Semantic Approach for Discovery and Visualization of Academic Information Structured with OAI-PMH

Joanna Alvarado-Uribe[1], Arianna Becerril García[2], Miguel Gonzalez-Mendoza[1], Rafael Lozano Espinosa[1], José Martín Molina Espinosa[1]

[1] Tecnologico de Monterrey, School of Engineering and Sciences, Av. Eugenio Garza Sada No. 2501 Sur, Col. Tecnológico, 64849, Monterrey, N.L., México, {A00987514, mgonza, ralozano, jose.molina}@itesm.mx

[2] Universidad Autónoma del Estado de México, Instituto Literario Ote. No. 100, Col. Centro, 50000, Toluca, Estado de México, México, abecerrilg@uaemex.mx

Abstract: There are different channels to communicate the results of a scientific research; however, several research communities state that the Open Access (OA) is the future of academic publishing. These Open Access Platforms have adopted OAI-PMH (Open Archives Initiative - the Protocol for Metadata Harvesting) as a standard for communication and interoperability. Nevertheless, it is significant to highlight that the open source knowledge discovery services based on an index of OA have not been developed. Therefore, it is necessary to address Knowledge Discovery (KD) within these platforms aiming at students, teachers and/or researchers, to recover both, the resources requested and the resources that are not explicitly requested – which are also appropriate. This objective represents an important issue for structured resources under OAI-PMH. This fact is caused because interoperability with other developments carried out outside their implementation environment is generally not a priority (Level 1 "Shared term definitions"). It is here, where the Semantic Web (SW) becomes a cornerstone of this work. Consequently, we propose OntoOAIV, a semantic approach for the selective knowledge discovery and visualization into structured information with OAI-PMH, focused on supporting the activities of scientific or academic research for a specific user. Because of the academic nature of the structured resources with OAI-PMH, the field of application chosen is the context information of a student. Finally, in order to validate the proposed approach, we use the RUDAR (Roskilde University Digital Archive) and REDALYC (Red de Revistas Científicas de América Latina y el Caribe, España y Portugal) repositories, which implement the OAI-PMH protocol, as well as one student profile for carrying out KD.

Keywords: the Semantic web; knowledge discovery; user profile ontology; ontology merging; OAI-PMH; visualization

1 Introduction

There are different channels to communicate the results of a scientific research; however, the Open Access (OA) is the future of academic publishing [2]. These Open Access Platforms have adopted OAI-PMH (Open Archives Initiative - the Protocol for Metadata Harvesting) as a standard for communication and interoperability. For instance, according to ROAR (Registry of Open Access Repositories), there are more than 4,000 repositories in the World that implement OAI-PMH [19]. Such figures show the consolidation of repositories as well as the large amount of scientific-academic resources following the philosophy of OA, which are available online for query. Nevertheless, it is significant to highlight that the open source knowledge discovery services based on an index of OA have not been developed [8]. Therefore, it is necessary to address the Knowledge Discovery (KD) within these platforms aiming students, teachers, or researchers to recover both useful resources requested and resources not explicitly requested by them – which are also appropriate –. This objective represents an important issue for structured resources under the OAI-PMH protocol. This fact is caused because interoperability with other developments carried out outside their implementation environment is generally not a priority (Level 1 "Shared term definitions") [15]. It is here, where the Semantic Web (SW) becomes very important for this work.

Accordingly, it is noteworthy that our research work arises in the context of the original vision of SW. That is, the vision of providing more meaning to Web information through the logical connection of terms in order to establish interoperability between systems [47]. On the one hand, we consider SW as part of the Open World Assumption (OWA) – also known as the Classical Paradigm [40] – stipulating that there may be unspecified information (considered as unknown) that can be inferred [40]. On the other hand, we take into account that SW is based on the idea of adding more machine-readable semantics to Web information through annotations written in Resource Description Framework (RDF) [26]. Such that, the incorporation of the RDF model, in this work, will bring about two key features: to gain the W3C design principles and some main features of SW, such as interoperability, extensibility, evolution, and decentralization; and to allow anyone can make statements about any resource [26].

Thus, our work arises from these facts, with the aim of designing and developing a solution to allow a user KD on structured resources with OAI-PMH, by applying the SW technologies. Likewise, it is to both to improve the information retrieval and also the visualization of outcomes within this approach. Therefore, it is relevant to develop technologies that support the discovery of interesting resources within the structured repositories with OAI-PMH for any user, taking into account his/her context information [5]. Analogously, due to a large amount of information gathered and integrated, it is necessary to incorporate a layer for data visualization allowing visual interpretation of the retrieved information in a quick, simple, and easy to understand manner [3]. Consequently, we propose OntoOAIV, a semantic approach for the selective knowledge discovery into structured information with OAI-PMH, focused on supporting the activities of scientific or academic research for a specific user.

Because of the academic nature of the structured resources with OAI-PMH, the field of application chosen is the context information of a student. The idea of building a user profile model is supported in the envisioning of some researchers, belonging to the user modeling community, that propose to use and to share the user models' information among applications. By using and sharing this information, we could integrate the preferences, interests, and characteristics of the user into the context of applications in order to enhance the service provided [35]. Hence, a student profile model is developed. This modeling is carried out using an ontology, because it provides common conceptualizations for data integration [47] and represents one of the two major approaches used to address the lack of interoperability in the user modeling [35]. Furthermore, keeping in mind these arguments and that Dublin Core (DC) (Level 1) is part of the OAI-PMH [15] protocol, we use some initiatives implemented in SW, such as DC (Level 2) [26, 13, 14, 15], Friend Of A Friend (FOAF) [26, 9, 1], and the DBPedia's PersonData [12], in order to provide more user context information to the proposed approach. According to the foregoing, we verify that the conceptualization and the SW's technologies are useful for dealing with the protocol's gaps.

Thereby, the main contribution of our work is the designed and implemented semantic approach, which allows KD within academic contents structured with OAI-PMH considering the user's context. Highlighting three specific contributions: the usage of the algorithm for merging ontologies proposed by Ameen et al. [4], the adaptation of the students' representation model based on the ontological approach presented by Panagiotopoulos et al. [39], and the incorporation of the tool for visualizing information stored in triples developed by Alvarado-Uribe et al. [3]. Accordingly, this work also contributes to the four defined rules within the Linked Data area "for publishing, sharing, and interlinking structured data on the Web" [51]. Finally, in order to validate the proposed approach, we use the RUDAR (Roskilde University Digital Archive) [45] and REDALYC (Red de Revistas Científicas de América Latina y el Caribe, España y Portugal) [44] repositories, which implement the OAI-PMH protocol, as well as one student profile for carrying out KD [5].

The paper is organized as follows. In Section 2, a review of the related work is provided. In Section 3, the semantic approach proposed is explained. In Section 4, the experiments of our approach are described. Then, in Section 5, the results of these tests are reported and discussed. The conclusions and future work are given in the last section.

2 Related Work

Scientific collaboration has long promoted the reuse and sharing of knowledge and data widely [29]; therefore, regarding non-commercial solutions, since 2001, a movement has been consolidating. Such a movement promotes free and unrestricted access to scientific content, especially when this content has been publicly funded. This movement is called OA and was formalized by means of three declarations: Budapest [11], Berlin [36], and Bethesda [10]. In this sense, it is important to mention that the beginning of the Directory of Open Access Journals (DOAJ) in 2003, developed by the Lund University in Sweden, ushered in the formalization and organization of OA for the case of scientific journals. Currently, DOAJ contains more than 9,000 open access journals from 128 countries [16].

Repositories, portals, and journals that are integrated into OA adopt – as a good practice – an interoperability protocol to exchange information in order to have communication rules and standards for structuring data. For instance, OAI-PMH is a low-barrier mechanism for the interoperability among repositories [31], which provides a framework for the independent interoperability of the application based on metadata harvesting. It is noteworthy that metadata to be transmitted via OAI-PMH should be coded in the Dublin Core (DC) format within an XML file, which usually includes several DC records depending on the configuration of each data provider and harvesters. The OAI-PMH interoperability protocol is established as a standard for publication in OA since various solutions of free and open software include the implementation of this protocol to build repositories and to manage scientific publications, such as DSpace [17] and Open Journal Systems (OJS) [41]. As of 2015, more than 32,000 OJS installations have been identified, of which, 8,286 contain at least 10 articles published, for a total of 2.8 million articles available through OAI-PMH [41].

Inasmuch as the main objective of this article is knowledge discovery; it is relevant to clarify this concept. Due to their essential goal, libraries have a strategic interest in the tools and technologies that facilitate the discovery and access to resources for the communities they serve. In this field, the resource discovery systems represent the next generation of OPAC (Online Public Access Catalog) within the Integrated Library Systems (ILS), which are commonly known as web-scale discovery services [8]. In this regard, Schonfeld defines discovery as "the process and infrastructure required for a user to find an appropriate item" [46]. However, that recovery ability is focused on user-generated searches, but not in the discovery based on semantic recovery or through inference, which would give the user a more smart or enriched information retrieval. Thus, KD is approached from the computational point of view. KD is the most desirable end product of computing, Frawley et al. [23] define KD as non-trivial extraction of implicit, previously unknown, and potentially useful information, from data. Therefore, considering the previous definitions, the OntoOAIV model follows the KD approach. In accordance with the foregoing, we have divided this section into these two main sub-areas in order to compare and to position our work.

2.1 Resource Discovery Systems for Libraries

Regarding commercial solutions named as "discoverers", in the libraries context, there are EBSCO Discovery Service [18], WorldCat Discovery Service [38], Summon [43], Ex Libris Primo [21], among others. These discoverers do not perform KD as we proposed for the OntoOAIV model. Hence, in Table 1, we present a review of some libraries resource discovery products and services against the OntoOAIV's aims in order to make clear the differences with our approach. This comparison is divided into the following five criteria: Does the approach use the Semantic Web technologies? (SW column); Is the approach specialized on indexing structured resources with OAI-PMH or considering the meta-data of this protocol to gain an aggregate value? (OP column); Does the approach enable the knowledge discovery? (KD column); Does the approach consider the user context through a user profile? (UP column); Does the approach allow visualizing the outcomes of a query? (V column).

Table 1

Comparison of the resource discovery-oriented approaches for libraries and the OntoOAIV's aims

Approach	Description	Status	SW	OP	KD	UP	V
Ex Libris Primo [21]	Product (software) that allows libraries to access their collections	Active (commercial)		X			
WorldCat Discovery Service [38]	Search platform of libraries resources and external databases	Active (commercial)		X			
BLUEcloud PAC [49]	Platform for libraries services	Active (commercial)		X			
BiblioCore [6]	New generation of OPAC for libraries	Active (commercial)					
AquaBrowser [42]	Product (software) that allows libraries to access their collections	Active (commercial)					
Summon Service [43]	Web-scale discovery service for libraries and other resources of an institution	Active (commercial)		X			
Encore [27]	Resource discovery solution for libraries	Active (commercial)		X			
EBSCO Discovery Service [18]	Search platform of libraries resources and external databases	Active (commercial)		X			
Blacklight [7]	Discovery platform framework for libraries	Active (open code)		X			
VuFind [54]	Library resource portal for	Active		X			

	searching and for retrieving library's resources	(open code)					
EXtensible Catalog [22]	Next generation software for libraries	Active (open code)		X			
Franklin [53]	Discovery tool that provides access to multiple collections	Active (open code)		X			

2.2 Semantic Approaches to Information Retrieval and Discovery

In this section, some tools focused on semantic searches are introduced. A recent approach, called intelliSearch, is presented by Mehta et al. [37], introducing an implementation of a semantic web search engine based on semantic relatedness. Similarly, other projects that follow the approach of using semantic search techniques are Evi [20], AquaLog [33], and Yummly Recipe Search [55, 30]. In the same way, SemSearch [32] is based on a semantic search engine that supports complex queries in terms of multiple keywords. In addition, an application called Semantic Search, proposed by Guha and McCool [25], uses the SW data to improve a web search. It is important to mention that some SW's engines were born free but after a few years they have become commercial solutions or have been acquired by large companies, such is the case of Sindice [52] and Freebase [24]. The development of "Google Knowledge Graph" is partly based on Freebase. The "Knowledge Graph" [48], a Google search feature, was considered as a first step in building next-generation search engines.

In Table 2, we present a review of some semantic approaches to information retrieval and discovery against the OntoOAIV's aims in order to evaluate each approach and to highlight the areas of opportunities of our work. This comparison is divided into the same five criteria defined in Section 2.1.

Table 2

Comparison of approaches oriented to the Knowledge Discovery and the OntoOAIV's aims

Approach	Description	Status	SW	OP	KD	UP	V
IntelliSearch [37]	Implementation of a semantic web search engine based on semantic relatedness	Active (commercial)	X		X		
Evi [20]	Mobile application based on SW that incorporates a question-answering engine	Active (non commercial)	X		X		
AquaLog [33]	Portable question-answering system for SW	Active (non commercial)	X		X		
Yummly [55, 30]	Semantic web search engine for food, cooking, and recipes	Active (open code)	X		X	X	
Semantic Search [25]	Application that uses SW to augment the search results based	Active (private)	X		X	X	

	on traditional information retrieval						
SemSearch [32]	Semantic search engine that supports complex queries in terms of multiple keywords	Active (open code)	X	X	X		
Sindice [52]	A semantic search engine of labeled resources with RDF, microformats, microdata, and RDFa	Inactive (it is part of a commercial solution)	X		X		
Freebase [24]	A knowledge base with a platform and an API to access it	Inactive (it was acquired by Google)	X		X		
Knowledge Graph [48]	A knowledge base used by Google to improve its search engine	Active	X		X		

In summary, Table 1 and Table 2 present 21 projects: 12 for libraries and 9 for computational knowledge discovery. Regarding the library approaches, 10 of them address OAI-PMH, a key aspect of our work; however, none uses the Semantic Web technologies and provides the desired Knowledge Discovery in the computational field and in our approach. Likewise, the user's context considered as an input to the KD process and the visualization of graphs generated as an output for the user's query are not included in these projects. Conversely, accounting for the various computational approaches, all contemplate the use of semantic technologies and therefore, yield KD. However, only 2 applications encompass the user's context and even more, 1 handles structured information with OAI-PMH. Again, the graphics-based visualization is not present. As a consequence, OntoOAIV arises precisely with the idea of allowing the Knowledge Discovery and the visual analysis in structured resources under the OAI-PMH specification through the incorporation of semantic technologies and graphs.

3 Description of the Proposal

For the explanation of our proposal, we divided this section into three subsections. The first part describes the methodology and implementation of the proposed semantic approach, emphasizing the usage of the algorithm for merging ontologies. The second subsection explains the adaptation of the student's representation model. While the third part presents the incorporation of the tool for exploiting the OntoOAIV's knowledge base through a visual representation.

3.1 Semantic Approach

This semantic approach is based on the work proposed by Becerril et al. [5]. Thus, we will present a brief description of this extended proposal, named OntoOAIV, a semantic approach to context-aware resource discovery and visualization over scholarly content structured with OAI-PMH. The extended methodology of this proposal is shown in Figure 1.

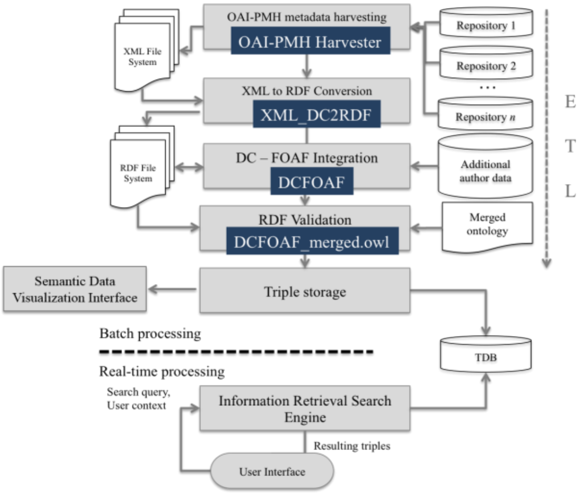


Figure 1

OntoOAIV's methodology

The first process involved in this approach is metadata harvesting of information resources available through repositories that implement the OAI-PMH protocol. OAI-PMH specifies the output information in XML file serialization format with DC; therefore, the output in XML is submitted to a transformation into RDF format. After that, those RDF files enter in the authorship information enrichment process, which is performed by following the specification of FOAF and by relating data retrieved from DBPedia, particularly from the Person-Data dataset [12]. As a result, for each harvested item, the creator tag – corresponding to the author or co-authors of the resource – is enriched. The RDF type (dc:creator) is specified as http://xmlns.com/foaf/0.1/Person, as well as the name (foaf:name), surname (foaf:givenName, foaf:surname), related person or co-author (foaf:knows), interest (foaf:topic interest), description of the creator (occupation, degree, among others) (dc:description), and birth date (onto:birthDate).

Subsequently, OntoOAIV validates the resulting knowledge using an ontology obtained from the merging of the DC and FOAF ontologies; such process is explained in Section 3.1.1. Later, the knowledge base gathered and validated is located as a repository, which can be exploited through a visualization interface (see Section 3.3). Finally, an information retrieval based on inference is performed, being this knowledge discovery the service provided to the end-user. This discovery process requires two inputs: a search query and the contextual user information. Such context is modeled through an ontological representation described in Section 3.2.

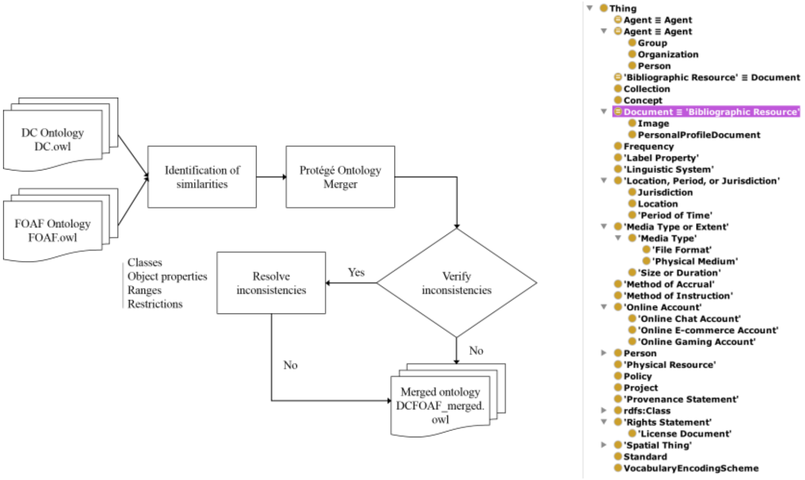
In summary, this approach is framed in the logic of the ETL (Extraction, Transformation, and Load) process, since the information is collected and organized, and then transformed into RDF, enriched and validated using ontologies, and finally loaded as triplets to a triple-store. Furthermore, this methodology is designed to add from 1 to *n* repositories, provided they fulfill the OAI-PMH specification. The technological contributions that OntoOAIV provides are a DCFOAF integrator, a DCFOAF_merged.owl ontology, an OntoOAIEstudiante.owl ontology and a visualization tool.

3.1.1 The Validation: Algorithm for Merging Ontologies

The purpose of the merging process is to generate a knowledge representation of the integration of data from the OAI-PMH repositories (DC) and other information sources, that allow enriching the data of the authors (FOAF), aiming to build a model to verify the consistency of such integration. Hence, the proposed ontological model integrates the namespace maintained by the Dublin Core Metadata Initiative (DCMI) [13], DC ontology constituted of 25 classes, and the FOAF namespace [9], ontology composed of 19 classes. Accordingly, the merged ontology can be used to model the properties of an information resource, such as a book, an article, among others, with its author or authors. For example, a researcher is an author of a publication (dc:creator) and at the same time is a person (foaf:person). Likewise, this ontology represents the authorship relation between one publication and one researcher as well as the co-authorship of one researcher with another, who shares a publication in common.

A merging of ontologies cannot be completely solved automatically due to a variety of factors. For instance, an insufficient specification of an ontology, which obstructs to find similarities with another ontology, thus, a merging is carried out manually or semi-automatically; where, a tool helps to find possible relations between items of different ontologies and an expert confirms these relations based on the ontology components' natural language description and his/her common sense. For this work, the semi-automatic approach is used since through Protégé [50] and its "Refactor>Merge Ontologies" function is carried out the merging of the two ontologies: DC and FOAF. Nevertheless, as Ameen et al. [4] mention, this automatic integration does not solve the inconsistencies generated after the process. In consequence, an adjustment is applied to the resulting ontology using the merging algorithm proposed by Ameen et al. [4]. This algorithm is illustrated in Figure 2.

Therefore, the automatic process in Protégé [50] identified (in both ontologies) and merged the Agent and rdfs:Class classes, as a result of their identical names. However, for the BibliographicResource (DC ontology) and Document (FOAF ontology) classes – equivalent class of CreativeWork –, the merging was performed manually, because there was not a coincidence in their names, even though their definitions are equivalent. In this way, the resulting ontology presents the class hierarchy shown in Figure 2.



- Thing
  - Agent ≡ Agent
  - Agent ≡ Agent
    - Group
    - Organization
    - Person
  - 'Bibliographic Resource' ≡ Document
  - Collection
  - Concept
  - Document ≡ 'Bibliographic Resource'
    - Image
    - PersonalProfileDocument
  - Frequency
  - 'Label Property'
  - 'Linguistic System'
  - 'Location, Period, or Jurisdiction'
    - Jurisdiction
    - Location
    - 'Period of Time'
  - 'Media Type or Extent'
    - 'Media Type'
      - 'File Format'
      - 'Physical Medium'
    - 'Size or Duration'
  - 'Method of Accrual'
  - 'Method of Instruction'
  - 'Online Account'
    - 'Online Chat Account'
    - 'Online E-commerce Account'
    - 'Online Gaming Account'
  - Person
  - 'Physical Resource'
  - Policy
  - Project
  - 'Provenance Statement'
  - rdfs:Class
  - 'Rights Statement'
    - 'License Document'
  - 'Spatial Thing'
  - Standard
  - VocabularyEncodingScheme

(a)

(b)

Figure 2

(a) Algorithm to merge ontologies based on the work of Ameen et al. [4] and (b) Class hierarchy of the merged ontology

Subsequently, the object properties were analyzed in order to identify semantic coincidences, as well as to verify and to match their ranges and domains. Regarding this analysis, the Creator class from DC and Maker class from FOAF were declared as similar. On the other hand, the knows and topic interest properties from FOAF also were studied due to its relevance in the description of the resources in OntoOAIV – foaf:knows represents the people with whom a relation is given, for example, a co-authorship; and foaf:topic interest describes the people's interests (authors/co-authors) –. Afterward, the data properties were addressed, where the title property was manually modified to establish a relation between Title (DC) and title (FOAF) since they present a variation in their names (capital letter). At the end, the merged ontology, called DCFOAF_merged.owl, is verified using a reasoner. This verification proves that the resulting ontology is consistent, considering that the asserted model and the inferred model are similar. Therefore, this ontology can be used to check the consistency of the RDF/XML files obtained in the harvesting, transformation, and integration phases of OntoOAIV. The metrics of DCFOAF_merged.owl are presented below: 42 Classes, 66 Object properties and 40 Data properties.

3.2 The User's Context: Student Profile Ontology

For this proposal, the user's context is conceptualized as the ability to perceive information about the user's environment. This definition arises in order to infer non-explicit facts about the user for later providing results more consistent according to such information. Therefore, it is necessary to carry out a modeling of the user's profile. According to the foregoing, one of our goals is to build, to populate, and to use an ontology related to the scientific-academic nature of the information addressed in this approach, such as publications' data, in order to deal with the user context. Since this context can include profiles of researchers, teachers, and students, the scope of this work is limited to modeling one of the profiles mentioned above: the student's profile.

An approach to modeling a student using an ontology was proposed by Panagiotopoulos et al. [39]. This ontology specifies four main classes, named “Student”, “StudentCourseInformation”, “StudentCurrentActivity”, and “StudentPersonalInformation”. A relevant aspect of this ontology is the inclusion of the student's personal information in addition to the student's academic information since this personal information provides mostly static and permanent student information. However, the ontology proposed by Panagiotopoulos et al. [39] was not available on the Web and was designed with some different features to our approach. Therefore, a new ontology is built based on the student ontology proposed by Panagiotopoulos et al. [39]. As a consequence, an ontology called OntoOAIEstudiante was developed, which represents the user's context from the information that is previously known and is defined as a static entry. It is noteworthy that our student representation does not contemplate information about interaction with the application.

Thereby, the proposed student ontology for our approach considers four classes: "Estudiante" represents any student; "Cursos" describes the subjects, the school, the program, and the program level (BA, MA, Ph.D.) at which the student is enrolled; "ActividadActual" provides information about the student current enrolled period, previous experience, course goals, modules, and period; and "InformaciónPersonal" gives information about the student's accessibility, demographics, and motivation.

The user profile represented in the "OntoOAIEstudiante" ontology is the input that provides the user's context to the OntoOAIV's inference engine.

3.3 The Visualization: Semantic Data Visualization Tool

In OntoOAIV, the incorporation of a layer for data visualization is considered because a large amount of information was gathered and integrated for our experiments, in order to provide any user a quick, simple, and easy interpretation of this information. This fact is justified by the idea expressed by Machová et al. [34]: "Information in an ontology is usually too extensive to be visualized globally in its whole complexity".

The integrated visualization approach into OntoOAIV was developed by Alvarado-Uribe et al. [3]. This work was proposed with the aim of exploiting semantic information through an individual and collective visual representation of resources, using an SPARQL endpoint. The main features of the tool are addressed below: a based keyword search engine (these terms denote the resources that constitute the knowledge base specified by the SPARQL endpoint), an SPARQL endpoint (the only and main entry to this tool), and a visual representation (a set of graphs, such as bar charts, a heat map, and a location map, as well as text).

Due to the location information (e.g. latitude and longitude) is not included in the OntoOAIV's knowledge base (described in Table 3), we modified the tool in order to obtain graphs more relevant for our goals, such as to define on the homepage as main properties to dc:type and dc:title.

4 Experiments

In order to test our approach, we have carried out three experimental phases. The first phase is for validating our knowledge base for then, exploiting it in the second and third phases through a query formulated by a user. Therefore, in Section 4.2, the knowledge discovery process is described while in Section 4.3, the visualization process is explained.

4.1 Data Collection and Knowledge Base

For validating the proposed model, we used the RUDAR [45] and Redalyc [44] repositories, which implement OAI-PMH, as our two data providers. In addition, we merged the FOAF and DC ontologies, as well as the "PersonData" dataset from DBPedia [12], in order to enrich and to validate our knowledge base. To finally obtain an enriched, validated, and stored knowledge base (in triples).

4.2 Selective Knowledge Discovery Process: Use Case

The user's profile defined and used in this experiment belongs to the student "Margret Fintz", identified by "univ:DL9510078". This student is enrolled in a "Master’s" degree program in "Pedagogy Educational Studies" at "Deutsche Universität", taking the "Humanism and Pedagogy" subject. Margret is interested in "Social apprenticeship" and "Johan Friedrich Herbart". In the same way, she specifies that "German", "English", and "Spanish" are her preferred languages, "article" and "thesis" are her preferred types of information resources, and "DE" is her demographic data. Regarding the application example is described that

Margret searches for "non-violence". Consequently, our approach provides the "Education for Active Non-Violence" thesis identified by "oai:rudar.ruc.dk:1800/2990", proposed by "Uski, Juha Janne Olavi", and published on "2008-01-17", based on the search term, as well as on Margret's profile. Since this resource was also located by coincidence with the program in which Margret is enrolled (Pedagogy Educational Studies), this result is considered as relevant for the student, which leads to the deductive reasoning process. The reasoning process takes as input the resource identifier (oai:rudar.ruc.dk:1800/2990), providing as a first deduction that Juha Janne Olavi Uski and Stephen Carney (who is specified as a contributor of this resource) are authors of this thesis; therefore, they are defined as interesting for Margret. Thus, a second deduction arises by considering the works belonging to these authors as pertinent for her. Subsequently, if these works are relevant, then the co-authors of these resources are valuable to her. Finally, another deduction is given when the works of these co-authors are determined as significant to Margret. As a result, a dataset, composed of 129 resources, represents the output of Margret's search.

4.3 Implementation of the Visualization Tool: Use Case

Because the only entry for this tool is an SPARQL endpoint, we had to generate an endpoint of the OntoOAIV's knowledge base. Hence, we used Dydra [28] to produce this SPARQL endpoint for our knowledge base; however, such repository only counts on a significant amount of resources (100), for our purposes, since the size of the original knowledge base has almost half a terabyte of information. Afterward, the generated SPARQL endpoint is provided to the tool's interface, which provides a view of the resources contained in this repository. Once this connection is established, the "Thesis" term is chosen to perform a search. In consequence, the tool provides the visualization of 100 resources of this type in both graphics and text.

5 Results and Discussion

As a first result, OntoOAIV provides a knowledge base composed of 7,917,081 facts from the academic resources structured with OAI-PMH, described in Table 3. This base contains information about 968,903 authors, 60,354 out of which were enriched. It is relevant to examine the composition of the knowledge base in order to know the data that were finally stored in the triple-store since some of them were discarded for errors found. For example, only 60,354 authors, out of the 60,927 identified, were enriched. Regarding resources, the knowledge base is composed of the 395,419 resources result from the conversion process to RDF, where only 394,776 resources have an identifier, 379,966 have a source, and

394,775 have a publication date. A curious fact is that 1,600,540 dc:subject were found, i.e., an average of 4.04 keywords per resource representing the topics that address each of them. Unfortunately, in the subject description, there is a great diversity of nomenclatures, formats, and languages used in the repositories, which can prevent to identify relationships and to have a greater semantics. On the other hand, the lack of identifiers (URIs - Universal Resource Identifiers) for resources (books, articles, thesis, authors, among others) on the Web leads to ambiguity and homonymy issues, which can produce a difficult, incomplete, and inconsistent integration process.

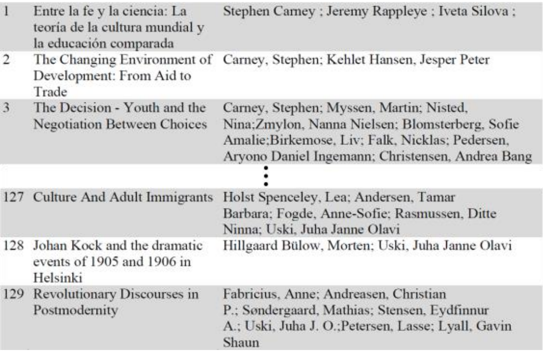
Table 3

OntoOAIV's knowledge base

Property	Total of Triples
http://xmlns.com/foaf/0.1/Person	60,354
http://purl.org/dc/elements/1.1/identifier	394,776
http://purl.org/dc/elements/1.1/source	379,966
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	60,354
http://purl.org/dc/elements/1.1/date	394,775
http://xmlns.com/foaf/0.1/name	60,354
http://purl.org/dc/elements/1.1/format	385,975
http://purl.org/dc/elements/1.1/description	361,577
http://purl.org/dc/terms/modified	394,772
http://dbpedia.org/ontology/birthDate	60,354
http://purl.org/dc/elements/1.1/type	405,463
http://purl.org/dc/elements/1.1/title	394,776
http://purl.org/dc/elements/1.1/publisher	382,412
http://purl.org/dc/terms/isPartOf	394,780
http://purl.org/dc/elements/1.1/subject	1,600,540
http://purl.org/dc/elements/1.1/rights	379,966
http://xmlns.com/foaf/0.1/surname	60,354
http://xmlns.com/foaf/0.1/givenName	60,354
http://purl.org/dc/elements/1.1/relation	381,980
http://purl.org/dc/elements/1.1/creator	968,903
http://purl.org/dc/elements/1.1/language	394,650

Regarding the results of the selective knowledge discovery process, an extract of this output in Figure 3 is provided since all of the results of this process cannot be presented for lack of space. It is noteworthy here that this process complies with our main goal: the knowledge discovery using information implemented with OAI-PMH and enriched with semantic technologies. This is verified when from 1 result is produced up to 129 outputs, all relevant to the user. One aspect that must be addressed is the update of the user's context because to in this work, this context is provided during the design phase and cannot be modified during execution.

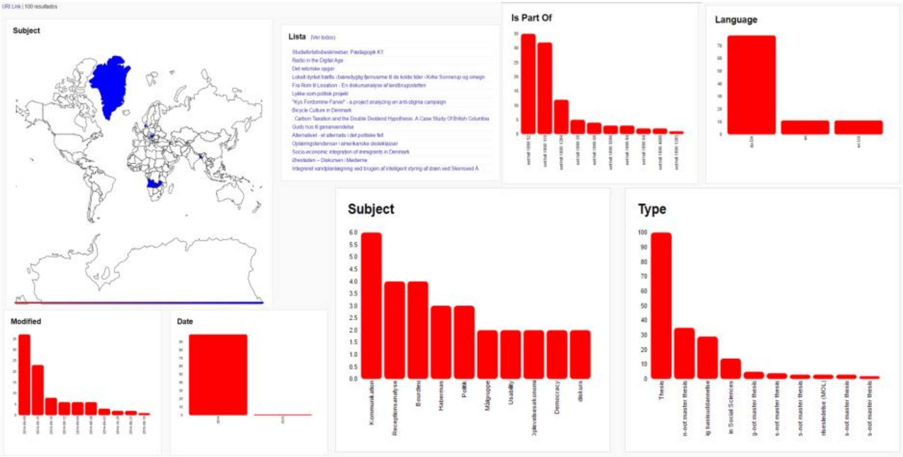
For the example of the visualization of the information contained in the OntoOAIV's knowledge base, Figure 4 is included. This visualization provides data clusters that allow a quick and easy exploration of the information. For example, it shows that all resources are Thesis published in 2014 and 2015, and written in "da-DK" (Danish (Denmark)), "en" (English), and "en US" (English (United States)). Although some results can be appreciated, the heat map reflects the lack of information as a result of only taking a sample of the original repository for this experiment. The endpoint used is https://dydra.com/joanna-au/oai-pmh-repository/sparql



1	Entre la fe y la ciencia: La teoría de la cultura mundial y la educación comparada	Stephen Carney ; Jeremy Rappleye ; Iveta Silova ;
2	The Changing Environment of Development: From Aid to Trade	Carney, Stephen; Kehlet Hansen, Jesper Peter
3	The Decision - Youth and the Negotiation Between Choices	Carney, Stephen; Myssen, Martin; Nisted, Nina;Zmylon, Nanna Nielsen; Blomsterberg, Sofie Amalie;Birkemose, Liv; Falk, Nicklas; Pedersen, Aryono Daniel Ingemann; Christensen, Andrea Bang
	⋮	
127	Culture And Adult Immigrants	Holst Spenceley, Lea; Andersen, Tamar Barbara; Fogde, Anne-Sofie; Rasmussen, Ditte Ninna; Uski, Juha Janne Olavi
128	Johan Kock and the dramatic events of 1905 and 1906 in Helsinki	Hillgaard Bülow, Morten; Uski, Juha Janne Olavi
129	Revolutionary Discourses in Postmodernity	Fabricius, Anne; Andreasen, Christian P.; Søndergaard, Mathias; Stensen, Eydfinnur A.; Uski, Juha J. O.;Petersen, Lasse; Lyall, Gavin Shaun

Figure 3

Example of the results found for the Margret's search



Studieforløbsbeskrivelser, Pædagogik K1
Radio in the Digital Age
Det retoriske opgør
Lokalt dyrket trafik i bæredygtig fjernvarme til de kolde tider i Kirke Sonnerup og omegn
Fra Rom til Lissabon - En diskursanalyse af landbrugsstøtten
Lykke som politisk projekt
"Kys Fordomme Farvel" - a project analyzing an anti-stigma campaign
Bicycle Culture in Denmark
Carbon Taxation and the Double Dividend Hypothesis: A Case Study Of British Columbia
Guds hus til genanvendelse
Alternativet - et alternativ i det politiske felt
Opløsningstendenser i amerikanske skoleklasser
Socio-economic integration of immigrants in Denmark
Drestaden – Diskursen i Medierne
Integreret vandplanlægning ved brugen af intelligent styring af åen ved Skenswad A

Figure 4

Visualization of the search performed using the "Thesis" term

This visualization also confirms that it is important to count on properties related to the location, such as latitude and longitude, aiming to populate the heat map, one of the most representative charts of the tool; otherwise, this graph would not be useful. Alternatively, because the tool depends on the endpoint provided, restrictions such as availability, maintenance and format can avoid the correct functioning of this tool. Consequently, if the endpoint's information repository is not updated, the application will not return useful information to users.

Conclusions and Future Work

In this paper, OntoOAIV is introduced aiming to verify that the incorporation of the Semantic Web technologies provides the interoperability that the Open Access platforms structured with OAI-PMH need to address for the selective knowledge discovery and in consequence, to enable students the recovering of resources not explicitly requested by them but that result potentially useful. Hence, it is proved based on the results that the incorporation of semantic technologies (algorithm for merging ontologies and ontology for providing the student's context) allows the knowledge discovery in structured information with OAI-PMH. Therefore, OntoOAIV is an approach that allows dealing with the interoperability issue presented by OAI-PMH, achieving satisfactory results in the knowledge discovery despite the issues and limitations faced (such as the validation of the merged ontology). Regarding the visualization tool, the inclusion of graphs in the query's results is an important feature that differentiates our work from those reviewed in the literature. This incorporation is relevant because the graphs allow the user to analyze large amounts of information through the simplified and understandable presentation of the data, rather than just getting a text.

OntoOAIV could be extended to take advantage of the information sources of Linked Open Data as other input for our approach. In addition, the proposal can be enhanced with the use of controlled vocabularies and/or multilingual ontologies to retrieve information in several languages. Regarding the user profile, the incorporation of a mechanism capable of learning and updating the information contained in this is established as an improvement, i.e., the inclusion of a dynamic user profile. Concerning the visualization approach, it is recommended to improve the clustering algorithm, resulting in more enriched graphs. For example, finding matches in different languages.

Acknowledgement

Joanna Alvarado-Uribe is currently supported by a CONACYT studentship.

References

[1] Allemang, D., Hendler, J.: Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. Morgan Kaufmann, 2nd edn. (May 2011)

[2] Alok Jha: Open access is the future of academic publishing, says Finch report. [Online]. Available: https://www.theguardian.com/science/2012/jun/19/open-access-academic-publishing-finch-report (current November 2016)

[3] Alvarado-Uribe, J., González-Mendoza, M., Hernández-Gress, N., Escobar-Ruiz, C.E., Hernández-Camacho, M.U.: Una herramienta visual para la búsqueda semántica rdf. Research in Computing Science 95, 9-22 (2015)

[4] Ameen, A., Rahman Khan, K.U., Rani, B.P.: Semi-automatic merging of ontologies using protégé. International Journal of Computer Applications 85(12), 35-42 (January 2014)

[5] Becerril, G.A., Lozano, E.R., Molina, E.J.M.: Enfoque semántico para el descubrimiento de recursos sensible al contexto sobre contenidos académicos estructurados con oai-pmh. Computación y Sistemas 20(1), 127-142 (2016)

[6] BiblioCommons: Bibliocore. [Online]. Available: http://www.bibliocommons.com/products/bibliocore (current November 2016)

[7] Blacklight: Blacklight. [Online]. Available: http://projectblacklight.org/ (current November 2016)

[8] Breeding, M.: The future of library resource discovery: A white paper commissioned by the niso discovery to delivery (d2d) topic committee. White paper, National Information Standards Organization (NISO), 3600 Clipper Mill Road, Suite 302. Baltimore, MD 21211 (February 2015)

[9] Brickley, D., Miller, L.: Foaf vocabulary specification 0.99. [Online]. Available: http://xmlns.com/foaf/spec/ (current September 2016)

[10] Brown, P.O., Cabell, D., Chakravarti, A., Cohen, B., Delamothe, T., Eisen, M., Grivell, L., Guédon, J.C., Hawley, R.S., Johnson, R.K., Kirschner, M.W., Lipman, D., Lutzker, A.P., Marincola, E., Roberts, R.J., Rubin, G.M., Schloegl, R., Siegel, V., So, A.D., Suber, P., Varmus, H.E., Velterop, J., Walport, M.J., Watson, L.: Bethesda statement on open access publishing. [Online]. Available: http://legacy.earlham.edu/~peters/fos/bethesda.htm (current October 2016)

[11] Chan, L., Cuplinskas, D., Eisen, M., Friend, F., Genova, Y., Guédon, J.C., Hagemann, M., Harnad, S., Johnson, R., Kupryte, R., Manna, M.L., Rév, I., Segbert, M., de Souza, S., Suber, P., Velterop, J.: Budapest open access initiative. [Online]. Available: http://www.budapestopenaccessinitiative.org/ (current October 2016)

[12] DBpedia: The dbpedia data set (2014). [Online]. Available: http://wiki.dbpedia.org/services-resources/datasets/dbpedia-data-set-2014 (current September 2016)

[13] DCMI: Dcmi metadata terms. [Online]. Available: http://dublincore.org/documents/2012/06/14/dcmi-terms/ (current September 2016)

[14] DCMI: Dublin core metadata element set, version 1.1. [Online]. Available: http://dublincore.org/documents/dces/ (current September 2016)

[15] DCMI: Metadata basics. [Online]. Available: http://dublincore.org/metadata-basics/ (current September 2016)

[16] DOAJ: Directory of open access journals (doaj). [Online]. Available: https://doaj.org/ (current October 2016)

[17] DuraSpace: Dspace. [Online]. Available: http://www.duraspace.org/ (current October 2016)

[18] EBSCO: Ebsco discovery service. [Online]. Available: https://www.ebscohost.com/discovery (current November 2016)

[19] Eprints: Registry of open access repositories. [Online]. Available: http://roar.eprints.org/view/type/ (current February 2016)

[20] Evi: Evi an amazon company. [Online]. Available: https://www.evi.com (current May 2016)

[21] Ex Libris: Primo discovery and delivery. [Online]. Available: http://www.exlibrisgroup.com/category/PrimoOverview (current November 2016)

[22] eXtensible Catalog Organization: extensible catalog. [Online]. Available: https://www.extensiblecatalog.org/ (current November 2016)

[23] Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J.: Knowledge discovery in databases: An overview. AI Magazine 13(3), 57-70 (September 1992), american Association for Artificial Intelligence

[24] Google: Freebase data dumps. [Online]. Available: https://developers.google.com/freebase/#freebase-wikidata-mappings (current May 2016)

[25] Guha, R., McCool, R.: Tap: A semantic web platform. Computer Networks 42(5), 557-577 (August 2003), Elsevier

[26] Gutierrez, C., Hurtado, C., Mendelzon, A.O.: Foundations of semantic web databases. In: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. pp. 95-106. PODS '04, ACM, New York, NY, USA (2004)

[27] Innovative Interfaces: Encore discovery solution. [Online]. Available: https://www.iii.com/products/sierra/encore (current November 2016)

[28] James Anderson and Arto Bendiken: Dydra. [Online]. Available: https://dydra.com/ (current November 2016)

[29] Kessler, C., dAquin, M., Dietze, S.: Linked data for science and education. Semantic Web 4(1), 1-2 (2013)

[30] Kessler, W.: Semantic search. [Online]. Available: http://wiltrud.hwro.de/teaching/semweb15w/supplements/4S_SemanticSearch.handout.pdf (current May 2016)

[31] Lagoze, C., Van de Sompel, H.: The open archives initiative: Building a low-barrier interoperability framework. In: Proceedings of the 1st ACM/IEEE-

CS Joint Conference on Digital Libraries. pp. 54-62. JCDL '01, ACM, New York, NY, USA (2001)

[32] Lei, Y., Uren, V., Motta, E.: Managing Knowledge in a World of Networks: 15th International Conference, EKAW 2006, Poděbrady, Czech Republic, October 2-6, 2006. Proceedings, Lecture Notes in Computer Science, vol. 4248, chap. SemSearch: A Search Engine for the Semantic Web, pp. 238-245. Springer Berlin Heidelberg (2006)

[33] Lopez, V., Pasin, M., Motta, E.: The Semantic Web: Research and Applications: Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29-June 1, 2005. Proceedings, Lecture Notes in Computer Science, vol. 3532, chap. AquaLog: An Ontology-Portable Question Answering System for the Semantic Web, pp. 546-562. Springer Berlin Heidelberg, Berlin, Heidelberg (May-June 2005)

[34] Machová, K., Vrana, J., Mach, M., Sinčák, P.: Ontology evaluation based on the visualization methods, context and summaries. Acta Polytechnica Hungarica 13(4), 53-76 (2016), BUDAPEST TECH BECSI UT 96-B, BUDAPEST, H-1034, HUNGARY

[35] Martinez-Villaseñor, M.d.L., Gonzalez-Mendoza, M., Hernandez-Gress, N.: Towards a ubiquitous user model for profile sharing and reuse. Sensors 12(10), 13249-13283 (2012)

[36] Max-Planck-Gesellschaft: Berlin declaration on open access to knowledge in the sciences and humanities. [Online]. Available: https://openaccess.mpg.de/Berlin-Declaration (current October 2016)

[37] Mehta, A., Makkar, P., Palande, S., Wankhede, S.B.: Semantic web search engine. International Journal of Engineering Research and Technology 4(4), 687-691 (April 2015)

[38] OCLC: Worldcat discovery. [Online]. Available: https://www.oclc.org/worldcat-discovery.en.html (current November 2016)

[39] Panagiotopoulos, I., Kalou, A., Pierrakeas, C., Kameas, A.: Artificial Intelligence Applications and Innovations: 8th IFIP WG 12.5 International Conference, AIAI 2012, Halkidiki, Greece, September 27-30, 2012, Proceedings, Part I, IFIP Advances in Information and Communication Technology, vol. 381, chap. An Ontology-Based Model for Student Representation in Intelligent Tutoring Systems for Distance Learning, pp. 296-305. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)

[40] Patel-Schneider, P.F., Horrocks, I.: Position paper: A comparison of two modelling paradigms in the semantic web. In: Proceedings of the 15th International Conference on World Wide Web. pp. 3-12. WWW '06, ACM, New York, NY, USA (2006)

[41] PKP: Open journal systems. [Online]. Available: https://pkp.sfu.ca/ojs/ (current October 2016)

[42] ProQuest: Aquabrowser. [Online]. Available: http://www.proquest.com/products-services/AquaBrowser.html (current November 2016)

[43] ProQuest: The summon service. [Online]. Available: http://www.proquest.com/products-services/The-Summon-Service.html (current November 2016)

[44] Redalyc: Sistema de información científica redalyc - red de revistas científicas de américa latina y el caribe, españa y portugal. [Online]. Available: http://www.redalyc.org (current May 2016)

[45] RUDAR: Rudar - roskilde university digital archive. [Online]. Available: http://rudar.ruc.dk/ (current May 2016)

[46] Schonfeld, R.C.: Does discovery still happen in the library? roles and strategies for a shifting reality. Report, Ithaka S+R, New York, NY (2014), http://www.sr.ithaka.org/wp-content/mig/files/SR_Briefing_Discovery_20140924_0.pdf

[47] Shadbolt, N., Hall, W., Berners-Lee, T.: The semantic web revisited. IEEE Intelligent Systems 21(3), 96-101 (May 2006), iEEE

[48] Singhal, A.: Introducing the knowledge graph: things, not strings. [Online]. Available: https://googleblog.blogspot.mx/2012/05/introducing-knowledge-graph-things-not.html (current May 2016)

[49] SirsiDynix: Bluecloud pac. [Online]. Available: http://www.sirsidynix.com/products/bluecloud-pac (current November 2016)

[50] Stanford University: Protégé. [Online]. Available: http://protege.stanford.edu/ (current November 2016)

[51] Subirats-Coll, I.: Seven things you should know about linked data. COAR Repository Observatory (2) (2014)

[52] Tummarello, G., Delbru, R., Oren, E.: The SemanticWeb: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings, Lecture Notes in Computer Science, vol. 4825, chap. Sindice.com: Weaving the Open Linked Data, pp. 552-565. Springer Berlin Heidelberg (2007)

[53] University of Pennsylvania: Franklin. [Online]. Available: http://franklin.library.upenn.edu/index.html (current November 2016)

[54] Villanova University's Falvey Memorial Library: VuFind. [Online]. Available: http://vufind-org.github.io/vufind/ (current November 2016)

[55] Yummly: Yummly api documentation. [Online]. Available: https://developer.yummly.com/documentation (current May 2016)