The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0001-253X.htm

Organizing open archives via lightweight ontologies to facilitate the use of heterogeneous collections

J. Alfredo Sánchez
Universidad de las Américas Puebla, San Andrés Choulula Puebla, Puebla, Mexico
María Auxilio Medina
Universidad Politécnica de Puebla, Cholula, Mexico
Oleg Starostenko
Universidad de las Américas Puebla, Cholula, Mexico, and
Antonio Benitez and Eduardo López Domínguez
Universidad Politécnica de Puebla, Cholula, Mexico

Abstract

Purpose – This paper seeks to focus on the problems of integrating information from open, distributed scholarly collections, and on the opportunities these collections represent for research communities in developing countries. The paper aims to introduce OntOAIr, a semi-automatic method for constructing lightweight ontologies of documents in repositories such as those provided by the Open Archives Initiative (OAI).

Design/methodology/approach – OntOAIr uses simplified document representations, a clustering algorithm, and ontological engineering techniques.

Findings – The paper presents experimental results of the potential positive impact of ontologies and specifically of OntOAIr on the use of collections provided by OAI.

Research limitations/implications – By applying OntOAIr, scholars who frequently spend many hours organizing OAI information spaces will obtain support that will allow them to speed up the entire research cycle and, expectedly, participate more fully in global research communities.

Originality/value – The proposed method allows human and software agents to organize and retrieve groups of documents from multiple collections. Applications of OntOAIr include enhanced document retrieval. In this paper, the authors focus particularly on document retrieval applications.

Keywords Information integration, Ontologies, Open archives, Distributed collections, Clustering, Information management, archives

Paper type Research paper

Aslib Proceedings: New Information Perspectives
Vol. 64 No. 1, 2012
pp. 46-66

0001-253X
DOI 10.1108/00012531211196701

The work reported here has received support from CONACYT under Grant No. W35804A. The authors gratefully acknowledge the participation of Metztli Ibañez Rivera, Ricardo Omar Chávez García and Adriana Gabriela Ramírez de la Rosa in various phases of the project. Initial support was also received from the Ibero-American Program for Science and Technology under Grant No. CYTED VII.19 RIBIDI.

1. Introduction

Open archives provide opportunities to promote research and knowledge dissemination throughout the world. They are of particular significance for individuals, organizations and countries with limited or no access to commercially available collections, as open archives may support scholarly work, and also provide mechanisms to expose research results via open repositories.

Availability, however, does not automatically imply accessibility or ease of integration, organization, and use. Scientists frequently need to spend many hours organizing the information space made up by the highly heterogeneous distributed collections such as those offered by the Open Archives Initiative (OAI)[1]. In this paper, we propose the construction of ontologies to facilitate the work of organizing and using documents offered by the OAI community.

Ontologies provide a shared understanding of a domain. In the field of digital libraries, ontologies can support tasks such as resource description, information integration, interoperability, searching, and browsing. Since acquiring the knowledge necessary to construct ontologies is a costly task that requires much time and many resources, ontology-learning methods have been developed for this purpose.

We present the OntOAIr method (Ontologies from Open Archives Initiative Repositories to support Information Retrieval), an ontology-learning method for the construction of ontologies. The OntOAIr method uses simplified representations of documents, an adaptation of an existing clustering algorithm, and ontological engineering techniques.

In the context of information systems, the literature presents varying definitions of ontologies. One of the most quoted definitions establishes that an ontology is an explicit specification of a conceptualization (Gruber, 1993). An extended version of this definition suggests that the conceptualization must be shared (Borst, 1997).

The literature distinguishes lightweight from heavyweight ontologies according to the degree of formality involved in their encoding (Lassila and McGuinness, 2001. The scope of our work is limited to lightweight ontologies. They range from an enumeration of terms to a graph or taxonomy of concepts with well-defined relationships among them, which provide a representation of an information space. The term lightweight indicates that the construction of ontologies does not involve domain experts. It also refers to tree-like structures where each node label is a language-independent propositional formula (Giunchiglia *et al.*, 2006).

In lightweight ontologies, there is not a strong distinction between terms and concepts. Terms make controlled vocabularies available for the classification of content and they are use-dependent. Some web search engines use lightweight ontologies (Gómez *et al.*, 2004; Plisson *et al.*, 2005).

Ontologies that we propose are aimed to create a data model able to:

- provide a shared terminology for accessing data providers that human and software agents can understand and use;
- define the meaning of each term of the model in an unambiguous manner;
- implement the semantics of the data model in a machine-accessible way; and
- index data providers to support information retrieval.

The OntOAIr method allows human and software agents to organize and retrieve groups of documents from multiple collections. This method uses simplified representations of documents, an adaptation of the Frequent Itemset-based Hierarchical Clustering algorithm (FIHC), and ontological engineering techniques. Schemas and namespaces are defined to formalize the constructed ontologies, and different levels of expressivity are explored by means of mappings using the XML, RDF – Schema and OWL-DL languages.

Our method to construct ontologies includes the following tasks: harvesting, representation, clustering, and formalization. The harvesting task obtains the documents from the digital collections. The representation task constructs a vector representation for each harvested document. The clustering task applies an exclusive hierarchical algorithm to the vectors to produce a tree of clusters. Finally, the formalization task transforms the tree of clusters into a lightweight ontology.

OntOAIr allows human and software agents to organize and retrieve information from collections. In order to demonstrate the potential of the proposed method, we are using similar collections to those provided by the Open Archives Initiative (OAI) as a test bed, with encouraging results.

The remainder of the document is organized as follows. Section 2 briefly presents fundamentals of Open Archives Initiative (OAI), which is central to our work from the perspective of contents. Then, section 3 discusses related work. Section 4 presents the OntOAIr method for constructing lightweight ontologies. Next, section 5 proposes a keyword-based information retrieval model as an application of the OntOAIr method. Section 6 describes a prototypical system that implements the proposed model whereas Section 7 describes the evaluation of OntOAIr method. Section 8 refers to supporting digital libraries in developing countries. Finally, section 9 includes conclusions and suggests future directions of our work.

2. The Open Archives Initiative

The Open Archives Initiative (OAI) develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. There are two types of participants in OAI: data providers that expose metadata of their resources in semi-structured documents termed records, and service providers that use records to offer value-added services (Lagoze and van de Sompel, 2001). At the time of this writing, there are over 1,200 data providers, which include universities, research institutes, or libraries.

2.1 Interoperability in OAI

OAI proposes a low level mechanism referred to as the Protocol for Metadata Harvesting (OAI-PMH) to support application-independent interoperability. This protocol provides external online access to records. Our work is based on version 2.0 of the protocol, which is the most current and uses Dublin Core (DC) as its default metadata standard. The protocol defines request verbs, which are utilized by service providers to harvest records. However, this protocol does not offer mechanisms for retrieving actual content, that is, request verbs do not implement similarity measures to harvest only relevant records. As a consequence a harvesting task implies the retrieval of either a specific record or all the records from a data provider. Thus, typically records are harvested, selected, and organized before determining their relevance.

2.2 OAI retrieval tools

The typical ways to find relevant records on OAI-compliant data providers are search engines and service providers that implement retrieval mechanisms. Some members of the OAI community have developed tools to retrieve information from data providers, the results of which are presented as lists of relevant records. With respect to our work, record lists exhibit two general drawbacks: First, these lists cannot satisfy the information needs of all users. Instead of lists of records, some users would be more interested in obtaining a view that enables them to explore data providers through the construction of groups of similar records. The second drawback is that most search engines and service providers search on a single data provider. Thus, when multiple data providers are involved, users must initiate several requests.

In our work, we posit that ontologies can help overcome the drawbacks of lists of independent records and the retrieval from a single data provider by providing a mechanism to organize result sets. We provide briefly an overview of this area in the following section.

3. Related work

Document clustering and ontologies have been very useful in supporting information retrieval tasks. This section presents salient contributions that aim at the utilization of data collections available to people and systems into some form of ontological knowledge representation.

3.1 Document clustering in OAI

Document clustering establishes a similarity function over documents. It automatically groups documents such that documents within a cluster have a high internal similarity whereas documents in different clusters are dissimilar (Halgamuge and Wang, 2005).

There are different methods to construct clusters, such as probabilistic clustering, clustering based on association rules, or clustering algorithms based on item sets. Some of the measures commonly used to choose an appropriate method include complexity, scalability, clustering precision, clustering recall, and clustering speed (Guojun *et al.*, 2007).

We are interested on unsupervised hierarchical clustering algorithms with high scalability and good clustering precision, which can be applied to organizing records into ontologies that facilitate browsing and exploiting collections. Previous works on reference collections have shown that hierarchical algorithms produce high quality clusters (Halgamuge and Wang, 2005; Xu and Wunsch, 2008).

(Brase and Nejdl, 2004) use DC metadata and ontologies to classify learning resources. Keywords are extracted from the subject element and they are compared with the ACM Computer Classification system. Specialized ontologies are used to identify topics and subtopics. Ontologies can be used in the work of (Brase and Nejdl, 2004) because a specific domain is considered. In general, however, records from data providers describe resources in several domains. OAI-PMH specifications do not require controlled vocabularies to reduce ambiguity of the terms found in DC elements. This lack of control in the selection of terms may cause inconsistencies in any simplified representation of a document. Instead of the DC attribute (which may be assigned to zero or multiple values), the extraction of the keywords in our work's retrieval models only use the title and description elements, as they refer to valuable

information about the topic of a resource. As it often occurs, we assume that the terms in the title also appear in the abstract section.

(Harrison *et al.*, 2004) describe a general purpose system called OAI-PMH aggregator which measures the cosine similarity between records from a data provider using the Vector Space Model (VSM) (Hamel, 2009). Ranking values range between 0 and 1, where 1 represents the maximal similarity. The results of similarity calculations are stored in an optional container of the harvested record.

The method proposed in (Harrison *et al.*, 2004) is based on a model that has proved to be very successful in several contexts. However, from our point of view, it has three main drawbacks:

(1) The computation of similarity is performed every time a query is introduced, which may have a significant impact on performance.

(2) The original records are modified by the inclusion of an element to hold similarity calculations, which goes against standard compliance.

(3) The method does not provide a general view of the content offered by data providers.

3.2 Frequent itemset-based hierarchical clustering

Frequent Itemset-based Hierarchical Clustering (abbreviated FIHC) is an agglomerative clustering algorithm proposed by (Fung *et al.*, 2006). This algorithm is based on the hypothesis that if a group of documents refers to the same topic, they should share a set of terms called frequent item *sets*.

FIHC uses feature vectors and frequent item sets to produce a hierarchical structure of non-overlapping clusters. This algorithm requires two mandatory input parameters, which are referred to as global support (the percentage of documents in a collection that contains a frequent item set), and cluster support (the percentage of documents in a cluster that contains a frequent item set). An optional parameter can be added to force FIHC algorithm to produce a fixed number of clusters.

We conducted several experiments on various data sets and, as a result, we proposed some modifications to FIHC in our implementation. In summary, the pruning processes (tree and child pruning), as well as the sibling merging process were removed from the construction of the tree of clusters. The constructed tree has the following characteristics:

(1) Clusters in the k-th level of the tree are identified by k-terms labels.

(2) Labels are formed with the most representative terms of feature vectors according to the score function.

(3) All the feature vectors of a cluster contain the terms of their cluster label.

(4) Feature vectors under the score function form a special cluster whose label is null. This cluster is a direct descendent of the root.

The FIHC algorithm has shown higher precision, more efficiency and greater scalability than the Unweighted Pair Group Method with Arithmetic (UPGMA), k-means and other clustering algorithms (Fung *et al.*, 2006). Other advantage of the FIHC algorithm is its low number of database passes made when searching the space.

3.3 Incorporation of ontologies into keyword-based information retrieval systems

Ontologies have been considered an alternative to overcome the drawbacks of keyword-based search engines. They have been combined with traditional information retrieval models to improve search effectiveness. For instance describes a portal that provides semantic data retrieval. Search requests return ontology instances or links to documents that reference instances. KeyConcept is another work that makes use of ontologies.

Some experiments have shown an improvement in search precision when semantics is used in addition to keywords for retrieval (Aitken and Reid, 2000; Cui and O'Brien, 2000). Experiments have shown that the ontology-based matching produce greater precision than the keyword matching. For example, in the DOME project (Cui and O'Brien, 2000) ontologies enable semantic interoperability between a distributed search engine and data providers aiming at user-transparent query solving via retrieving and integrating information. A shared ontology is maintained for all the available data providers. In this project, the construction and mapping of the shared ontology are semi-automated processes.

However, a new-shared ontology needs to be developed to integrate new data providers. As a result, reusability and flexibility are reduced. In contrast, our work proposes the construction of an ontology for each data provider. The construction process requires human intervention solely for the purpose of establishing the value of three input parameters, as explained in Section 3.2. In this approach, clustering algorithms are used to group items that may be semantically close to each other, but it is the user who decides whether the proposed groups are meaningful.

3.4 Ontology learning methods

At the time of this writing, a completely automatic construction of ontologies has not been reported in the literature. However, there are several contributions in this direction, which are discussed next.

We are particularly interested on methods for learning ontologies from texts. For example, OntoLearn (Navigli and Velardi, 2004) is an ontology learning method that applies a hierarchical algorithm to a set of documents from dedicated web sites and document warehouses. OntoLearn uses the output of the algorithm as well as WordNet to construct domain ontologies. This method proposes a semantic interpretation of terms, in such a way that the constructed ontologies describe concepts with terms composed by two or more words. From our perspective, however, the main drawback of OntoLearn is that it assumes the existence of previously classified documents.

(Diederich and Balke, 2007) propose the semantic grow-bag approach to create lightweight topic, categorization systems. The approach uses the keywords provided by authors of digital objects to compute a co-occurrence metric, finds relations between keywords based on the co-occurrence metric, and constructs graphs that represent the neighborhood of the keywords. The approach is computed off-line, and re-run periodically, to update the graphs, according to the added author keywords.

Table I summarizes the characteristics of some salient ontology learning methods. The comparison framework takes into account whether the documents are accessible via OAI-PMH, the domain, and type of these documents, the basis of the method used to construct ontologies. In the second column, the value "compliant" means that the work uses OAI-PMH records as documents. Otherwise, the value "non compliant" is shown. In

this case, documents do not have a common structure, that is, they contain free text. In the third column, the value "specific" means that the documents belong to a specific domain such as business or medicine. If this is not the case, the value "general" is used. The fourth column includes the type of the documents. In the fifth column, the method used to cluster documents is presented. In the fifth column, the value "clustering" is followed of the name of an algorithm; the value "conceptual clustering" means that keywords of document clusters are compared with concepts of the WordNet ontology.

The ontology learning methods described in Table I implement a pre-processing step that consists of the application of natural language processing techniques to complete documents. As a result of this step, simple representations of documents are obtained. Then, a clustering algorithm makes use of these representations to propose a meaningful organization of documents. In contrast, the OntOAIr method presented in the following section takes advantage of the explicit structure of metadata records and only applies the pre-processing step to some of their elements. As a consequence, fewer data are taken into account by the clustering algorithm to organize metadata.

4. The OntOAIr method

This section describes the OntOAIr method (Ontologies from Open Archives Initiative Repositories to Support Information Retrieval), a method that we first proposed in (Medina and Sánchez, 2008) to semi-automatically construct ontologies from data providers called ontologies of records. These are hierarchical structures formed by disjoint groups of records. Ontologies of records have two main uses:

(1) To organize records based on their content.

(2) To support information retrieval tasks from multiple data providers.

The process of constructing ontologies of records consists of four main tasks: harvesting, representation, clustering, and formalization. The harvesting task obtains the documents from the digital collections. The representation task constructs a vector representation for each harvested document. The clustering task applies an exclusive hierarchical algorithm to the vectors to produce a tree of clusters. The formalization task transforms the tree of clusters into a lightweight ontology.

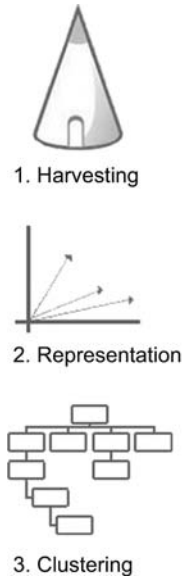
Figure 1 shows the sequence of these tasks, which are described with further detail in the following sections.

4.1 Harvesting

Harvesting uses the request verbs to obtain metadata records from data providers. In our work, harvesting has been delegated to software entities called harvester agents.

Table I.
Main characteristics of some ontology learning methods

Work	OAI-PMH	Domain	Documents	Method
Diederich and Balke, 2007	Not compliant	General	Free text	Clustering (grow-bag)
Karouri *et al.*, 2006	Not compliant	Specific	Web pages	Clustering (*k*-means)
Ljubič *et al.*, 2005	Not compliant	Specific	Free text	Clustering (*bisecting*)
Plisson *et al.*, 2005	Not compliant	Specific	Free text	Clustering (*k*-means)
Navigli and Velardi, 2004	Not compliant	Specific	Web sites, warehouses	Conceptual clustering, Wordnet




```
<?xml version="1.0" Encoding="UTF8"?>
<ontologyOfRecords>
<algorithm
        globalsupport="25"
        clustersupport="30">
        <name>FIHC</name>
</algorithm>
<cluster>
        <label>physics</label>
        <level>1</level>
        <record>
                <title>General Physics</title>
        ...
        </record>
        ...
        </cluster>
...
</ontologyOfRecords>
```

4. Formalization

Figure 1.
Main tasks to construct ontologies of records

This notion was first discussed in our early work (Medina *et al.*, 2004), which included a detailed description of these agents.

Harvester agents are modeled with Gaia, a general-purpose methodology that supports the analysis and design of multi-agent systems.

4.2 Representation

Once records from data providers have been harvested, simplified representations of records are produced. In our work, content information of records is extracted from the DC elements title and description.

DC elements contain free text. Before processing, any case sensitivity in the data are removed. Then, feature vectors are generated. A feature vector is a simplified representation of a record formed by keywords (all terms other than stop-words extracted from title and description) and weights (numeric values that represent the relevance of keywords). The name is adopted from (Fung *et al.*, 2006).

Our work adds two new elements to a feature vector: an identifier to discriminate between records and a URL with the location of its data provider. Both elements uniquely identify a record in OAI; they do not add content information.

4.3 Clustering: adaptation of FIHC

The selection of FIHC algorithm for document clustering is based on the following assumption: "the terms of cluster labels can be used to construct a vocabulary that describes the main topics of a collection".

We implemented FIHC algorithm as proposed by (Fung *et al.*, 2006), however, we experimentally determined that pruning child and sibling clusters does not achieve a better clustering and they were removed from the construction of the tree of clusters.

Child pruning is addressed to reduce the width of the tree of clusters until a user-specified number of clusters is reached. In our experience, this process does not

contribute to improve the cluster accuracy since the number of main clusters in collections is unknown.

Sibling pruning is addressed to reduce the depth of the tree of clusters considering only the non-leaf nodes at level two or greater. Since the F and entropy measures take the nodes at level one into account, under these measures, clustering accuracy would not be affected by the elimination of this process.

Therefore, we propose an adaptation of FIHC algorithm where the child pruning and the sibling pruning are removed from the construction of the tree of clusters.

4.4 Formalization

The formalization task refers to the transformation of the tree of clusters into an ontology of records represented in a machine-accessible language. We have used XML, RDF-Schema, and OWL-DL representations to explore different levels of expressivity. Some common characteristics of these representations are the following:

- at least one cluster is needed;
- the record element represents a resource in the OAI initiative; and
- each cluster has a label, a level and zero or more records.

Representing an ontology of records in XML inherits all the advantages of the language itself such as its simplicity and readability. However, there are also some shortcomings. For example, there is no standard way to assign meaning to nested tags, thus leading to ambiguity when interpreting XML documents. From an ontological point of view, the main drawback is that the semantics of XML documents is accessible to humans but not to machines.

In order to add machine-accessible semantics to ontologies of records, we have extended XML representation to get a Resource Description Framework Schema (RDF Schema) implementation. RDF-Schema is a semantic extension of Resource Description Framework (RDF), which is a framework to depict web metadata (Gómez *et al.*, 2004). By using RDF Schema, it is possible to define vocabularies in RDF documents, specify properties and restrict its range to model assumptions about any particular application domain. Thus, RDF Schema offers a unique interpretation of ontologies of records. In RDF Schema, properties are defined globally, that is, they are applied to all classes. However, the expressiveness of RDF Schema is limited mainly because only binary relationships can be represented. Thus it is considered a primitive ontology language. To extend the expressivity of RDF Schema, we propose the use of the Web Ontology Language Description Logic (OWL-DL) to take advantage of tools that enables efficient reasoning about the knowledge represented in the ontologies of records and to maintain compatibility with RDF ontologies. Part of the knowledge about an ontology of records that can be represented in OWL-DL is described in the following statements:

(1) An ontology of records is formed by clusters (a cluster is part of an ontology of records).

(2) A cluster contains records (a record is contained in a cluster).

(3) A cluster is described by a label (a label describes a cluster).

(4) A cluster has a level in the ontology.

(5) A record has title, subject, description, identifier, URL, DataProvider, MetadataFormat, and Datestamp elements. These elements are OWL-DL classes.

(6) The properties isFormedBy and contains are transitive.

In the previous list, the words in italics are names of properties, and sentences in parenthesis represent the inverse of a property. To verify the feasibility of the OntOAIr method, we propose the use of ontologies to support a keyword-based retrieval model. The next section describes this model.

5. Keyword-based retrieval model

The keyword-based model retrieves clusters of records from XML ontologies. Records are treated as documents, queries as specifications of clusters of records, and ontologies as the structures that organize records of data providers. In this model searching relevant clusters is conducted via matching operations between queries and cluster labels. Natural language queries are converted to keyword-based queries. Query keywords are extracted after a stopword elimination process. The stopword list used, is formed by intersecting the stopword lists, of the CACM, and Time collections[2]. Keywords are case-insensitive.

Figure 2 shows an overview of the keyword-based retrieval model. At this point, it is assumed that an ontology of records of a data provider has been constructed. The user introduces a query (request) through a web interface. The result is obtained after matching operations between query keywords and cluster labels; this includes links to access records.

Mapping from queries to ontologies of records is hidden to the users. The search process on the ontology of a data provider comprises the selection of potential clusters (clusters for which a similarity measure between query keywords and the cluster label is equal or greater than a threshold).

The similarity measure is only applied to cluster labels instead of all records. Intuitively, labels with more terms represent more specific topics than labels of ancestor clusters.

A retrieval task using multiple collections requires handling of multiple ontologies. For this purpose, a list of potential clusters extracted of each ontology is used.

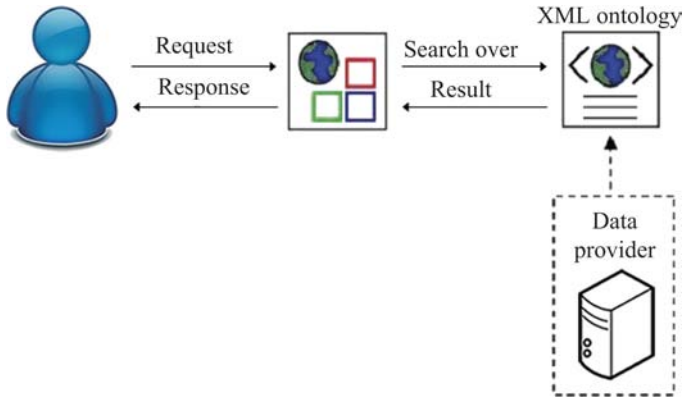


Figure 2.
Elements of the keyword-based model applied to a data provider

6. OntoSIR: a prototypical implementation of OntOAIr

The OntoSIR system (Ontology-based Information Retrieval System) is a demonstration prototype system of the method to construct ontologies of records (section 4). It implements the information retrieval model presented in section 5.

The main goal of implementing OntoSIR is to provide users with an environment to construct ontologies of records, and an interface to explore and retrieve information from multiple collections. Figure 3 shows a general case of use of OntoSIR. In this scenario, a user selects data providers and introduces a query through a web-based interface. Then, OntoSIR harvests metadata records from these data providers by using OAI-PMH request verbs and automatically analyzes records. Harvested records, represented as feature vectors and the FIHC algorithm are used to construct an ontology of records for each data provider.

OntoSIR provides a query-driven recall mechanism for the constructed ontologies, which are centrally maintained.

There are two interaction modes between users and the keyword-based model: synchronous and asynchronous. The former is used when the required ontologies have been constructed. Then, the retrieval tasks are performed on line. In contrast, the latter uses e-mail to send the response.

The graphical user interface of OntoSIR allows users to carry out the following tasks:

- searching relevant groups of records from multiple data providers; and
- clustering the records from a data provider.

The results of these tasks are returned as XML files, in such a way that they can be used by parsers, and other XML processing software.

Figure 4 shows the interface of OntoSIR when the search option is selected from the menu (left panel). The right panel contains a selection box and text fields through which data providers and the query can be entered.

OntoSIR is a web-accessible system implemented in Java that requires a servlet container. Apache Tomcat 5.0.29 is used in this capacity. The current version of OntoSIR employs MySQL 5.0.4 as its database management system. The next section discusses experimental results of the OntOAIr method.

7. Evaluation

Assuming that a set of records has been harvested, experiments were conducted to test the feasibility of the OntOAIr method and the implementation of the retrieval model.

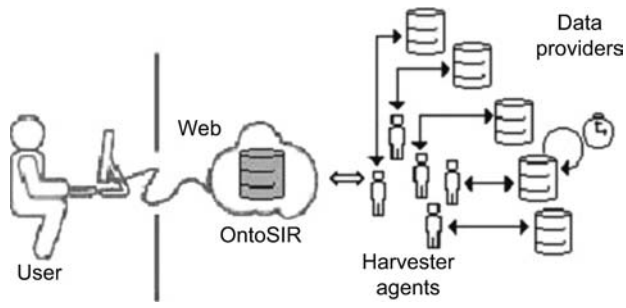


Figure 3.
A general use case of OntoSIR

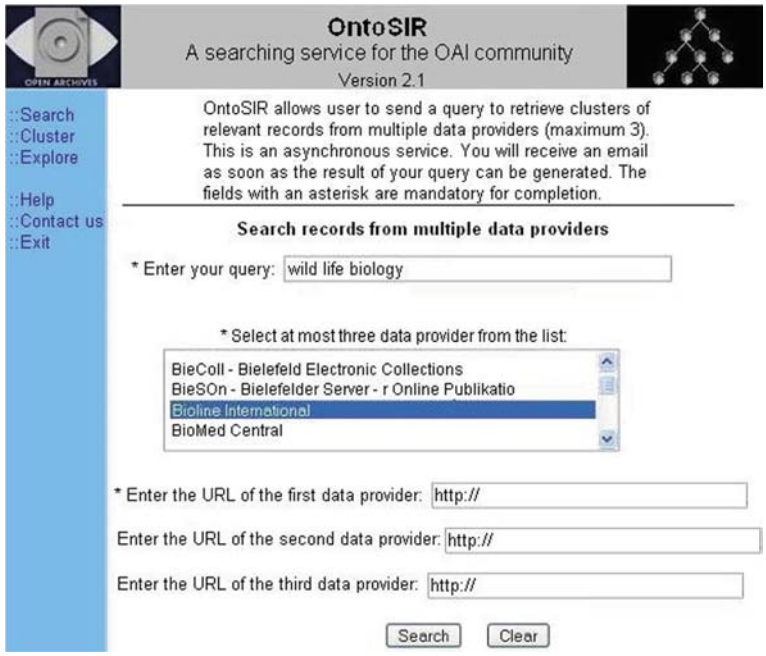


Figure 4.
Graphical user interface of OntoSIR to search records

The evaluation is focused on representative situations that are plausible in actual information retrieval contexts.

7.1 Clustering evaluation

Clustering evaluation is aimed at quantifying the goodness of a clustering algorithm, which may be judged differently depending on which measure is used (Berry and Castellanos, 2010). This section presents the comparison of our adaptation of the FIHC algorithm with other widely used clustering algorithms: UPGMA (Jain and Dubes, 1988; Lambert *et al.*, 2010) and bisecting *k*-means (Jain and Dubes, 1988; Lambert *et al.*, 2010; Berry and Castellanos, 2010).

In our comparison, we use the vector space model with TF*IDF-method to represent the texts and the cosine measure for calculating similarity between text and clusters. DocCluster (doccluster, 2007) is used as the implementation of FIHC algorithm, while Cluto produces the result of UPGMA and bisecting *k*-means.

7.1.1 Clustering measures. The evaluation method uses two clustering measures: overall *F* measure and entropy. *F* measure estimates the accuracy of the produced clustering solutions taking reference collections into account. It is assumed that the documents of reference collections belong to a single class or topic called natural class. Entropy provides a measure of goodness for non-nested clusters or for the clusters at one given level of a hierarchical clustering (Berry and Castellanos, 2010). By using the *F* measure, each cluster is treated as the result of a query and each natural class as the relevant set of documents for the query.

7.1.2 Data sets. Our work uses the data sets of Table II, these are heterogeneous in terms of document size, cluster size, number of classes and documents distribution.

Classic4 is formed by the CACM, CISI, CRAN and MED abstracts[3]. Hitech and Wap are from the San Jose Mercury newspaper articles[4] and the Yahoo! subject hierarchy web pages[5], respectively. Reuters and Re0 are extracted from newspaper articles (Lewis, 1997). For Reuters, only the articles that are uniquely assigned to a natural class are used. In all of the data sets, stop words have been removed.

7.1.3 Results. Table III shows the *F* measure values with different user specified number of clusters, where a dash indicates that the algorithm is not scalable to run. For our implementation of FIHC algorithm, a cluster support of 25 per cent and a global support of 5 per cent are used. A minimal support of 3 per cent in the extraction of frequent item set is used. The values of these parameters were experimentally determined.

The main disadvantage of UPGMA is that it is not scalable for large data sets. Bisecting *k*-means and FIHC are scalable.

FIHC is robust enough to produce consistently high quality clusters in many cases. Intuitively, each class in FIHC has a "core" vocabulary that acts as a simple disambiguation mechanism. However, these core vocabularies may overlap. Table IV shows the total entropy. The number of desired clusters was introduced as another input parameter. Bisecting *k*-means and FIHC are the best algorithms as they have similar behavior with respect to entropy, whereas UPGMA does poorly. Thus, we conclude that the clusters provided by FIHC are useful. They could be used as

Table II.
Summary description of data sets

Data Set	Number of documents	Number of classes	Class size	Number of words
Classic4	7,094	4	1,033-3,203	12,009
Hitech	2,301	6	116-603	13,170
Re0	1,504	13	11-608	2,886
Reuters	8,649	65	1-3725	16,641
Wap	1,560	20	5-341	8,460

Table III.
Overall F measure comparison

Data set (number of classes)	Number of clusters	FIHC	UPGMA	Bisecting *k*-means
Classic4 (four)	15	0.50	–	0.41
	30	0.51	–	0.45
	60	0.49	–	0.29
Hitech (six)	15	0.40	0.37	0.43
	30	0.39	0.51	0.31
	60	0.39	0.49	0.24
Re0	15	0.43	0.53	0.37
	30	0.41	0.47	0.37
	60	0.38	0.38	0.30
Reuters	15	0.59	–	0.44
	30	0.58	–	0.37
	60	0.6	–	0.33
Wap	15	0.54	0.59	0.55
	30	0.53	0.58	0.46
	60	0.52	0.57	0.39

reasonable aid for classification in OAI contexts and might even provide new insights into what collections are about.

7.2 Task-based evaluation

In absence of a commonly agreed-on schema for analyzing the properties of an ontology, a common way to proceed is to evaluate it within some existing application. This section describes the competency of the ontologies of records to support information retrieval. The test bed is the OntoSIR system. Hereafter, the XML implementation of the ontologies is used.

The OntOAIr method is compared against the Vector Space Model (VSM) (Hamel, 2009). Precision and recall measure effectiveness. The implementation of the VSM, which is used in the experiments is provided by the Apache Lucene search engine (McCandless *et al.*, 2010).

At the time of this writing, OAI does not suggest a standard test collection, thus we chose the CACM collection[6] because it can be regarded as an OAI-compliant data provider. This is a collection of titles and abstracts from the CACM magazine. CACM collection consists of 3,204 records; each record includes information about the author, title, abstract (manually assigned) keywords and information about cites.

The CACM collection includes queries formed by experts and their relevance judgments. In our work, keywords are extracted from the titles and the abstracts. In the experiments, all the query keywords were given a weight of 1.

Table V shows the results of the comparison between the OntOAIr method and VSM. It includes averaged values of precision to standard values of recall.

Table V shows an improvement for every standard level of recall (average percent 17.8 per cent). Precision is interpolated to standard levels of recall and averaged on the number of queries (15).

Data set	Number of clusters	FIHC	UPGMA	Bisecting *k*-means
Classic4	4	1.45	1.59	1.37
Hitech	6	1.62	1.87	1.65
Re0	13	1.83	1.98	1.71
Reuters	65	2.01	2.03	1.92
Wap	20	1.66	1.48	1.77

Table IV.
IV. Entropy comparison

Recall	VSM precision	OntOAIr precision
0.1	0.53	0.61
0.2	0.41	0.51
0.3	0.32	0.36
0.4	0.25	0.28
0.5	0.20	0.25
0.6	0.15	0.19
0.7	0.10	0.13
0.8	0.07	0.09
0.9	0.03	0.04
1.0	0.01	0.02

Table V.
V. Recall and precision results

In order to produce the recall-precision curve of Figure 5, recall and precision were normalized taking the best and the worst case into account (Hamel, 2009).

The experiments show that the OntOAIr method is stable and feasible to support information retrieval. It improves slightly the retrieval effectiveness in comparison with VSM, especially at the top documents retrieved. However, more experimentation and statistical analysis are needed to generalize this argument. The experiments were conducted on a Windows XP system running on a 2.8GHzPentium (R) processor with 1GB of memory.

8. Supporting digital libraries use in developing countries

Digital libraries have a key role to play as means to help individuals and organizations in the process of transforming the information they have access to into knowledge that will make them competitive at the regional, national and international levels. In this sense, open archives and open access collections may support the work of researchers, lecturers and students from developing countries while keeping costs at manageable levels and providing the foundations for constructing their own repositories and participating in the knowledge society. However, much work is needed to assist these users so their insertion into a world academic community may happen in an expedite manner.

Technically, the implementation and use of a digital library implies access to multiple databases that hold either or both metadata and full-text documents. Freely available methods for finding, selecting, organizing, filtering, and presenting information from digital libraries have been made available. For example, digital library platforms such as JeromeDL, BRICKS (Ryszard and McDaniel, 2010) or Greenstone (Witten and Bainbridge, 2007) have long been available and were designed to facilitate the construction and deployment of digital libraries. Still, institutions in

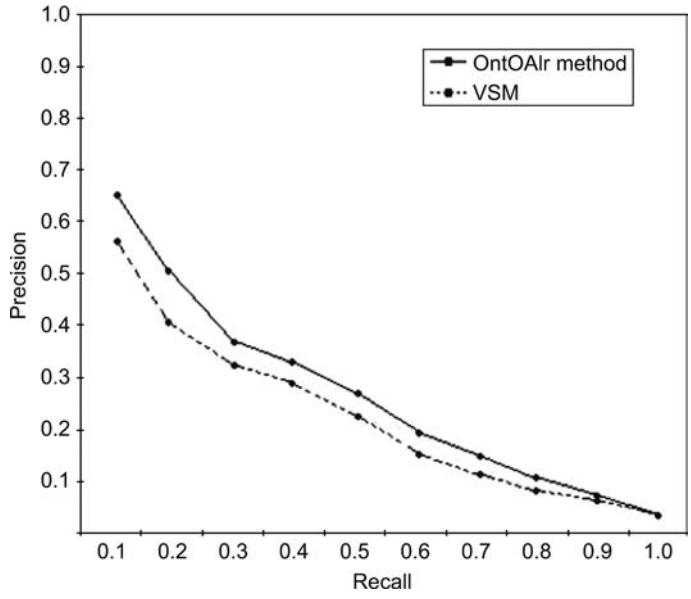


Figure 5.
Precision and recall of OntOAIr and VSM

developing countries must necessarily devise strategies to adapt methods or platforms to take advantage of existing services and, more importantly, to assist users in the processes of building, exploiting and disseminating their own collections. Unfortunately, users in these countries have traditionally perceived themselves as consumers of information and not as producers. Thus, for digital libraries to have a greater impact on development, they must empower users to understand their contents and to use it to generate new knowledge. Simple, consistent interfaces and document models are needed so users can realize the potential of digital libraries not only as information repositories, but also as individual and collective information spaces that allow users to select, annotate and organize existing documents to support the collaborative construction of new collections.

Our work on OntOAIr aims precisely at assisting users to gain a better understanding of the knowledge represented in existing open archives. If used among research communities, we expect mechanisms such as those provided by OntOAIr for semi-automating the organization of large information spaces to leverage the utilization of knowledge and, eventually, the generation of new knowledge that is much needed to foster progress in diverse areas of developing countries.

9. Conclusions

We have presented work that addresses issues in the semi-automatic organization of information spaces derived from accessing open archives. We introduced OntOAIr, an ontology learning method that is robust enough to construct lightweight ontologies from multiple collections of documents with minimal human intervention. This method is based on four main tasks: harvesting, representation, clustering, and formalization.

The harvesting task obtains the documents from the collections. The request verbs of OAI-PMH protocol were used for harvesting, although other methods and techniques for distributed systems could carry out this task.

The representation task constructs a vector representation for each harvested document after a stopword elimination process and the assignment of weights to keywords. The OntOAIR method uses the TF*IDF method, but other methods such as Jaccard index can also be used. The clustering task applies the FIHC algorithm to produce a tree of labeled clusters. We proposed an adaptation of this exclusive hierarchical algorithm, which removes the child pruning process and the sibling merging process from the original version proposed by (Fung *et al.*, 2006). The selection of values for the input parameters (support of an item set, global support and cluster support) strongly influences the accuracy of the clustering; however, experimental evaluation (using the F measure and the entropy), have shown that the FIHC algorithm is as at least as good as UPGMA algorithm and bisecting k-means algorithm.

The formalization task transforms the tree of clusters into a lightweight ontology. The XML, RDF and OWL-DL languages were used to explore different alternatives to formalize the ontologies constructed by the OntOAIr method. The XML language produces simple, readable, and reusable ontologies, but their semantics is only accessible to humans, not to machines. The RDF language enables the definition of the vocabulary and the specification of properties, but the expressivity to represent relationships is limited. The OWL-DL language offers RDF compatibility, computational completeness, and decidability. However, the OWL-DL model is more complex than the XML Schema and the RDF schema.

In comparison with other ontology learning methods, the OntOAIr method has the following advantages:

- uses general domain documents;
- produces a hierarchy of labeled clusters; and
- supports scalable information retrieval models (similarity functions only involve queries and cluster labels).

The main disadvantages are the following:

- the determination of appropriate values for input parameters;
- the assignment of keywords weights do not take into account the structure of documents; and
- the representation of concepts of a single term (the current version of the method does not manage phrases).

Ontologies of records are an information retrieval model on their own because they support indexing and retrieval. The indexing process is based on two tasks:

(1) The selection of informative elements of each record.

(2) The organization of the records in data structures that enable the search.

A retrieval process has been implemented through the keyword-based retrieval model. This model assumes that a record that does not match any term in the query is not relevant.

An algorithm to establish similarity between queries and groups of records has been developed for the keyword-based retrieval model. We have conducted some small-scale experimentation and, as a result, we have demonstrated that the effectiveness of the keyword-based retrieval model is similar to that of the VSM. Averaged values show a recall of 87 per cent and a precision of 71 per cent. However, further experimentation and larger document sets are needed to test and improve our method.

The prototypical system called OntoSIR was implemented to validate the potential of the OntOAIr method and the feasibility of the retrieval model. OntoSIR can be regarded as a searching service for the Open Archives Initiative community. In addition to OAI-compliant data providers, OntoSIR can use other collections. The main disadvantage of this system is its asynchronous operation.

The OntOAIr method can be used to support manual construction of ontologies, to cluster the responses of search engines, or as a basis to support reasoning in Semantic Web contexts. Testing the OntOAIr method with large collections is the first of our planned tasks for the immediate future. Then, open research lines to be considered are:

(1) Adding inference mechanisms that search through the ontologies and deduce results in an organized manner.

(2) Defining tasks to support the maintenance of ontologies.

Inference is a useful tool to complete missing information. OntoSIR could be used for implicit query expansion or to automatically maintain the consistency of the ontologies of records.

Ontologies are rarely static, thus we propose two tasks for their maintenance: First, the inclusion of a set of records in a previously constructed ontology without compromising the accuracy of the clustering or the effectiveness of the retrieval, and second, the management of versioning, which should help to keep track of the evolution of ontologies.

The construction of ontologies of records is a step towards the formal encoding of the content of collections. We expect the reuse of ontologies of records by other applications and ontologies to support Semantic Web tasks such as knowledge acquisition, knowledge management. and similarity-based retrieval.

We still need to systematically observe communities of practice as they rely on semi-automatically constructed ontologies as the basis for organizing their information spaces. However, our initial findings indicate that the approach promoted by OntOAIr may be of significant help for accessing and organizing large open collections by scientists and specialists as well as by more general user communities that need support for knowledge-intensive activities.

Notes

1. The list of data providers is available at: www.openarchives.org/Register/BrowseSites
2. The collection of Communications of the ACM (CACM) and the Time collection are available at: http://ir.dcs.gla.ac.uk/resources/test_collections) (accessed February 14 2010).
3. Classic4 collections are available at: ftp://ftp.cs.cornell.edu/pub/smart/ (accessed February 14 2010).
4. Hitech and Wap collections are available at Text REtrival Conference TIPSTER. 1999. Available at: http://trec.nist.gov (accessed February 14 2010).
5. Yahoo! subject hierarchy web pages are available at Yahoo! available at: www.yahoo.com (February 14 2010).
6. CACM collections are available at: www.dcs.gla.ac.uk/idom/ir_resources/test_collections/ (accessed February 14 2010).

References

Aitken, S. and Reid, S. (2000), "Evaluation of an ontology-based information retrieval tool", in Gómez, A., Benjamins, V.R., Guarino, N. and Uschold, M. (Eds), *Workshop on the Applications of Ontologies and Problem-Solving Methods. European Conference on Artificial Intelligence (ECAI'00), Berlin, September*, pp. 34-43.

Berry, W.M. and Castellanos, M. (2010), *Survey of Text Mining II: Clustering, Classification, and Retrieval*, Springer-Verlag, London.

Borst, W. (1997), *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*, Centre for Telematica and Information Technology, University of Twente, Enschede, technical report.

Brase, J. and Nejdl, W. (2004), "Ontologies and metadata for elearning", in Staab, S. and Studer, R. (Eds), *Handbook on Ontologies*, 2nd ed., International Handbooks on Information Systems, Springer, Dordrecht, Heidelberg, London and New York, NY, pp. 555-74.

Cui, Z. and O'Brien, P. (2000), "Domain ontology management environment", *Proceedings of the 33rd Hawaii International Conference on System Sciences '00 (HICSS'00), Hawaii, January*, pp. 8-15.

Diederich, J. and Balke, W. (2007), "The semantic growbag algorithm: automatically deriving categorization systems", *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries '07 (ECDL'07, Budapest, Hungary, September). Lecture Notes in Computer Science, Vol. 4675*, Springer, Berlin, pp. 33-40.

doccluster (2007), *Clustering C++ Library Documentation*, available at: http://wikipedia-clustering.speedblue.org/download/ClusteringDoc/main.html (accessed November 9, 2009).

Fung, B.C.M., Wang, K. and Ester, M. (2006), *Hierarchical document clustering, The Encyclopedia of Data Warehousing and Mining, Volume I*, Idea Group Reference, Hershey, PA, London, Melbourne and Singapore, pp. 555-9, available at: www.cs.sfu.ca/~wangk/pub/FWE05dwm.pdf

Giunchiglia, F., Marchese, M. and Zaihrayeu, I. (2006), "Encoding classifications as lightweight ontologies", *Journal of Data Semantic*, Vol. VIII, Winter.

Gómez, A., Fernández, M. and Corcho, O. (2004), *Ontological Engineering*, Springer-Verlag, London.

Gruber, T.R. (1993), "A translation approach to portable ontology specification", *Knowledge Acquisition*, Vol. 5 No. 2, pp. 199-220.

Guojun, G., Chaoqun, M. and Jianhong, W. (2007), *Data Clustering: Theory, Algorithms, and Applications*, ASA-ISAM, Philadelphia, PA, ASA-SIAM Series on Statistics and Applied Probability.

Halgamuge, S.K. and Wang, L.P. (2005), *Classification and Clustering for Knowledge Discovery*, Springer-Verlag, Berlin and Heidelberg.

Hamel, L. (2009), *Knowledge Discovery with Support Vector Machines*, John Wiley & Sons, Chichester, Wiley Series on Methods and Applications in Data Mining.

Harrison, T.L., Elango, A., Bollen, J. and Nelson, M. (2004), *Initial Experiences Re-exporting Duplicate and Similarity Computation with an OAI-PMH Aggregator*, available at: http://arxiv.org/abs/cs/0401001

Jain, A.K. and Dubes, R.C. (1988), *Algorithms for Clustering Data*, Prentice Hall, Upper Saddle River, NJ.

Karouri, L., Aufaure, M.A. and Bennacer, N. (2006), "Context based hierarchial clustering for the ontology learning", *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), Hong Kong, December*, pp. 420-7.

Lagoze, C. and van de Sompel, H. (2001), "The open archives initiative: building a low-barrier interoperability framework", *Proceedings of the Joint Conference on Digital Libraries (JCDL'01), Roanoke, VA*, pp. 54-62.

Lambert, M.S., Tennoe, T.M. and Henssonow, F.S. (2010), *UPGMA*, VDM Verlag Dr Mueller Ag & Co. KG, Saarbrücken.

Lassila, O. and McGuinness, D.L. (2001), "The role of frame-based representation on the semantic web", *Electronic Transactions on Artificial Intelligence*, Vol. 6 No. 5, available at: www.ep.liu.se/ea/cis/2001/005/cis01005.pdf

Lewis, D.D. (1997), "Reuters-21578 text categorization test collection", *Distribution 1.0*, September 26, available at: http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html (accessed June 28, 2008).

Ljubič, P., Lavrač, N., Plisson, J., Mladeniae, D., Bollhalter, S. and Jermol, M. (2005), "Automated structuring of company competencies in virtual organizations", *Proceedings of the Conference on Data Mining and Data Warehouses 2005 (SiKDD 2005), Ljubljana, October*, pp. 190-3.

McCandless, M., Hatcher, E. and Gospodnetić, O. (2010), *Lucene in Action*, 2nd ed., Manning Publications, Stamford, CT.

Medina, M.A. and Sánchez, J.A. (2008), "OntOAIr: a method to construct lightweight ontologies from document collections", *Proceedings of the Ninth Mexican International Conference on ComputerScience 2008 (ENC 08), Baja California, México, October*.

Medina, M.A., Sánchez, J.A., Chávez, A. and Benitez, A. (2004), "Designing ontological agents: an alternative to improve information retrieval in federated digital libraries", *Proceedings of the Atlantic Web Intelligence Conference 2004 (AWIC'04, Cancún, México, May). Advances in Web Intelligence, Lecture Notes in Computer Science. Vol. 3034*, Springer, Berlin, pp. 155-63.

Navigli, R. and Velardi, P. (2004), "Learning domain ontologies from document warehouses and dedicated web sites", *Computational Linguistics*, Vol. 30 No. 2, pp. 151-79.

Plisson, J., Mladeniae, D., Ljubič, P., Lavrac, N. and Grobelnik, M. (2005), "Using machine learning to structure the expertise of companies: analysis of the Yahoo! business data", *Conference on Data Mining and Data Warehouses (SiKDD 2005 Proceedings). 7th International Multi-conference on Information Society IS'05*, pp. 186-9.

Ryszard, K.S. and McDaniel, B. (2010), *Semantic Digital Libraries*, Springer-Verlag, Berlin and Heidelberg.

Witten, I.H. and Bainbridge, D. (2007), "A retrospective look at Greenstone: lessons from the first decade", *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '07), Vancouver, BC, June 18-23*, ACM, New York, NY, pp. 147-56.

Xu, R. and Wunsch, C.D. II (2008), *Clustering*, Wiley/IEEE Press, New York, NY.

Further reading

Fung, B.C.M., Wang, K. and Ester, M. (2005), "Hierarchical document clustering using frequent itemsets", *Proceedings of the Third SIAM International Conference on Data Mining (SDM'03), San Francisco, CA, May*, pp. 59-70.

Witten, I.H., Bainbridge, D. and Nichols, D. (2009), *How to Build a Digital Library*, 2nd ed., Morgan Kaufmann Publishers, Amsterdam, Boston, MA, Heidelberg, London, New York, NY, Oxford, Paris, San Diego, CA, San Francisco, CA, Singapore, Sydney and Tokyo.

About the authors

J. Alfredo Sánchez is Professor of Computer Science and Director of the Laboratory of Interactive and Cooperative Technologies (ICT) at Universidad de las Américas Puebla (UDLAP). He holds MSc and PhD degrees in Computer Science from Texas A&M University, and a BEng degree in Computer Systems from UDLAP. Since 1996, he has conducted R&D projects in areas such as digital libraries, human-computer interaction and computer-supported cooperative work. Results from these projects have been reported in more than 90 refereed and invited publications. He has been a visiting professor at the University of Waikato, New Zealand, and a visiting scientist at the Center for Botanical Informatics of the Missouri Botanical Garden. Professor Sánchez serves on the editorial board of the *International Journal of Digital Libraries* and coordinates the Digital Libraries Community of the Mexican Internet 2 Consortium. He also has served as president of the Mexican Computer Science Society and has been a member of the National Researchers System in Mexico. J. Alfredo Sánchez is the corresponding author and can be contacted at: J. Alfredo.Sánchez@gmail.com

Maria Auxilio Medina is an Associate Professor of Computer Science at Universidad Politécnica de Puebla (UPP). She holds MSc and PhD degrees in Computer Science from Universidad de las Américas Puebla (UDLAP) and a BEng degree in Computer Systems from

Benemérita Universidad Autónoma de Puebla (BUAP). She has participated in projects related to the development of software agents and their applications to digital libraries. Currently, her research topics are information retrieval, knowledge representation based on ontologies, semantic web, and information and communications technologies. Dr Medina coordinates ICT's academic community at the UPP.

Oleg Starostenko is Professor of Computer Science at Universidad de las Américas Puebla (UDLAP). He holds a BEng and MSc degrees from Lviv State University "Lvivska Polytechnica" and a PhD degree in Optoelectronics from Benemérita Universidad Autónoma de Puebla (BUAP). He is a member of the National Researchers System (SNI).

Antonio Benitez is an Associate Professor of Computer Science at the Universidad Politécnica de Puebla (UPP). He holds MSc and PhD degrees in Computer Science from Universidad de las Américas Puebla (UDLAP) and a BEng degree in Computer Systems from Benemérita Universidad Autonoma de Puebla (BUAP). He has participated in projects related to the development of robotics and description of virtual environments. His current research areas include computer perception and virtual reality. Dr Benitez coordinates the graduate education department at the UPP. He is a member of the National Researchers System (SNI).

Eduardo López Domínguez has been Associate Professor of Computer Science at the Universidad Politécnica de Puebla (UPP). He holds MSc and PhD degrees in Computer Science from Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOEP) and a BEng degree in Computer Systems from Instituto Tecnológico de Tehuacán (ITT). Since 2004, he has participated in projects related to the development of mobile distributed systems, partial order algorithms and multimedia synchronization.