Building CORTUPP: a digital collection of technical reports with semantic features

Ma. Auxilio Medina, Argelia B. Urbina Nájera, Antonio Benitez R., J. de la Calleja, E. López D., Rebeca Rodríguez H.

Universidad Politécnica de Puebla
Tercer Carril del Ejido Serrano S/N
Juan C. Bonilla, Puebla, México
{mmedina, aurbina, abenitez, jdelacalleja, elopez, rrodriguez }
@uppuebla.edu.mx,
WWW home page: http://informatica.uppuebla.edu.mx/
~mmedina, ~aurbina, ~abenitez, ~jdelacalleja, ~elopezd, ~rrodriguez,

Abstract. The construction of a digital collection from the beginning implies technical decisions such as choosing format and design of documents, the selection of search and browsing mechanisms to access data and metadata, and the use of an architecture which support collaborative work of authors. This paper describes CORTUPP, a digital collection of technical reports. CORTUPP uses REC, an external service to support collaborative labeling and ranking of documents.

1 Introduction

Learning is a continuous process supported by daily activities; peer and student - teacher interactions enrich and accelerate this process [1]. Since 2004, the Universidad Politécnica de Puebla (UPPuebla) has adopted the competency based education model [2]. The paper describes our work in integrating educational resources into a digital collection of technical reports called CORTUPP. A technical report is a document that describes a research project or a technological solution, this is constructed as final works of students.

The paper is organized as follows. Section 2 briefly describes a semantic digital library. Section 3 explains the architecture of the collection. Section 4 describes the semantic features of CORTUPP. Section 5 explains how REC, an external service that supports collaborative labeling and ranking of documents is integrated to CORTUPP. Finally, Section 6 includes conclusions and suggests future directions of our work.

2 Semantic digital libraries

Semantic digital libraries refer to systems build upon research on digital libraries, semantic web, social networking and human computer interaction: they integrate knowledge organization systems, delivered by classic digital libraries, with the semantic web and social networking (Web 2.0) technologies [3].

Authors believe that semantic web technologies can support the development of valuable collections and services required in educational institutions. The particular interest is a semantic digital library, that according to [3] is formed by materials, tools and meanings. Some of the goals of a semantic digital library are the following ones:

- Anyone can use it
- Knowledge is accessible from the semantic digital library
- Resources are available with the modality anytime anywhere
- Friendly and multi-modal interfaces
- Multiple connected devises

Although freely distributed software exists around the world to construct semantic digital libraries such as Greenstone [1] or Jerome DL [2], we decided to implement an independent component in order to take into account the work flows implemented at the UPPuebla. Authors believe that CORTUPP can serve as a basis to construct a semantic digital library for the UPPuebla.

3 Architecture of CORTUPP

CORTUPP collection consists of a database, a web interface, assessment instruments, a common structure of documents and search mechanisms. This is available at `http://server3.uppuebla.edu.mx/cortupp/`. Figure 1 shows the architecture of our collection. This is an adaptation of an architecture proposed by [3].

The content of CORTUPP is formed by technical reports, registers of assessment committees, assessment instruments and calendar of activities. The data about users and count of users are also part of the content. The main users are teachers and students at the UPPuebla that make use of services of access, storage and search. Next sections describe the components of the architecture.

3.1 Structure of technical reports

We propose a common structure of the documents that introduces a common semantic by itself. This structure is formed by the following mandatory chapters: 1)research propose, 2)theoretical marc, 3)research design, 4)implementation, 5)results and 6)conclusions. Support material of the research project such as interviews, questionnaires or large tables can be added in appendixes.

Authors are free to propose the structure of each chapter, except the first one that refers to the research propose which is formed by the following mandatory sections: introduction, general objective, specific objectives, justification,

[1] http://www.greenstone.org/

[2] http://www.jeromedl.org/

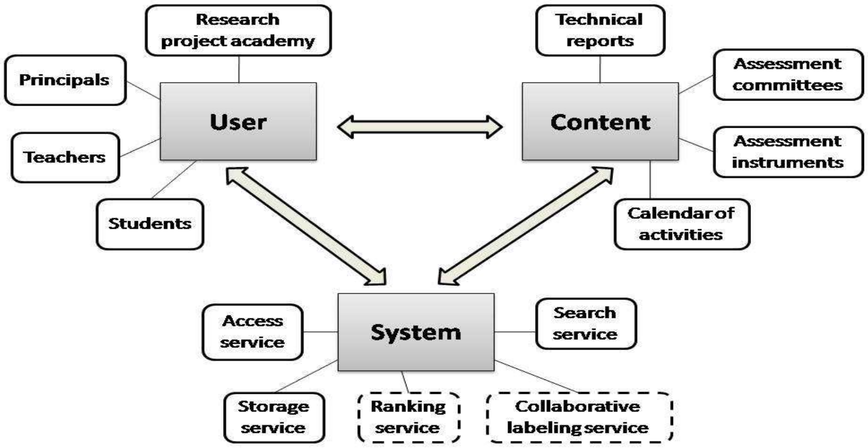


Fig. 1. Architecture of CORTUPP

chronogram of activities, hardware and software requirements and scopes and limits of the research project. The document structure has been defined as a Latex template. The BibTex file format is used to create the bibliography [3]. A technical report is described itself as a techreport entry. .

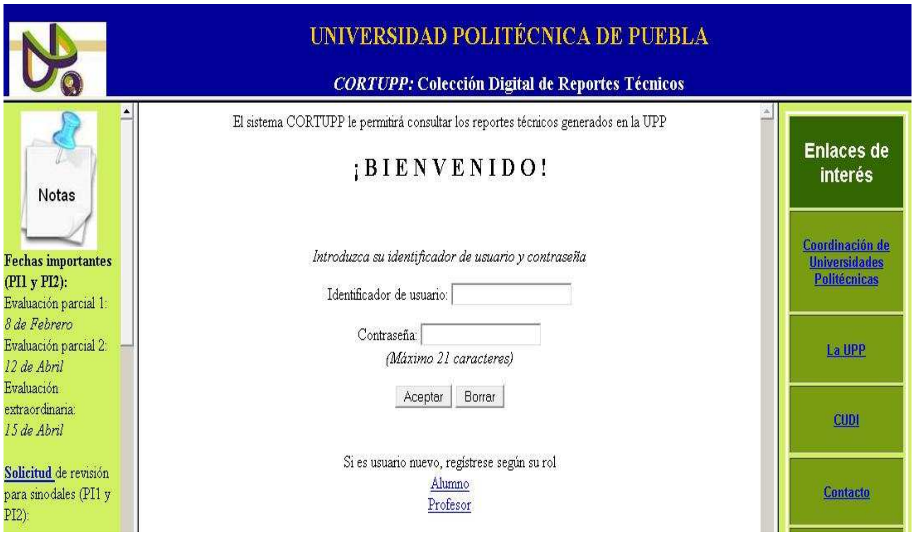
We identify internal users that belong to the UPPuebla community, they are students, teachers, managers and staff of the diffusion department; and external users who are members of another academic communities or visitors of the collection.

3.2 Keyword-based services

CORTUPP has a web-based interface to access the documents stored at the database, this makes use of hyperlinks to explore documents, to download the assessment instruments or to access to relevant web pages. Technical reports are stored as PDF files in the web server.

At CORTUPP users can carry on two types of searches: 1) keyword-based searches and 2) authority search (search by author or by a participant of the assessment committee). An assessment committee is formed by three teachers who play the role of advisor, secretary and vocal. This committee validates the content of the document. Figure 2 shows the interface of CORTUPP.

[3] http://www.kfunigraz.ac.at/ binder/texhelp/bibtx-7.html



UNIVERSIDAD POLITÉCNICA DE PUEBLA

CORTUPP: Colección Digital de Reportes Técnicos

El sistema CORTUPP le permitirá consultar los reportes técnicos generados en la UPP

¡BIENVENIDO!

Introduzca su identificador de usuario y contraseña

Identificador de usuario:

Contraseña:

(Máximo 21 caracteres)

Aceptar Borrar

Si es usuario nuevo, regístrese según su rol

Alumno

Profesor

Notas

Fechas importantes (PI1 y PI2):
Evaluación parcial 1:
8 de Febrero
Evaluación parcial 2:
12 de Abril
Evaluación extraordinaria:
15 de Abril

Solicitud de revisión para sinodales (PI1 y PI2):

Fig. 2. Interface of CORTUPP

4 Semantic features

CORTUPP uses existing legal metadata in semantically enabled libraries. Technical reports are described with the Dublin Core (DC) elements of Table 1. These elements are associated to the elements of the Latex template.

Table 1. DC elements used to describe a technical report

DC element	Description
`Creator`	Indicates the name of the first author
`Date`	Indicates the delivery date
`Description`	Contains the abstract of the technical report
`Identifier`	This is a number used to identify the technical report in the collection
`Language`	Language of the content (Spanish)
`Publisher`	Contains the name of the university as the entity responsible for publishing the technical report
`Subject`	Keywords of the technical report according to a research area
`Title`	A given title to the technical report

CORTUPP is represented in a structure called ontology of records that maintain an organization by content. This is a hierarchical structure that provides a unique and unambiguous interpretation of the document elements. This has concept-term relationships useful for search based on free text. The main characteristics of an ontology of records are the following ones:

1. Technical reports are clustered by similarity
2. Clusters in the k-level have labels of k-terms
3. All documents of a cluster share the terms of its label

The features of the ontology of records can be found in [4]. Then, semantic information is represented by metadata attached to each document and by the ontology of records. CORTUPP design corresponds to the levels of knowledge proposed by [3]:

1. Organization of the information in databases
2. Organization of the information in the documents
3. Organization of the metadata
4. Organization of the topics treated in the documents
5. Organization of the concepts, terms and relations

5 REC: an external service with semantic features

Adding semantic features for digital collections is a topic of interest in research areas such as collaborative labeling, web 2.0 and semantic digital libraries. For example, [5] describes the potential of tagging systems to support knowledge organization or [6] investigate social book marking in digital libraries and derive the design requirements to incorporate social book marking.

A tag is a keyword that acts like a subject or category for the associated content [3]. Tags are user added metadata, tagging is the establishment of a relationship between an online information resource and a user.

In social contexts, such as Flickr [4], facebook [5], del.icio.us. [6] and Soboleo [7], traditional measures of information retrieval are not important, else the opinion and experience of previous users. In this sense, we have decided to integrate REC, an open software that allow users of CORTUPP to add tags to take into account subjective information of users.

REC [7] makes use of the "induced tagging" technique design to improve the quality of automatic markers. It offers *collaborative labeling* through the resulting tags that produce recommendations and a *ranking documents* service where labels are useful for helpful content recommendation.

REC allow domain experts and members of the community to assign meaning labels to the technical reports of CORTUPP. Using REC, users construct a different organization of the documents. The integration of REC to CORTUPP makes it a community information space through functionality for selection, annotation, authoring/contribution and collaboration.

[4] http://www.flickr.com/

[5] http://www.flickr.com/

[6] http://delicious.com/

[7] http://www.soboleo.com/

6 Conclusions

CORTUPP allows users to reuse the content of documents, the collection integrates research activities of students under the supervision of a team of teachers. This has the following advantages: distribution of assessment instruments, extension of document descriptions with labels. However, there are several challenges in the construction of a semantic digital library such as providing for more usability and inference mechanisms. At the date, CORTUPP can be perceived as a result of collaborative content production at the UPPuebla.

Currently, search services at CORTUPP are keyword-based. As future work, we plan to expand those services in order to have semantic search engines, such engines can be used to improve the quality of keyword-based search engines by taking into account the meaning of the words. We conclude with further possibilities of organization and recommendation that arise from the use of REC.

References

1. I., H.: Role of information technologies in teaching learning process: perception of the faculty. Turkish Online Journal of Distance Education - TOJDE **9**(2) (2008)
2. Lindley, W.I.: Constraints and potentials of training mid-career extension profesionals in africa, part 2 (1999)
3. Kruk, S.R., McDaniel, B.: Semantic Digital Libraries. Springer-Verlag, Berlin, Heidelberg (2009)
4. Medina, M.A., Sánchez, J.A.: Ontoair: A method to construct lightweight ontologies from document collections. In: ENC '08: Proceedings of the 2008 Mexican International Conference on Computer Science, Washington, DC, USA, IEEE Computer Society (2008) 115–125
5. Li, Q., Lu, S.C.Y.: Collaborative tagging applications and approaches. IEEE Multimedia **15** (2008) 14–21
6. Puspitasari, F., Lim, E.P., Goh, D.H.L., Chang, C.H., Zhang, J., Sun, A., Theng, Y.L., Chatterjea, K., Li, Y.: Social navigation in digital libraries by bookmarking. In: ICADL'07: Proceedings of the 10th international conference on Asian digital libraries, Berlin, Heidelberg, Springer-Verlag (2007) 297–306
7. Sánchez, J.A., Arzamendi-Pétriz, A., Valdiviezo, O.: Induced tagging: promoting resource discovery and recommendation in digital libraries. In: JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, New York, NY, USA, ACM (2007) 396–397

Acknowledgments

We thank to the staff of Programa Académico de Ingeniería en Informática 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.