Ontology-based descriptions of image collections

Ma. Auxilio Medina[1], J. Alfredo Sánchez[2], J. de la Calleja[1], Antonio Benitez[1]

[1]Universidad Politécnica de Puebla
Tercer Carril del Ejido Serrano S/N
Juan C. Bonilla, Puebla, México
[2]Universidad de las Américas Puebla
Ex-Hacienda Santa Catarina Mártir S/N
San Andrés Cholula, Puebla, México
{mmedina,abenitez,jdelacalleja}@uppuebla.edu.mx
j.alfredo.sanchez@gmail.com

Abstract. Thesaurus have been widely used to provide structured vocabularies to describe books, however, they only provide part of the required knowledge for semantic web contexts. This paper proposes the use of ontologies and standard metadata to model semantic descriptions of image collections that refers to book covers. Ontologies represent keyword-based organizations. The classes and the hierarchical relationships of the ontology allow users to query images by topic with common search engines. The *greenBookC* collection is used as a test bed. This collection is published and maintained in Greenstone, a suite of software for digital libraries.

Keywords: digital libraries, ontologies, semantic descriptions

1 Introduction

Users of digital libraries employ keyword based search engines, browsing mechanisms or recommendation systems to find relevant information. In these tools, words are considered as "units of meaning". However, there is a gap between the contents and meanings in multimedia data [4].

Different approaches exist to organize multimedia collections that vary from processing low level features to associate concepts to media. The results of these approaches often require knowledge of experts or specialized users. Most of the users of digital libraries interact with image collections through search engines that allow them to query images by title, date, author, format, publisher or with a feature associated with the file that stores the image. These descriptors are associated with image content.

Frequently, image collections are also described using a set of terms of a keyword-based organization system or with natural language descriptions. In order to extend the implicit syntactic level of these descriptions, semantic descriptions with predefined structures and control vocabularies are considered richer structures to store background knowledge [5].

Our interest is focused on describe images that refers to book covers. Traditionally, thesaurus have been widely used to provide structured vocabularies to describe books, however, they only provide part of the required knowledge for semantic web contexts. In semantic digital libraries (SDLs), materials, tools and meanings are addressed to offer benefits such as: anyone can use it, knowledge is accessible from the SDL, resources are available with the modality anytime anywhere, there are friendly and multi-modal interfaces with multiple connected devises [2].

This paper explores an alternative to associate background knowledge to image descriptions that refers to book covers. The alternative makes use of ontologies and metadata standards. The classes and the hierarchical relationships of ontologies can be exploited by keyword-based search engines. A collection of images is built as a test bed. The collection is published in Greenstone, a suite of software for digital libraries.

The paper is organized as follows. Section 2 presents brief descriptions of software tools that allow users to construct image collections in the semantic web. Section 3 explains the ontologies and metadata management. Section 4 describes the test bed collection. Section 5 presents preliminary results. Finally, Section 6 includes conclusions and suggests future directions of our work.

2 Related work

Open freely distributed software exists around the world to construct image collections for educational, government and commercial institutions. This section describes some of the most common tools used in semantic web applications and depicts some representative collections.

DSpace[1] allow users to create repositories of digital content in multiple formats. This is a widely popular tool that supports the customization of interfaces to fit different user needs [6]. "The Spanish Image collection" is an example of a collection constructed in DSpace. The collection is formed by 1196 JPEG files[2]. The images can be query with descriptors such as author, title or date; a list of categories is used to organize images. The description of images at DSpace is syntax-based.

Fedora commons repository software[3] is a general purpose, open-source digital object repository system under the Apache License. This is a centered platform that enables storage, access and management of digital content, although does not support indexing, discovery and delivery application mechanisms [3]. The modules offer interoperability and extensibility of data. Image collections can be constructed upon Fedora such as the "Maryland Map Collection". This collection is formed by JPEG files that depicts Maryland, the Chesapeake Bay,

[1] Dspace home page is available at http://www.dspace.org/

[2] The Spanish Image Collection is available at:
http://dspace.nitle.org/handle/10090/1267

[3] Fedora home page is available at: http://fedora-commons.org/. This work is licensed under a Creative Commons Attribution-Share Alike 3.0 Unported License

and the surrounding region from 1590 to the present. The interface visualizes images and metadata. The Libraries' Catalog and the ArchivesUM are used to organize this collection.

Greenstone[4] is a suite of software to build and construct collection of semantic digital libraries. This is produced by the New Zealand Digital Library Project at the University of Waikato, and developed and distributed in cooperation with UNESCO and the Human Info NGO. This is an open-source, multilingual software issued under the terms of the GNU General Public License. Greenstone supports a variety of formats that represents text, images, audio and video [6]. The "Historic Campus Architecture Project"[5] developed by the Council of Independent Colleges (CIC) in the United States, is an example of an image collection in Greenstone. In the project, 5000 images of hard copy photographs of 4 x 6 and 8 X 10 inches were transformed into JPEG files. Metadata is used as the descriptors to explore the collection. Metadata is stored in XML files.

After analyzing these software tools, we choose Greenstone to implement the ontology-based alternative to describe image collections. The features of Greenstone related with the support of semantic descriptions of images are the following ones [6]:

- The digital library server runs in different platforms
- Full text is searchable by default
- Servers and harvesters data compliance with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
- Export and ingest data from DSpace
- Collections can be updated anytime without disturbing users
- Dublin Core (DC) is the default metadata format when a new collection is constructed

3 Building semantic descriptions

Although thesaurus have been a typical classification system of digital libraries, they are not able to accomplish the information requirements of the semantic web. Ontologies have larger representation power than thesaurus [2]. An adaptation of the steps proposed [5] was done to construct a lightweight ontology used to build semantic descriptions.

The steps to construct semantic descriptions based on metadata and ontologies that refers to book covers are the following ones:

1. *Construction of a description template for book covers.* This template answer questions such as: what kind of information does users want to record for a particular image, how users query an image collection of book covers, what format is appropriate to store images, how can images be organized,

[4] Greenstone is available at: http://www.greenstone.org/

[5] The History Campus Arquitecture Project is available at http://puka.cs.waikato.ac.nz/cgi-bin/cic/library

what metadata standard is useful to support semantic descriptions, will the descriptions be used locally or a mechanism of information sharing must be considered

2. *Linking images to ontology classes* in such a way that classes can be used as values for a metadata standard. Hierarchical relationships should be used as a query expansion mechanism

3. *Describing additional domain and background knowledge* in metadata standard elements

Each case study might use of a different ontology as well as a distinct metadata standard as input. The goal in this work is to exploit ontological characteristics as the basis to improve image descriptions and retrieval mechanisms.

Table 1 shows an excerpt of an ontology that was used to construct semantic descriptions of book covers that form a collection called *greenBookC*. This collection correspond to book covers of physical books of the library of the UPPuebla. The areas of knowledge proposed by the "Consejo Nacional de Ciencia y Tecnologa" (CONACYT) are used as the main classes of this ontology.

Table 1. An excerpt of an ontology for book covers

1. Mathematical physicist and Earth sciences
2. Biology and chemistry
3. Medicine and health sciences
4. Humanities and behavior sciences
5. Social sciences
6. Biotechnology and food sciences
7. Engineering Computer science Economics Electronics Financial engineering Industrial engineering - Materials - Mechatronics Science and technology Telecommunications
8. Others

The ontology of Table 1 has 8 main classes and 29 subclasses organized in 3 levels. Each class has a label, a level and one or more instances (images). An additional class is used to hold images of specific types of literature: biographies, dictionaries and novels. The classes were constructed using knowledge of experts. Semantic information is represented by metadata attached to each image; in particular, the dc:subject element is used to support the organization of topics, this means that the names of the classes are used to fill this element. Table 1 shows the DC elements used to form a semantic description. An institutional policy help to maintain a controlled vocabulary for the semantic descriptions.

Table 2. DC elements used to construct a semantic description

DC element	Description	Fixed values
`Creator`	Indicates the name of the first author	
`Date`	The year of the book	
`Format`	The digital manifestation of the book	JPEG
`Identifier`	Identifier of the physical book	
`Language`	Language of the content	Spanish, English or French
`Type`	Categorizes the nature of the content	image
`Subject`	A class name	
`Source`	A reference to the owner of the resource	Universidad Politécnica de Puebla
`Title`	The title extracted from the book cover	

The hierarchical relationships of the ontology are useful for free text search. For example, a cover book that belongs to the materials class, it is also considered as a member of the industrial engineering class.

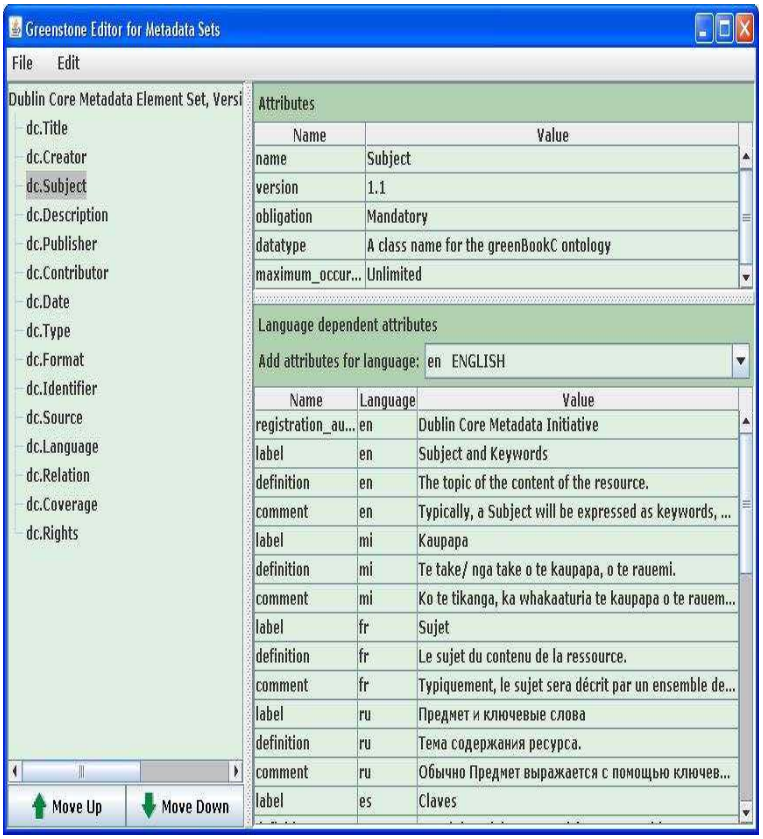
4 greenBookC: an image collection with semantic descriptions at Greenstone

This section describes the use of the ontology described in Section 3 to enrich an image collection of book covers. GreenBookC collection uses existing legal metadata of Greenstone. Semantic description of images are described with the unqualified Dublin Core (DC) elements of Table 2 as shown in Figure 1. After a metadata standard is added, metadata elements need to be filled as illustrated in Figure 2. These data can be assigned in different languages in order to improve collection accessibility.

According to [1], semantic descriptions are stored in standard metadata formats and knowledge representation languages such as XML, RDF, RDF-Schema and OWL. At greenBookC, these descriptions are stored as XML files. These files can be exported to DSpace or be processed by another XML tools.

5 Preliminar results

The greenBookC collection is formed by 1504 images of book covers in JPEG format. JPEG is a standard to represent compressed continuous-tone images [6]. The images are normalized as follows: 32 bits for colours, each one 288 per 352 pixels. The size of the files varies between 33 and 98 Kbytes. A document camera was used to get the images in order to enhance text and graphics.



Dublin Core Metadata Element Set, Versi

- dc.Title
- dc.Creator
- dc.Subject
- dc.Description
- dc.Publisher
- dc.Contributor
- dc.Date
- dc.Type
- dc.Format
- dc.Identifier
- dc.Source
- dc.Language
- dc.Relation
- dc.Coverage
- dc.Rights

Attributes

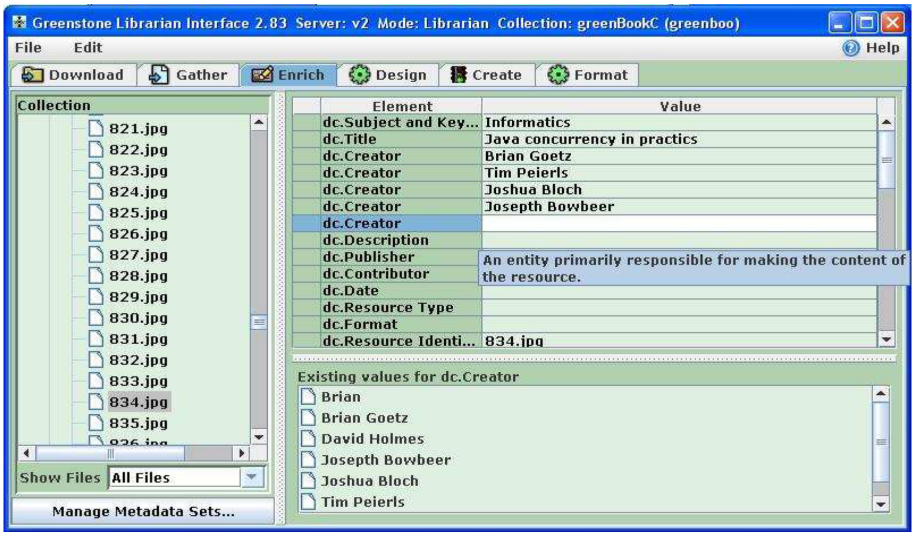
Name	Value
name	Subject
version	1.1
obligation	Mandatory
datatype	A class name for the greenBookC ontology
maximum_occur...	Unlimited

Language dependent attributes

Add attributes for language: en ENGLISH

Name	Language	Value
registration_au...	en	Dublin Core Metadata Initiative
label	en	Subject and Keywords
definition	en	The topic of the content of the resource.
comment	en	Typically, a Subject will be expressed as keywords, ...
label	mi	Kaupapa
definition	mi	Te take/ nga take o te kaupapa, o te rauemi.
comment	mi	Ko te tikanga, ka whakaaturia te kaupapa o te rauem...
label	fr	Sujet
definition	fr	Le sujet du contenu de la ressource.
comment	fr	Typiquement, le sujet sera décrit par un ensemble de...
label	ru	Предмет и ключевые слова
definition	ru	Тема содержания ресурса.
comment	ru	Обычно Предмет выражается с помощью ключев...
label	es	Claves

Fig. 1. Using DC elements at Greenstone



Element	Value
dc.Subject and Key...	Informatics
dc.Title	Java concurrency in practics
dc.Creator	Brian Goetz
dc.Creator	Tim Peierls
dc.Creator	Joshua Bloch
dc.Creator	Joseppth Bowbeer
dc.Creator	
dc.Description	
dc.Publisher	An entity primarily responsible for making the content of the resource.
dc.Contributor	
dc.Date	
dc.Resource Type	
dc.Format	
dc.Resource Identi...	834.jpg

Existing values for dc.Creator

- Brian
- Brian Goetz
- David Holmes
- Joseppth Bowbeer
- Joshua Bloch
- Tim Peierls

Fig. 2. Filling metadata for an instance of the greenBookC collection

The create panel at the librarian interface of Greenstone is used to construct the greenBookC collection. The simple image collection (image-e) from the demo Greenstone collections was used as a template. The ImagePlugin of Greenstone processed all the images appropriately on a Windows XP operative system.

greenBookC has a web-based interface to search the images. Descriptors such as topic, author or the browse list can be used to query the images. Users can carry traditional keyword-based searches in different DC elements or use the metadata descriptions whose values comes from the ontology. Distinct languages can be used to query the collection. At the date, the work has been address to construct the semantic descriptions and empirically the descriptions have improve collection accessibility. However, a lot of effort is still required to evaluate the proposed approached.

6 Conclusions

The use of lightweight ontologies to enrich metadata standard is a basic but simple policy to organize images by content and an alternative to construct semantic descriptions. Ontologies and metadata enable information sharing between academic communities. Ontologies provide controlled vocabularies that help to reduce ambiguity from natural language or keyword based descriptions.

The visualization of book covers at Greenstone allow users to recognize visual features of book covers. The use of existing legal metadata enables information sharing and improves interoperability in digital libraries. Different metadata standards or additional elements can be used to store relevant information of book covers as the abstract or the synopsis of a book, the publisher and the physical location, respectively.

There are several challenges in the construction of collections for semantic digital libraries, however there are software tools that can help users to develop this task successfully. As future work, we plan to represent the classes, subclasses, individual, properties and restrictions of the ontology formally in order to integrate reasoning capabilities.

References

1. Allemang, D., Hendler, J.: Semantic web for the working ontologist. Morgan Kauffman Publishers (2011)
2. Kruk, S.R., McDaniel, B.: Semantic Digital Libraries. Springer-Verlag (2009)
3. Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: an architecture for complex objects and their relationships. International Journal on Digital Libraries V6(2), 124–138 (Apr 2006)
4. Ricardo Baeza-Yates, B.R.N.: Modern information retrieval. Addison Wesley (2011)
5. Wielinga, B.J., Schreiber, A.T., Wielemaker, J., Sandberg, J.A.C.: From thesaurus to ontology. In: Proceedings of the 1st international conference on Knowledge capture. pp. 194–201. K-CAP '01, ACM, New York, NY, USA (2001), http://doi.acm.org/10.1145/500737.500767
6. Witten, I.H., Bainbridge, D., Nichols, D.M.: How to Build a Digital Library. Morgan Kaufmann, Burlington, MA, 2. edn. (2010)

Acknowledgments

We thank to the staff library at the UPPuebla for their help and cooperation in the construction process. This work is partially supported by PROMEP grant Biblioteca Digital Semántica de Recursos Educativos /103-5/09/4023.