The use of Taxonomies as a way to achieve Interoperability and improved Resource Discovery in DSpace-based Repositories

Miguel Ferreira, Ana Alice Baptista
(mferreira@dsi.uminho.pt, analice@dsi.uminho.pt)

Department of Information Systems,
University of Minho,
4800-Guimarães
Portugal

Abstract

In this paper, we present the ongoing work being developed at University of Minho in the context of Institutional Repositories, more precisely the ones based on the DSpace platform (developed jointly by the Massachusetts Institute of Technology and Hewlett-Packard). As part of our research, we have created an add-on for DSpace to ensure authority control over the keywords that human cataloguers may use to describe their items of information. These keywords are used by the visitors of the repository for searching and browsing the catalogue. The keywords are organised in a taxonomy that results from the combination of several specialised thesauri, one for each community of users (e.g., Computing, Engineering, Architecture, etc.). Each of these off-the-shelf thesauri is described by a simple XML file. The first thesaurus we have imported into our system was the publicly-available *Association for Computing Machinery (ACM) Computing Classification System (CCS)*. The success of the add-on exceeded our expectations given that there was a broad acceptance by the community of users, probably driven by its simplicity and ease of use. ACM contributed to our work by validating our XML version of the CCS and by publishing it on their Web site. Our most recent endeavour is centred on the conversion of the ACM CCS to OWL (Web Ontology Language). Through the use of taxonomies we expect to achieve better interoperability between analogous systems as well as improve the discovery of resources in our repository. Future work will be focused on the improvement of the add-on to support more complex structures, such as thesauri or ontologies.

Keywords: Controlled Vocabularies, Taxonomy, Thesaurus, Ontology, Institutional Repositories, DSpace.

Introduction

As the amount of globally accessible information grows, so does the inherent difficulty in finding desired items of information[1] [1, 2]. In the context of document archives, one of the most applied techniques to facilitate the discovery of items, both in traditional manual systems as in newer computerized systems, has been *indexing* [2].

Indexing consists in the assignment of values to predefined attributes to serve as a basis for searching [2, 3] and resource discovery. The combination of these attributes and values should constitute sufficient information to successfully characterise the contents of a document and enable the future retrieval of that document by solely looking at this information [2]. Examples of commonly found attributes are: *author*, *title*, *subject*, *abstract*, etc. These are generally referred to as *metadata*.

Metadata can be attained either automatically or manually. Fully automated methods generally result in the creation of a full-text catalogue, i.e., an index containing the most frequent words found in the document [4]. Other automatic procedures analyse the contents of the document and attempt to assign pre-established concepts to metadata attributes [2, 5].

Manual approaches generally consist in scanning a document for keywords that are considered relevant or more adequate to describe the subject of the document. These keywords may be selected from a predefined set of keywords, also know as controlled vocabulary, or self-thought by the person who is indexing document. Manual indexing usually results in better quality metadata which will later reflect the precision and recall of search results [2]; however, manual indexing, especially the one carried out by trained professionals is highly expensive, time-consuming and evermore unfeasible given the amount of information being produced today. Some repository systems combine the two approaches [2] and use manually inserted metadata as well as automatically generated full-text indexes to facilitate the discovery of items.

Given the ever-growing call for quality metadata and the amount of information being produced today combined with the insufficient funding necessary to finance professional indexing, a new paradigm has emerged: the *self-archiving* [2]. Self-archiving means that the producer of a document is the main responsible for the creation of its metadata. Doing it so, the task of

[1] An information item is defined broadly as an object that contains information. The representation can be in different media types such as text, image, video, etc. An information item is also referred to as a *document* inside the corpus of this paper.

creating quality metadata resorting to human expertise becomes less onerous as the task is distributed by several people. This change in the paradigm implies that documents will now be indexed by amateurs who know little or nothing about indexing [2].

The keywords chosen by users to describe their works can be extremely variable resulting in a poor inter-indexer consistency [2]. This results in an increased difficulty in finding similar items. Furnas, et al [6] showed that the probability of two amateur indexers using the same keywords to describe a specific object is less than 20%. It is crucial that both manual and automatic indexing procedures use controlled vocabularies to standardise document descriptions and to simplify subsequent searches by establishing a common language in a given domain.

In this paper we describe an ongoing project that consists in the implementation of a cross-domain controlled vocabulary used to describe self-archived items in DSpace-based repositories.

This paper is organised as follows: section 2 attempts to shed light on some of the concepts that will be used throughout the corpus of this paper; section 3 provides a quick overview on the genesis of institutional repositories; section 4 describes some of the work being developed at University of Minho in the context of institutional repositories; section 5 explains in detail how we implemented an add-on that enhances DSpace with controlled vocabularies; on section 6 we draw some conclusions about our work and discuss some proposals for future work.

Taxonomies, Thesauri and Ontologies

For the sake of clarity, we felt important to include a section in this paper to describe some of the concepts that will be used throughout the corpus of this piece.

Taxonomy

Taxonomy consists in a subject-based classification that arranges the terms in a controlled vocabulary into a hierarchy without doing anything further [7].

Almost anything – animate objects, inanimate objects, places, and events – may be classified according to some taxonomic scheme [8].

Mathematically, a taxonomy is a tree structure of classifications for a given set of objects. At the top of this structure is a single classification – the root node – that applies to all objects. Nodes below this root are more specific classifications that apply to subsets of the total set of classified objects [8].

Hence, a taxonomy is a collection of terms used to describe *things* that are grouped together in a tree structure. We are able to identify parent-child relationships between the terms in the controlled vocabulary.

Thesaurus

Thesauri basically take taxonomies as described above and extend them to make them better able to describe the world by not only allowing subjects to be arranged in a hierarchy, but also allowing other statements to be made about the subjects [7].

The following properties and relationships are incremented by thesauri:

- *Scope Note* – A string property attached to the term explaining its meaning within the thesaurus.
- *Use* – Refers to another term that is to be preferred instead of a certain term; implies that the terms are synonymous.
- *Related Term* – Refers to a term that is related to a given term, without being a synonym or a broader/narrower[2] concept.

Ontology

With ontologies the creator of the subject description language is allowed to define the language at will. Ontologies in computer science came out of artificial intelligence, and have generally been closely associated with logical inferencing and similar techniques, but have recently begun to be applied to information retrieval [7].

Ontologies extend the concept of thesaurus by enabling the creator of the controlled vocabulary to define new properties and relationships between terms.

Institutional repositories

In recent years, the development of institutional repositories has emerged as a new strategy for the preservation, publishing and dissemination of scholarly communications [9]. Universities and other research institutions throughout the world are actively planning and implementing institutional repositories aiming at providing its members a set of new services such as the archiving of research results (article preprints and post-prints, theses, and dissertations);

[2] Parent-child relationship.

management of digital collections; preservation of digital documents; housing of teaching materials and electronic publishing of journals and books [10]. Institutional repositories also fulfil the important task of levering scholars from the burden of administering their own publishing system (i.e. personal Web site) [9].

The digital repository genesis has been short, beginning in the late 2000 when the UK's University of Southampton released a software package called EPrints [10]. Since then, the movement to establish digital repositories has gained momentum, encouraged by a convergence of dropping costs for online storage, the proliferation of broadband networking technologies and the development of metadata standards to describe repository content [9, 10].

Other initiatives, such as the Open Archives Initiative (OAI) [11] sustain the general acceptance and proliferation of institutional repositories. The OAI is a collaborative effort towards the development and promotion of standards and solutions such as the OAI Protocol for Metadata Harvesting [12], which allows an institution to create descriptive metadata for the items in its custody and making it available to others who wish to use it [9, 10].

As well as EPrints, other repository systems have been developed by different organisations. DSpace is a good example of a general-purpose repository designed to capture the intellectual output of research organizations that is rapidly gaining wide-range acceptance, especially by universities and research institutes.

The development of DSpace is a responsibility of the MIT Library and Hewlett-Packard. Unlike EPrints, DSpace supports the ingestion of a wide range of digital material types [10]. The system itself does not present any restrictions on the formats that should be accepted; although, such restrictions can be set up by the administrator of the repository. In addition, DSpace is open-source allowing everyone in the community to contribute by building original enhancements and customisations.

DSpace Development at University of Minho

The University of Minho (UMinho) was the first institution in the Portuguese speaking world to use a translated version of DSpace. DSpace related activities at University of Minho started in April 2003 and since then many developments have been made.

The first step has been taken by the UMinho Documentation Services (SDUM) with the translation to Portuguese of the entire DSpace system and its implementation in the RepositóriUM[3]. This version of DSpace has been downloaded for use in many other institutions in Portugal and Brazil.

This same version was used as a basis for the Papadocs[4] system. Papadocs was first created to provide access to all assignments made by students of the Department of Information Systems. Some changes had to be made to the original version, in particular to what concerns the metadata. Additional fields in the area of education were appended to the basic Dublin Core [13] element set. Examples of newly created fields are:

- *Creator.Identifier* – i.e., student's id number;
- *Contributer.Teacher* – i.e., the identification of the main teacher responsible for a particular assignment or discipline.
- *Grade* – i.e., assignment classification;

This customized version of DSpace served as test-bed for a series of add-ons that enhance the platform in many ways. A Web site called *DSpace-Dev @ University of Minho*[5] was created with the purpose of sharing these add-ons with the interested community and generate discussion about other research projects currently in hands. All the source-code can be downloaded on this Web site.

The following sections provide a short description of some of the add-ons that resulted from this project.

Commenting Add-on

The Commenting Add-on consists in a set of classes, servlets and custom tags that bring informal communication capabilities to the DSpace environment. The informal communication is assured by a threaded forum that can be attached to any DSpace resource: web-page, community, collection, submitted item or e-person.

Recommendation Add-on

The Recommendation Add-on consists in a set of custom-tags that provide suggestions of resources related to a given selected resource. The most relevant custom-tag receives a parameter identifying the selected resource (type and id) to which the suggestions/recommendations shall be given. The

[3] http://repositorium.sdum.uminho.pt
[4] http://papadocs.dsi.uminho.pt
[5] http://dspace-dev.dsi.uminho.pt

add-on iterates through the database collecting items, e-persons and comments that are considered to be relevant.

Web of Communication Add-On

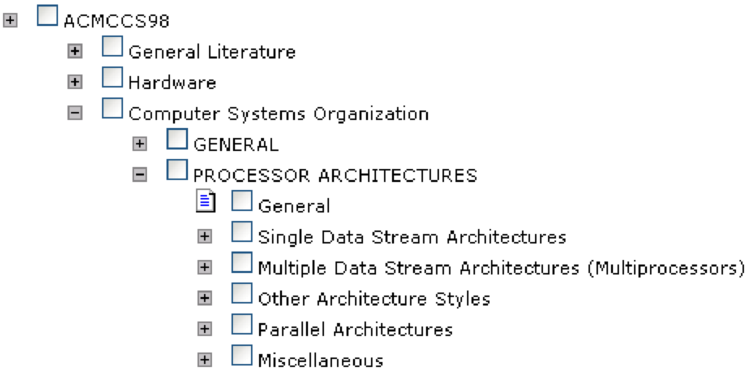
The 3D Web of Communication allows the user to discover hidden relationships between items, comments and people. It works by displaying a VRML 3D web of resources involved in a communication process. The user is also able to jump to specific items on the environment thus providing a 3D navigational system over DSpace

Controlled Vocabulary Add-on

Many repository systems favour self-archiving and DSpace is no exception. Users are stimulated to submit their works to the repository and generate themselves the appropriate descriptive metadata. Users that submit items to the repository are allegedly rewarded by an increased visibility of their work.

In most archiving scenarios, it is natural that a certain degree of ambiguity and heterogeneity will be found in the metadata provided by different users to documents with similar content. This can also be observed in archives where items have been indexed by trained professionals. To downsize this problem, we have developed an add-on for DSpace that restricts the keywords that users may employ during indexing stages of self-archiving.

During submission, users are asked to enter the keyword(s) that best describe their works. With our add-on in place, users are presented with a taxonomy that displays the terms that are allowed to be used as descriptors. For each community of users that interact with the repository a different taxonomy is presented. Each of these taxonomies is rendered to the user as an expandable tree (see Figure 1).



- ⊞ □ ACMCCS98
  - ⊞ □ General Literature
  - ⊞ □ Hardware
  - ⊟ □ Computer Systems Organization
    - ⊞ □ GENERAL
    - ⊟ □ PROCESSOR ARCHITECTURES
      - □ General
      - ⊞ □ Single Data Stream Architectures
      - ⊞ □ Multiple Data Stream Architectures (Multiprocessors)
      - ⊞ □ Other Architecture Styles
      - ⊞ □ Parallel Architectures
      - ⊞ □ Miscellaneous

Figure 1: Excerpt of the ACM Computing Classification System taxonomy presented to the user.

The development and maintenance of these domain-specific taxonomies is not our main concern. We have elected publicly available classification systems, one per each scientific community, to be used in the repository. Since these are highly used classification systems, the interoperability between similar repositories is simplified as the probability of finding other systems that use the same controlled vocabularies becomes higher.

DSpace is compatible with the Open Archives Initiative Protocol for Metadata Harvesting [12], a protocol that allows the creation of centralised catalogues of metadata to facilitate the discovery of items in physically distributed repositories. In this context, an agreement on the set of keywords used to describe the items in custody is of considerable importance.

The first controlled vocabulary we have imported into our system was the ACM Computing Classification System (1998 version) [14]. This controlled vocabulary is being used by the students of the Department of Information Systems to describe their academic projects. Recent contacts with other departments also interested in publishing their students' projects have resulted in the opening of the Papadocs repository to the Civil Engineering and Architecture communities. This event conducted to the adoption of two other taxonomies appropriate to describe the items submitted by members of these communities – we are now using a sub-set of the Engineering Index Thesaurus [15] and negotiating the possibility of using the Art & Architecture Thesaurus [16].

How does it work?

The add-on works by loading all the included taxonomies from independent XML files (stored on the server's file system) and rendering them as trees to the user. The structure of these XML files is very straightforward. We use four different elements to represent the whole structure of the taxonomies: *node*, which contains information about a specific term; *isComposedBy*, a wrapper element that contains a list of child nodes; *isRelatedTo*, an element that contains links to other related nodes in the taxonomy; and *hasNote*, an element that allows the inclusion of a small descriptive note about the term (see Figure 2).

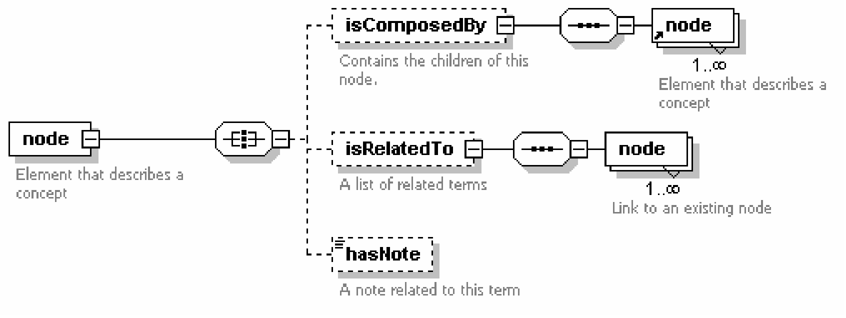
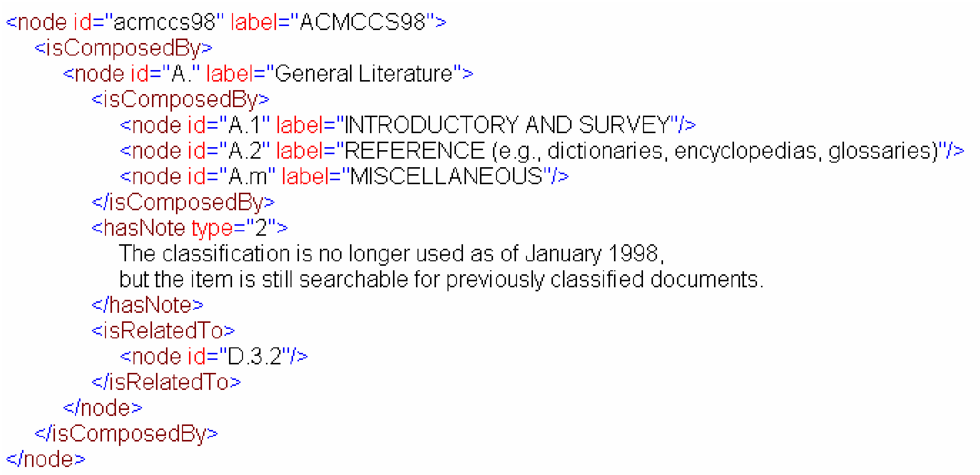

isComposedBy
Contains the children of this node.
node
1..∞
Element that describes a concept
node
Element that describes a concept
isRelatedTo
A list of related terms
node
1..∞
Link to an existing node
hasNote
A note related to this term

Figure 2: Schema that validates our XML taxonomy.

As we can see, this schema goes beyond the representation of simple taxonomies. The relationship *isRelatedTo* and the property *hasNote* allow the description of thesauri according to the ISO standard 2788:1986 [17]. However, at the moment our rendering system does not process these relationships so the user is presented with a limited view of the thesaurus actually described in the XML file. The reason for this approach is not technological, but social. We wanted our classification system to be simple and user-friendly to students that are not used to manipulate complex structures. Adding further dimensions to the taxonomy would probably steer users away from its use.

In Figure 3 is shown a small sample of the ACM Computing Classification System compliant with our schema.



```
<node id="acmccs98" label="ACMCCS98">
   <isComposedBy>
      <node id="A." label="General Literature">
         <isComposedBy>
            <node id="A.1" label="INTRODUCTORY AND SURVEY"/>
            <node id="A.2" label="REFERENCE (e.g., dictionaries, encyclopedias, glossaries)"/>
            <node id="A.m" label="MISCELLANEOUS"/>
         </isComposedBy>
         <hasNote type="2">
            The classification is no longer used as of January 1998,
            but the item is still searchable for previously classified documents.
         </hasNote>
         <isRelatedTo>
            <node id="D.3.2"/>
         </isRelatedTo>
      </node>
   </isComposedBy>
</node>
```

Figure 3: A sample of the ACM CCS taxonomy in XML.

Finding items in the repository

As stated previously, the keywords used to describe the submitted items are selected from a tree of terms. When a term is picked from this tree, the full-path between the root node and selected node is used to represent the chosen keyword. For example, a book on the programming language Java could generally be described by the following keywords:

- ACMCCS98/Software/PROGRAMMING TECHNIQUES/Object-oriented Programming
- ACMCCS98/General Literature

The advantage of this approach is twofold. It removes the ambiguity inherent to certain concepts by accompanying them with the correct context and allows the realisation of more general queries. For example, the book described above will be included in the list that results from querying the repository for items whose subject is *"ACMCCS98/Software/ PROGRAMMING TECHNIQUES"*. The concept matching is accomplished easily due to the fact that the most general concept is a mere substring of the most specific descriptor.

Conclusions and Future work

In this paper we present an Add-on for DSpace that enables repository administrators to compel its users to use a controlled set of keywords to describe self-archived items of information. The advantages of using controlled vocabularies are enumerated throughout the corpus of this document.

It is our belief that the introduced controlled vocabulary system is adequate to our objectives due its simplicity. Users can easily find the terms they are looking for by expanding just a few branches of the taxonomy. Further research should be performed to make sure our beliefs are truthful.

Much can be done in the future to improve the add-on. First, we could render the XML as a true thesaurus by exposing the *isRelatedTo* relationships as links to the user. Secondly, we could upgrade the thesaurus model to support other types of relationships and/or properties. This would mean start using ontologies to describe concepts and their relationships. In certain contexts this would be very useful, but in the context of our repository, the augmented number of relationships and complexity introduced by this new class of structures could be dissuasive for most users.

Interesting work could also be developed in the area of automatic cataloguing. The add-on could be enhanced to suggest keywords to the user by analysing the contents of the document being submitted or by comparing it with other documents already in the repository.

If the system keeps on being adopted by different communities, we will soon come to a state where some branches of the incorporated taxonomies will overlap. When this happens, we must start using techniques to merge taxonomies [18].

The main purpose of the Papadocs repository was to serve as a test-bed for the research we are performing in the field of institutional repositories. We have come to a point where some of the technology we have produced is being considered to be included in the RepositoriUM – the official institutional repository of University of Minho – where it will be used by a greater number of users. The efficiency and the scalability of our solutions are now being subject to a higher degree of consideration. Furthermore, some of the add-ons we have developed are being considered for inclusion in the official DSpace source-code.

It is also worth noting that our XML version of the ACM Computing Classification System is now being used by ACM it self and is publicly available for download on their Web site[6]. Our most recent endeavour is centred in the conversion of the ACM CCS from our XML format to OWL (Web Ontology Language) [19] in order to make it suitable for a greater number of users.

References

[1] D. Teixeira, M. Ferreira, and V. Verhaegh, "An Integrated Framework for Supporting Photo Viewing Activities in Home Environments," presented at European Symposium on Ambient Intelligence, Eindhoven, The Netherlands, 2003.

[2] Y.-M. Chung, W. M. Pottenger, and B. R. Schatz, "Automatic Subject Indexing Using an Associative Neural Network," presented at 3rd ACM International Conference on Digital Libraries (DL'98), 1998.

[3] C. Meadow, "Text Information Retrieval Systems," *Academic Press Inc.*, 1992.

[4] D. Cunliffe, C. Taylor, and D. Tudhope, "Query-based Navigation in Semantically Indexed Hypermedia," presented at Conference on Hypertext, 1997.

[5] K. Taghva, T. A. Nartker, and J. Borsack, "Recognize, Categorize, and Retrieve," presented at Symposium on Document Image Understand Technology, 2001.

[6] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary Problem in Human-System Communication," *Communications of the ACM*, vol. 30, pp. 964-971, 1987.

[7] L. M. Garshol, "Metadata? Thesauri? Taxonomies? Topic Maps!," Ontopia, 2004.

[8] Wikipedia, "Taxonomy," 2004.

[9] C. A. Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in Digital Age," *ARL Bimonthly Report*, 2003.

[10] H. F. Cervone, "The Repository Adventure," *Library Journal*, 2004.

[11] OAI, "Open Archives Initiative," vol. 2004.

[6] http://www.acm.org/class/1998/

[12] OAI, "The Open Archives Initiative Protocol for Metadata Harvesting," vol. 2004, 2002.

[13] DCMI, "Dublin Core Metadata Initiative," vol. 2004.

[14] N. Coulter, J. French, E. Glinert, T. Horton, N. Mead, R. Rada, A. Ralston, C. Rodkin, B. Rous, A. Tucker, P. Wegner, E. Weiss, and C. Wierzbicki, "The ACM Computing Classification System [1998 Version]," vol. 2004: Association for Computing Machinery, 1998.

[15] E. E. I. Inc., "Ei Thesaurus," 1998.

[16] G. R. Institute, "Art & Architecture Thesaurus," vol. 2004, 2000.

[17] ISO, "Guidelines for the establishment and development of monolingual thesauri, International Organization for Standardization - ISO 2788:1986," 1986.

[18] M. Sintichakis and P. Constantopoulos, "A Method for Monolingual Thesauri Merging," *Communications of the ACM*, 1997.

[19] W3C, "Web Ontology Language (OWL)," 2004.