the names.

Results in the field of publishers' names, mentioned in table 1, showed that just 19% of publisher names were documented by a standard or common list. Nonetheless, since some databases were reviewed by researchers the publishers' names were homogeneous. Using uniform names by publishers helps keep homogeneity at the content level.

In the field of subject analysis, the status of studied DLs was as table 2.

Table 2. Methods of subject analysis in studied DLs

Methods of subject analysis	No.	Percent
Using subject headings	22	85
Free indexing by author keywords	10	38
Free indexing by indexer keyword	5	19
Indexing by thesaurus	14	54
Indexing by ontology	0	0

As it is obvious in Table 2, most of the DLs use subject headings (85%) to allocate subjects for information resources as they do for books in libraries before. Also indexing by thesaurus (with 54%) is mostly used. Nonetheless, other ways such as using free indexing are also used to average. This shows that a diversity of tools is used to demonstrate information resources subjects. Also, the highly used tools in allocating subjects are Library of Congress Subject Headings (LCSH) with 73%, Persian subject headings with 69%, and Cataloging in Print (CIP) with 58%. Also, a wide variety of thesauri is used for indexing by studied DLs.

5. Proposed architecture for semantically integrating DLs

Semantic integration requires information systems' perception of the semantics of the user's information request and those of information resources and uses mediation or information brokering to satisfy the information request as well as it can (Sheth, 1998). In principle, there are two paradigms in DL integration: data warehousing, which is known as harvesting, and on-demand retrieval, which is known as federated search (Vdovjak & Houben, [2001]) (Hoffer, Ramesh, & Topi, 2011). In the data warehousing approach, all necessary data files are collected in a central repository before the user's query is issued. For example, the OAI-PMH protocol is used by DL systems to collect and harvest metadata from scattered DLs or other information repositories. On-demand driven approach collects the data from integrated sources dynamically during query evaluation. As an example, SRU/W protocols are used by DL applications to collect library information resources' metadata. It is essential to know that metadata standards based on XML (Extensible Markup Language) should be used by DLs. These two paradigms are classified as the syntactic interoperability of DLs. This level of interoperability is essential before trying to create semantic integration (Shen, 2006).

Based on the findings of this study, presented in the previous section, and previous research (Alipour-Hafezi, 2008) (Alipour-Hafezi, Horri, Shiri, & Ghaebi, 2010), that are mentioned in the literature review section, three layers architecture proposed. The data layer, Inference machine layer, and Application layer are the three mentioned layers. These three layers are base layers that each interoperability system needs at the semantic level. The following sections try to describe the proposed architecture in each mentioned layer (presented in figure 3).

5-1. Data Layer

In Iran, like in other countries, there are heterogeneous DLs that use different library resource management software (Alipour-Hafezi, 2008). Moreover, as mentioned in the findings section, different applications store data in different ways, and as a result, different outputs are provided by the studied applications. These outputs are generated in the software transaction layer. Moreover, DLs generally

cover a variety of library resources including books, articles, theses, research projects, audio files, video files, image files, and more. There are a few metadata standards that can cover all of these diverse formats. On the other hand, the studied DLs, based on the findings, do not provide standard outputs. Moreover, all the library resources' management systems can provide their outputs in XML format (Alipour-Hafezi, Horri, Shiri, & Ghaebi, 2010). However, some of the studied DL management systems do not allow other software to collect automatically, their metadata. Therefore, the syntactic integration requires the use of a hybrid model which can cover Harvesting and on-demand retrieval models simultaneously (Alipour-Hafezi, Horri, Shiri, & Ghaebi, 2010)[1]. Based on this fact, semantic integration should have the ability to work in such a situation. So, all the DLs with each level of data access can participate in this network and become a node in the DL network that prepares a semantic integrated retrieval service.

As semantic integration and retrieval services work at the metadata level, to prepare semantic integration we need a metadata translator to translate diverse outputs of DLs to a unique form. The study directs us to use the Dublin Core (DC) metadata standard with the mentioned 14 related elements: Title, Creator, Subject, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights. As a result, we will have a database that includes all of the DL metadata in the DC standard and its data model (DCMI Usage Board, 2012).

It is important to remember that, some of the studied DLs do not provide standard outputs. Therefore, our database could not cover their metadata. As a result, we would have two data resources: a) standard metadata databases that are accessible by data warehouse crawlers to collect data, and b) distributed DLs that work on-demand and do not allow crawlers to collect their metadata. So far, we have been able to syntactically prepare integrated access to their metadata. The other studied DLs that do not allow crawlers to harvest their metadata, offer their metadata by SRW servers. This connection is established via SRW protocol. A study in the DL interoperability

[1]These findings is related to syntactic interoperability that researcher obtained in the previous research.

protocols directs us to use SRW or SRU. Also, the SRW protocol is proposed in this architecture because it uses the SWOP (Simple Object Access Protocol) protocol on web service side. Moreover, this solution has high security for SRU (Veen & Oldroyd, 2004).

As demonstrated in the right hand of figure 2, the data layer includes two groups of DLs. One group lets the middleware collect their metadata. In such circumstances, the OAI-PMH protocol can be used to harvest their metadata in DC format by the middleware. OAI protocol is a suitable protocol for harvesting models and coordinates with the DC format (Yu, Chen, & Chang, 2005). Since the DLs use different metadata formats and the proposed system needs DC metadata based on the 14 mentioned fields, all the collected data from cloudy DLs should be translated to DC metadata standard. On the other hand, metadata from the other group, which does not let the middleware collect directly their metadata, needs to be collected on-demand by SRW protocol.

These metadata need metadata translation to DC metadata standard. Therefore, they integrate with the results of the data warehouse in the way of responding to requests. At all, a metadata translator is needed on both sides of data collection – harvesting and on-demand models.

5-2. Inference Machine Layer

The next section requires an inference machine. Based on the results, subject headings and thesaurus are mostly used to identify descriptions of information resources. Also, the document list of celebrities' names, authors, and publishers is mostly used to document their names. Therefore, a database is needed to cover these materials and their semantic relations as a network of concepts (linked data). We call these databases, semantic repositories. These infrastructures, factually provide the possibility of conceptual relationships. Metadata is stored in RDF as demonstrated in figure 2. Thus a transaction is automatically implemented to transfer data from XML to RDF. In this step, the RDF schema is needed to transfer metadata. RDF helps us to make links between mentioned metadata elements. In this way, the RDF schema plays a key role (Nilsson, 2008).The other transaction in this step is synchronizing meaningful elements (such as Title, and Subject) and name elements (such as Creator and Contributor) in the semantic store with a data warehouse. This can help the semantic store

to identify the existing data about the users' requests. Therefore, it can answer users initiatively without sending requests to the data warehouse. Also, requests for on-demand systems could be sent by SRW protocol, and their responses receive after the metadata translation stage. The translation filter tries to translate responses to 14 elements of DC as presented in figure 3.

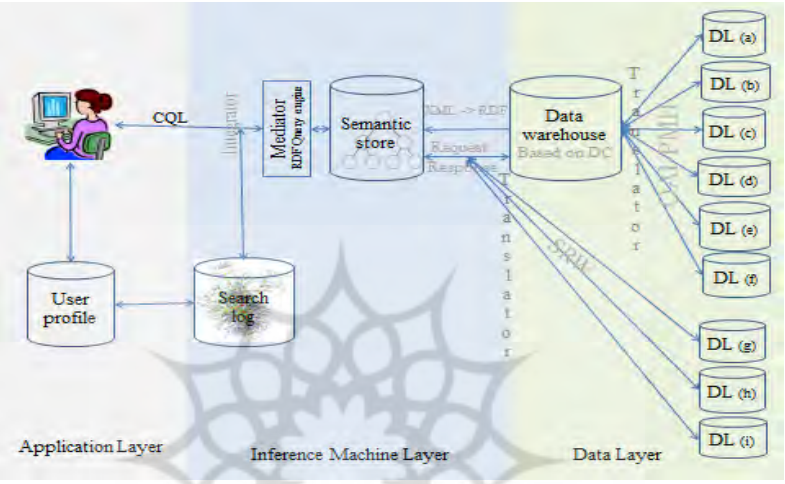

Figure 3. Proposed architecture facilitating semantic integration in heterogeneous DLs

As the data were stored in the RDF schema, a mediator is needed to translate the user query to the RDF query. It is important to know that users' requests firstly change to RDF query to search in semantic store. Findings are combined with the results of on-demand requests and displayed to users in the form of semantic relations. If users request descriptive metadata, it would be presented directly, and for the full-text s/he is directed to the related DLs.

Another transaction that simultaneously occurs, is recording the user's search logs. These search logs are recorded in a separate database along with user preferences in access to data. After a distinct time (it is proposed to use 6 months period) we have the ontology produced by user preferences, which helps the system to promote its responses. When saturated, this service can be integrated with the system responses. This functionality helps the system recommend keywords, resources, and so on. This service can be an added value service that promotes the system response in suggesting search

keywords or recommending semantically related information resources.

For instance, as a scenario, suppose a user login to the system and send a request for a keyword such as "cloud computing". First of all, the system translates the query to an RDF query. Afterward, the query is sent to the semantic store database. The diverse links are created by the existing semantic corpus. If a user is discovered by system, his/her search log can help system to recommend related keywords, subject or resources based on the search log and user profile. As mentioned above, system ontology makes links and recommend some semantically related resources. Moreover, the query and its semantically related keywords send to the data warehouse and other DLs to receive related metadata. Afterward, systems respond to the request and send back-related metadata records. The middleware system combines responses and recommends them to users. Access to the full image of each selected record will be provided after authentication.

5-3. Application Layer

In this layer web-based, the application receives user requests and transmits them to the next layer. After receiving the results, the application uniforms the results and displays them. Another task of the application identifies licensed users to use the integrated network. This database covers all the user's profiles which are members of DLs in the network. Therefore, the rate of access to digital objects and license agreements are checked in this step. Also, the database has been in direct connection with the search log database to store and classify search logs. User profile database collect and up-to-date its metadata about DL applications that are a member of the integrated network. Interoperability in this level has been done simply because of using uniform metadata.

5-4. Model validation

The "Demonstration" method is used to validate the proposed model (Vaishnavi & Jr., 2008). In this way the proposed conceptual model along with the problem statement and objects of this study presented to experts and specialists in information technology, library and information science, and system analysis domains. They accurately matched the proposed architecture with the problem

statement and objectives of this study and finally confirmed the model.

6. Discussions

Results in the case of the status of semantic metadata show that Title, creators (author, co-author/s, and so on), and publisher are the fields that studied DLs try to fill their metadata contents in their descriptive metadata systems. Nevertheless, they had documented fewer data fields by standard tools. This shows that DLs should pay more attention to the filling and also documenting of meaningful data fields. This is the primary element in making homogeneous data in semantic integration (DELOS, 2005). Moreover integrating metadata needs semantic tools to link data. Whether DLs use documenting tools such as thesauruses, subject headings, and so on in their bibliographic elements, as described in the proposed architecture, they can be used as knowledge map tools in linking their data. Therefore, documenting data plays an important role in semantic integration.

Preparing semantic integration is a suitable way to promote DL services. Logically serving semantic retrieval services at least need 3 layers and try to reduce complexity (Mayer, Mutschke, & Petras, 2008) (Pasad & Madalli, 2008) as it is respected in the proposed architecture, while some proposed architectures such as Martinez-costa, et al. do not respect this structure (Martinez-costa, Kalra, & Schulz, 2014). The proposed architecture in this study, factually, has some qualifications:

First of all, the architecture tries to simultaneously cover DL networks by harvesting and on-demand retrieval (Data layer). In practice, most integration activities need to pay more attention to DL's interests. As it is respected in the proposed architecture some of the studied DLs authorize crawlers to access their content and in return others make limitations in harvesting. So, suitable architecture must respect this fact.

Second, semantic integration needs uniformity at the metadata level. While each DL preserve its metadata in a different structure and offer output with different metadata standard. Therefore semantic interoperability architecture needs a step to convert diverse standards to an agreed standard. Diversity in metadata collection complicates the solution. The proposed architecture use translator in the way of both data collection models, harvesting, and on-demand retrieval. The task

of the translator is to convert diverse metadata standards to DC by the 14 mentioned elements.

Third, semantic retrieval services need a knowledge map to make semantic relations (Pasad & Madalli, 2008). The semantic store (inference machine layer) does the mentioned task in the proposed architecture. According to the findings, subject headings, thesauruses, and authority files that are used by DLs to organize their information resources are proposed to be used in the semantic store knowledge base. Therefore, the mentioned knowledge store, based on RDF, can be used to prepare a semantic retrieval service. The mediator stays on the inference machine layer and sends users' query to the semantic store, and consequently, the map of search keywords is created and sent to the data warehouse and other DLs by SRW protocol. Then the system receives responses and uniforms them in needed categories to present to users.

Forth, machine learning ability is another feature of the proposed architecture. The search log database (inference machine layer) can perfectly do this. This database has a connection with the user profile (application layer) and records users' search logs. The logs are recorded for a specified time for example 180 days. This recording helps the system to keep connections between user keywords and received documents. So the system can recommend to users some related documents by RSS service and also have the ability to recommend search keywords.

7. Conclusion

Integrated access to information resources in DLs and other scientific databases is an important solution in getting access to scattered information resources in the current era. DLs and other scientific databases should prepare integrated services to empower their existence and respond to users' increasing requests. It is important to know that this is the best way for DLs to keep their effective place in the information life cycle. Creating integrated access to DLs needs agreements in syntactic and semantic layers (Shen, 2006). As mentioned above, syntactic integration is base for semantic integration. In fact, in syntactic integration service, software, hardware, and NetWare of integration are created. Also, in semantic integration, the concepts of semantically meaningful metadata fields are considered.

In this study, the current status of Iranian DLs is studied in the case of their ability for serving semantic retrieval. Findings showed that the DLs are not in good condition. But some of them use documenting tools for meaningful data and other data that can be used in linking data. This fact demonstrates that other DLs should pay more attention to documenting their data. It is important to know that without documenting data, especially in the 14 studied fields, DLs cannot take part in semantic integration networks even if they have the ability in syntactic interoperability level. However, based on findings architecture of semantic interoperability is proposed. The proposed architecture tries to cover some related parts in the syntactic layer. Also, the model is based on the hybrid paradigm of system integration: data warehousing and on-demand retrieval. The overall architecture consists of three layers: Data, Inference machine, and application layers. The inference machine layer is the most impressive part of the proposed model. The functionality of the inference machine layer in using RDF schema is similar to the functionalities of linked open data projects (Vdovjak & Houben, [2001]) (Bizer, Heath, & Berners-Lee, Linked Data - The Story So Far, 2009) (Moon & Han, 2016) (Hidalgo-Delgado, Xu, Jesús Mariño-Molerio, Febles-Rodríguez, & Abel Leiva-Mederos, 2019) and the approach of OCLC in their Worldcat database (OCLC Worldcat, 2014).

All the proposed models were designed by using the current status of metadata contents in the studied DLs, their interactive capabilities, and interoperability models that are used in the DLs and suggested and examined in research reports that some important of them mentioned above. The model can semantically answer users' information requests. Moreover, the machine learning ability of the system helps it besides offering semantic retrieval, can learn and prepare accurate responses based on changing users' requests over time. In addition, given that machine learning is done through user query keywords, it can provide and develop new semantic relations more than the previous existing relations in subject headings and thesauruses.

Also supporting a mixed method in syntactic structure, based on the current situation of the studied DLs (Alipour-Hafezi, 2008), can improve its usage in the same situations other than the studied context. On the other hand, the proposed model uses a simple metadata standard, DC with the 14 main data fields. Various forms of metadata

in digital libraries have led to the use any other formats to increase system complexity.

Reference

Agosti, M., Ferro, N., & Silvello, G. (2016). Digital library interoperability at high level of abstraction. *Future Generation Computer Systems, 55*, pp. 129-146. doi:10.1016/j.future.2015.09.020

Ahmad Khan, S., & Bhatti, R. (2018). Semantic Web and ontology-based applications for digital libraries: An investigation from LIS professionals in Pakistan. *The Electronic Library, 36*(5), pp. 826-841. doi:10.1108/EL-08-2017-0168

Alipour-Hafezi, M. (2008). Interoperability between library software: a solution for Iranian libraries. *The Electronic Library, 26*(5), 726-734.

Alipour-Hafezi, M., Horri, A., Shiri, A., & Ghaebi, A. (2010). Interoperability models in digital libraries: an overview. *The Electronic Library, 28*(3), 438-452.

Arms, W. (2000). *Digital libraries.* MIT press.

Arms, W., Hillmann, D., Lagoze, C., Krafft, D., Marisa, R., Saylor, J., . . . Van de Sompel, H. (2002). A spectrum of interoperability. *D-Lib Magazine, 8*(1). Retrieved June 10, 2013, from http://www.dlib.org/dlib/january02/arms/01arms.html

Berners-Lee, T. (2009, June 18). *Linked Data*. Retrieved July 25, 2014, from W3.org: http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C. (2009). The emerging web of linked data. *Intelligent Systems, 24*(5), 87-92.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems, 5*(3), 1-22.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far . *International journal on semantic web and information systems, 5*(3), 1-22.

Bornmann, L., & Mutz, R. (2014). Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*. Retrieved September 14, 2014, from http://arxiv.org/ftp/arxiv/papers/1402/1402.4578.pdf

Chan, L. (2004). Supporting and Enhancing Scholarship in the Digital Age: The Role of Open Access Institutional Repository. *Canadian Journal Of Communication, 29*(3). Retrieved July 27, 2014, from http://cjc-online.ca/index.php/journal/article/view/1455/1579

Chen , Y.-N. (2015). A RDF-based approach to metadata crosswalk for semantic interoperability at the data element level. *Library Hi Tech, 33*(2), pp. 175-194. doi:10.1108/LHT-08-2014-0078

Chen, H. (1999). Semantic Research for Digital Libraries. *D-Lib Magazine, 5*(10). Retrieved June 10, 2013, from http://www.dlib.org/dlib/october99/chen/10chen.html

Chen, H., Finin, T., & Joshi, A. (2003). An Ontology for Context-Aware Pervasive Computing Environments. *Knowledge engineering review, 18*(3), 197-207.

Creating a European library space telematics for libraries programmes 1990-1998. (2000). Retrieved Dec. 25, 2014, from Cordis: http://cordis.europa.eu/libraries/en/intro.html

Data Strategy and Linked Data. (2014). Retrieved July 23, 2014, from oclc.org: http://www.oclc.org/data.en.html

DBpedia. (2014, July 21). Retrieved July 27, 2014, from Wiki.dbpedia.org: http://wiki.dbpedia.org/About

DCMI Usage Board. (2012, 06 14). *DCMI Metadata Terms*. Retrieved March 12, 2015, from http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#

DELOS. (2004). *Welcome to the DELOS network of excellence*. Retrieved November 18, 2011, from DELOS: http://www.delos.info/index.php?option=com_frontpage&Itemid=1

DELOS. (2005). *Semantic Interoperability in Digital Library Systems.* UKOLN: University of Bath, European Commission within the Sixth Framework Programme. Retrieved October 25, 2014, from http://opus.bath.ac.uk/23606/1/SI_in_DLs.pdf

Fafalios, P., Petrakis, K., Samaritakis, G., Doerr, K., Kritsotaki, A., Tzitzikas, Y., & Doerr, M. (2021). FAST CAT: Collaborative Data Entry and Curation for Semantic Interoperability in Digital Humanities. *Journal on Computing and Cultural Heritage, 14*(4), pp. 1-20. doi:10.1145/3461460

GeoNames Ontology. (2012, November). Retrieved July 22, 2014, from Geonames.org: http://www.geonames.org/ontology/documentation.html

Guha, N. (2006). *Semantic Digital Library Services.* Retrieved Jully 15, 2011, from I3s: http://www.l3s.de/kweb/kwepsy2006/FinalSubmissions/kwepsy2006_guha.pdf

Han, Y. (2006). ARDF-based digital library system. *Library Hi Tech, 24*(2), 234-240. doi:10.1108/07378830610669600

Hidalgo-Delgado, Y., Xu, B., Jes´us Marí˜no-Molerio, A., Febles-Rodr´ıguez, J., & Abel Leiva-Mederos, A. (2019). A Linked Data-based Semantic Interoperability Framework for Digital Libraries. *Revista Cubana de Ciencias Informaticas, 13*(1), 14-30.

Hoffer, J. A., Ramesh, V., & Topi, H. (2011). *MODERN DATABASE MANAGEMENT* (10 ed.). New Jersey: Prentice Hall.

Huang, S.-H., Ke, H.-R., & Yang, W.-P. (2005). Enhancing semantic digital library query using a content and service inference model (CSIM). *Information processing and management*(41), 891-908.

Hunter, J. (2003). Enhancing the semantic interoperability of multimedia through a core ontology. *Circuits and Systems for Video Technology, IEEE Transactions on, 13*(1), 49-58.

Issac, A., Schlobach, S., Matthezing, H., & Zinn, C. (2008). Integrated access to cultural heritage resources through representation

and alignment of controlled vocabularies. *Library Review, 57*(3), 187-199. doi:10.1108/00242530810865475

Klyne, G., & Carroll, J. J. (Eds.). (2004, February 10). *Resource Description Framework (RDF): Concepts and Abstract*. Retrieved July 25, 2014, from W3.org: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

Lesk, M. E. (1997). *Practical Digital libraries, byteel and bucks.* San Francisco: Morgan Kaufman.

Loutas, N., Kamateri, E., & Tarabanis, K. (2011). A semantic interoperability framework for cloud platform as a service. *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on.* IEEE.

Macedo, M., & Isaías, P. (2013). Standards Related to Interoperability in EHR & HS. In M. Ángel Sicilia, & P. Balazote (Eds.), *Interoperability in Healthcare Information Systems: Standards, Management, and Technology* (pp. 19-44). IGI global.

Mai, J.-E. (2003). The future of general classification. In N. J. Williamson, & C. Beghtol, *Knowledge organization and classification in international information retrieval* (pp. 3-12). NewYork: Haworth information press.

Martín, A., León, C., & López, A. (2015). Enhancing semantic interoperability in digital library by applying intelligent techniques. *2015 SAI Intelligent Systems Conference (IntelliSys)* (pp. 904-911). London: IEEE explore. doi:10.1109/IntelliSys.2015.7361251

Martinez-costa, C., Kalra, D., & Schulz, S. (2014). Improving EHR semantic interoperability future vision and challenges . *EHealth - For Continuity of Care: Proceedings of MIE2014* (pp. 589-593). IOS Press.

Martínez-Costa, C., Menárguez-Tortosa, M., & Tomás Fernández-Breis, J. (2013). Interoperability of EHR Systems Based on Semantic Representation and Transformation Models. In M. Ángel Sicilia, & P. Balazote (Eds.), *Interoperability in*

Healthcare Information Systems: Standards, Management, and Technology (pp. 59-81). IGI global.

Mayer, P., Mutschke, P., & Petras, V. (2008). Reducing semantic complexity in distributed digital libraries: treatment of term vagueness and document re-ranking. *Library Review, 57*(3), 213-224. doi:10.1108/00242530810865484

Mayer, W. E., Stumptner, M., Grossmann, G., & Jordan, A. (2013). *Semantic Interoperability in the Oil and Gas Industry: A Challenging Testbed for Semantic Technologies.* Association for the Advancement of Artificial Intelligence.

Moen, W. (2001). Mapping the interoperability landscape for networked information retrieval. *JCDL '01 Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries* (pp. 50-51). New York: ACM.

Moon, H.-k., & Han, S.-k. (2016). A study of Reference Model of Smart Library based on Linked Open Data . *Journal of the Korea Institute of Information and Communication Engineering*, *20*(9), pp. 1666-1672. doi:10.6109/jkiice.2016.20.9.1666

Nilsson, M. (2008, 01 14). *Expressing Dublin Core metadata using the Resource Description Framework (RDF).* Retrieved May 25, 2015, from Metadata Innovations: http://dublincore.org/documents/dc-rdf/

Nisheva-Pavlova, M., Shukerov, D., & Pavlov, P. (2015). Design and implementation of a social semantic digital library. *Information Services & Use, 35*(4), pp. 273-284. doi:10.3233/ISU-150784

OCLC Worldcat. (2014). (OCLC.org) Retrieved July 25, 2014, from OCLC: http://www.oclc.org/worldcat.en.html

Pasad, A., & Madalli, D. P. (2008). Faceted infrastructure for semantic digital libraries. *Library Review, 57*(3), 225-234. doi:10.1108/00242530810865493

Sahay, R., Zimmermann, A., Fox, R., Polleres, A., & Hauswirth, M. (2013). A Formal Investigation of Semantic Interoperability of HCLS Systems. In M. Ángel Sicilia, & P. Balazote (Eds.),

Interoperability in Healthcare Information Systems: Standards, Management, and Technology (pp. 148-183). IGI Global.

Shen, R. (2006). *Applying 5S framework to integrating digital libraries.* Virginia: Virginia polytechnic Institute and State University.

Sheth, A. P. (1998). *Changing focus on interoperability in information systems: from system, syntax, structure to semantics.* Kluwer. Retrieved June 25, 2013, from http://lsdis.cs.uga.edu/lib/download/S98-changing.pdf

SweoIG/TaskForces/CommunityProjects/LinkingOpenData. (2013, November 26). Retrieved July 25, 2014, from W3.org: http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Szyma´nski, J. (2011). Cooperative WordNet Editor for Lexical Semantic Acquisition. *IC3K 2009* (pp. 187-196). Berlin: Springer-Verlag.

Tennant, R. (2001). Different paths to interoperability. *Library Journal, 126*(3), 118-119.

Vaishnavi, V. K., & Jr., W. K. (2008). *Design science research methods and patterns: innovating information and communication technology.* NewYork: Taylor and Francis Group.

Vdovjak, R., & Houben, G.-J. ([2001]). *RDF Based Architecture for Semantic Integration of Heterogeneous Information Sources.* iwayan. Retrieved June 25, 20013, from http://iwayan.info/Research/Interoperability/Papers_Research/SemanticMediation/wiiw01.pdf

Veen, T. V., & Oldroyd, B. (2004). Search and retrieval in the European library: a new approach. *D-lib Magazine, 10*(2). Retrieved May 15, 2015, from http://www.dlib.org/dlib/february04/vanveen/02vanveen.html

Veltman, K. H. (2001). Syntactic and semantic interoperability: New approaches to knowledge and the semantic web. *New Review of Information Networking, 7*(1), 159-183.

Vetere, G., & Lenzerini, M. (2005). Models for semantic interoperability in service-oriented architectures. *IBM Systems Journal*, *44*(4), 887-903.

Wallace, L. K. (2004). *Libraries, Mission & Marketing: Writing Mission Statements that Work.* American library association.

Warren, P., & Alsmeyer, D. (2005). Applying semantic technology to a digital library: a case study. *Library management, 26*(4/5), 196-205. doi:10.1108/01435120510596053

Yu, S.-C., Chen, H.-H., & Chang, H.-W. (2005). Building an open archive union catalog for digital archives. *The Electronic Library, 23*(4), 410-418.

Zeng, M. L., & Chan, L. (2004). Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology*, *55*(5), 377-395.

How to Cite: Alipour Hafezi, M. (2022). Semantically integrating Digital Libraries: Proposed Architecture, *International Journal of Digital Content Management (IJDCM)*, 3(5), 51-77.
DOI: 10.22054/dcm.2022.67524.1083