Community Tales

An Infrastructure for the Collaborative Construction of Digital Theses Repositories

Lourdes Fernandez Ramirez
J. Alfredo Sánchez
Universidad de las Américas, Puebla
{*lulu, alfredo*}*@mail.udlap.mx*
Cholula Puebla, Mexico
biblio.udlap.mx/tesis

Keywords: Requirements, repository construction, interoperability

Abstract

"Tales" (pronounced tä'lez in Spanish) is the codename for the ongoing project aimed at the construction of the digital theses collection and associated services at Universidad de las Américas, Puebla (UDLA). UDLA is the first university in Mexico to approve a digital thesis requirement for all its academic programs and only one out of three in Latin America that have similar projects. We would like to share how our project has evolved from an early prototype onto a full-fledged SYSTEM involving automated upload facilities and complex search, navigation and federation mechanisms. The Libraries Division started the Tales project in the fall of 1999, when one pilot academic department committed to requiring digital theses and to incorporate them into the repository. Today, Tales implements an institutional policy, which requires that all university departments participate in some way in the construction of the repository. Our thesis repository participates in NDLTD and the OAI initiative.

Introduction

Digital libraries (DLs) are changing radically the ways in which traditional libraries managed their resources and offered services to users. DLs allow for the development of new ways of representing, using, generating and disseminating knowledge. They also promote new work practices and new ways of expressing and solving problems. We have conducted work on a university-wide project to produce a repository of digital theses and a wide range of retrieval and navigation services. This work is part of a long-term digital libraries program. In this paper, we describe how the introduction of a digital theses requirement is changing work practices and opening new opportunities for collaboration among students, academic departments and librarians inside and outside our institution.

University Digital Libraries for All (U-DL-A)

Our work is framed by an ambitious digital libraries program we have termed University Digital Libraries for All (U-DL-A), documented elsewhere [Sánchez 2001, Sánchez and Arias 2003]. U-DL-A has produced a digital libraries environment that now comprises various collections, services and user interfaces. One of our most important repositories is the collection of digital theses, which has observed a significant growth starting the current semester. Other collections include the university publications and historic archives in the libraries's special collections. Services vary from information retrieval methods to agent services to navigation mechanisms. Finally, user interfaces include personal and group spaces, visualization aids and user agents (publications concerning this work can be found at [ICT 2003]). Our work on digital theses was initiated in the context of the Networked Digital Library of Theses and Dissertations initiative (NDLTD) [Fox et al. 2001].

Theses at UDLA

Every year, about 900 theses are generated by graduating students of Universidad de las Américas, Puebla (UDLA), out of which approximately 10% are produced by graduate students (the education SYSTEM in Mexico requires a thesis for all the five-year bachelor's degree programs). Due to space limitations, UDLA's library has received and maintained hardcopies only of graduate theses. Printed and bound copies of undergraduate theses have been made available via the academic departments. Under these circumstances, finding theses on a specific topic can be a strenuous and often frustrating task, as departments do not usually keep their catalogs up to date, loan policies vary and are seldom enforced, and office hours and storage space are typically limited.

We thus set out to build *Tales*, the technological and organizational environment that facilitates the dynamic, collaborative construction of the digital theses component of our digital library. Among the advantages we have kept in mind as we developed Tales, the following have been key to foster user involvement:

- Availability. Theses can be made available for the entire university community or to the general public. The digital collection does not have restrictions as for number of available copies, physical space or office hours.
- Electronic storage. It is no longer necessary for the departments to store physically the documents. Some departments do not have enough space for storing them or personnel for their administration.
- Cost reduction. Student save money typically spent for printing and binding.

- Searching and browsing facilities. Ease of retrieval and navigation are but two of the advantages of the digital substrate that make digital theses appealing to be used for supporting new research endeavors.

Tales has evolved significantly from its inception as a prototype to its current university-wide deployment, which has allowed for the establishment of a digital theses requirement for all the institution's academic programs.

In the remainder of this document, we briefly describe this evolution and the main features of the resulting environment.

Tales: From prototype to deployment

The Tales project started in 1999 as one of the components of our digital libraries program. Tales aimed at contributing to the construction of digital collections in Spanish and serving as a platform for research of open issues in digital libraries. Our approach was to first demonstrate the benefits and advantages of creating a digital collection while developing and testing tools to support the workflow and general browsing and searching services. We invited one academic department to function as pilot group. The process began with the students, who submitted their theses on floppy or compact disks, in at least one of the allowed formats (MS-Word, HTML and LaTeX). We selected HTML as display and unification format. Every submitted thesis was converted into this format. Documents were parsed and stored according to the ETD-ML definition [Kipp et al. 1997]. Once each element was identified and stored we were able to reconstruct the thesis on demand to support various navigation and searching mechanisms. After two semesters of successful operation, we also invited our master's programs to participate in the effort. Our collection grew modestly but this stage allowed our group to concentrate on the development of required facilities to convert among formats, to parse and to upload documents and generally to gain experience and to work with other areas in the definition of policies and procedures for a wider use of Tales.

During the pilot operation, we built facilities for automating various processes and we also explored interoperability issues in the context of NDLTD and the Open Archives Initiative. Finally, the digital thesis requirement was approved by the academic council for all the academic programs starting the 2003 Spring Semester.

UDLA is the first university in Mexico to establish a digital thesis requirement for all its academic programs and, at the time of this writing, only one out of three in Latin America that have similar projects. As the volume of theses to be handled every semester has increased, it has been necessary to design facilities for allowing all involved parties to collaborate in the construction of the theses collection. Thus, responsibilities have been assigned to participants as follows:

- The registrar's office provides course and student listings related to thesis projects
- Academic departments update thesis committees and also validate that the requirement has been satisfied prior to graduation,
- Students upload documents,
- Advisors review and approve theses for publication in one of various available modalities, and
- Library personnel provides technical support, maintains the collection and runs processes to make the documents available via the web.

Tales conceptual design

Tales has been designed to facilitate collaboration throughout the theses process, from document submission to publication and use, including functionality for data and information retrieval, browsing and searching services and interoperability with other collections. Every semester, a setup stage feeds databases with information on students, courses and professors. Academic departments or theses coordinators define the composition of thesis committees. Student submit their documents and advisors may approve the document for publication. Each of the participants interact with Tales via web interfaces customized according to their roles. We selected HTML as storage format though we are in the process of moving to XML. At present, the display format could be HTML or PDF (generated from HTML). Additional issues involved in the design of the Tales environment are provided below.

Data Model

Two aspects were considered when designing Tales' data model: integration with other U-DL-A components and the structure of the theses themselves. The latter was based on the data type definition for Electronic Theses and Dissertations (ETD-ML). Only those elements relevant for our project were selected considering thesis regulations for our university. As for its incorporation into U-DL-A, the roles played by users had to be modeled along with data such as majors, academic departments, and organizational relationships. The Dublin Core metadata has been considered.

Authors

All graduating students are registered as authors in the Tales environment. They are required to submit each section of their documents (preliminary pages, tables of contents, chapters, appendices, references and so on) in two formats: PDF and its source format (typically MS Word). Authors are allowed to update the descriptive metadata of their thesis (title, date and place of exam defense, email). Files can be uploaded individually or as a compressed package. Students are notified by email whether the process has been completed successfully or corrections are needed.

Academic departments

Academic departments' administrators are in charge of registering all the evaluation committees for the students of all majors associated to the department. Initially this task is assigned to the instructor in charge of coordinating all thesis students. Other users may be added to do this task.

Thesis advisors

Once authors have submitted their theses, advisors are also notified. Advisors are expected to revise the submission and make sure authors submitted a complete and authorized version of the document. After revising and approving the thesis, the advisor will specify the access level for the thesis: *restricted* (only advisor and authors can see it); *local* (for campus only) or *global* (open to the general public). Once this has been done, the advisor releases the thesis for publication.

Library personnel

Once a thesis is deemed complete, it becomes part of the collection and its contents become accessible to users as specified by the thesis' advisor. The digital librarian initiates additional processes for structural recognition so tables of contents and indices that will facilitate content-based queries are constructed.

User Services

We decided to provide access to the collection through two main interfaces: one for searching and another for browsing. Additionally, we aimed to provide transparent access to other collections in a federation.

Two basic types of searching mechanisms were considered for Tales. The first type considers data retrieval, which entails direct matching of query terms with specific text in the database; the second type involves using mechanisms from the field of Information Retrieval (IR). In this respect we have taken advantage of ongoing developments in the context of U-DL-A, such as a server and library of information retrieval models.

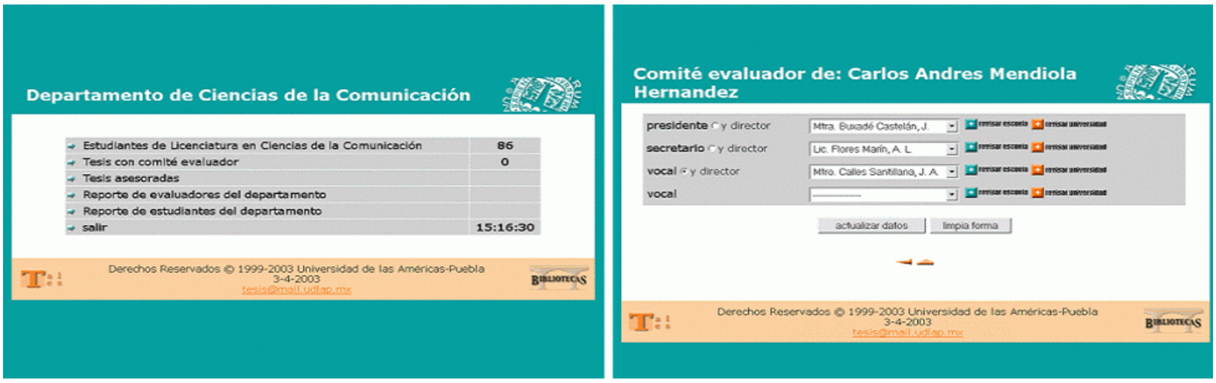
As for navigation means, we opted for providing access through lists of theses references linked to the contents, which are properly segmented for easy browsing. Each page must contain the complete bibliographic reference and links to other sections of the thesis to facilitate navigation and references to specific sections of the document. This is of particular importance if we consider that search engines can find any portion of a thesis, so every page must give contextual information. Segmenting also reduces downloading time and helps users to view documents quickly.

From its inception, we conceived of our digital thesis effort as a participant of a growing international community. We adhered to the NDLTD project and devised mechanisms for integration under the guidelines of the Open Archives Initiative (OAI).

Implementation

All the tools were implemented using Java. Access to services only requires a web browser, as interfaces were implemented as Java Servlets. Display of PDF files requires Adobe's Acrobat Reader. Our database management SYSTEM is MySQL, which is accessed by means of other services developed within U-DL-A.

Let us illustrate how the Tales environment facilitates the construction of a collection of digital thesis. The process begins with the academic department's administrators. They are responsible for registering the evaluation committees. Figure 1 shows the SYSTEM options available for these users. On the left side, the main options include registering committees, updating committee's data, changing participants' roles, and viewing the department's data (professors and students). The right side shows how the registration of the committee can be done: the selection boxes include all the department's professors; extra buttons allow for adding professors from other departments or from other institutions.



Departamento de Ciencias de la Comunicación

Estudiantes de Licenciatura en Ciencias de la Comunicación	86
Tesis con comité evaluador	0
Tesis asesoradas	
Reporte de evaluadores del departamento	
Reporte de estudiantes del departamento	
salir	15:16:30

Derechos Reservados © 1999-2003 Universidad de las Américas-Puebla
3-4-2003
tesis@mail.udlap.mx

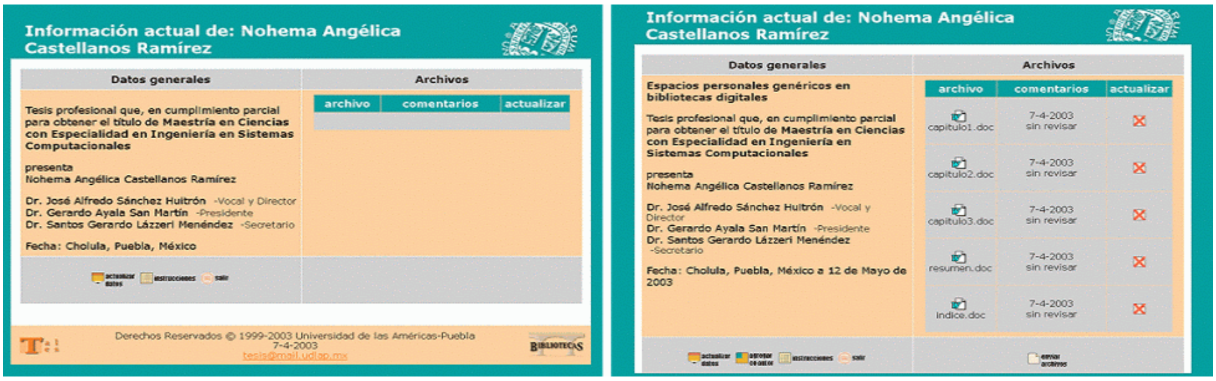
Comité evaluador de: Carlos Andres Mendiola Hernandez

presidente ○ y director	Mtra. Buxadé Castelán, J.	revisar escuela revisar universidad
secretario ○ y director	Lic. Flores Marín, A. L.	revisar escuela revisar universidad
vocal ● y director	Mtro. Calles Santillana, J. A.	revisar escuela revisar universidad
vocal	--------------	revisar escuela revisar universidad

Derechos Reservados © 1999-2003 Universidad de las Américas-Puebla
3-4-2003
tesis@mail.udlap.mx

Figure 1: Academic department's interface.

Students are expected to update their general information (email address, thesis title, place and date of the exam), may add coauthors and submit their files. As shown in Figure 2, files are uploaded in groups of five; compressed and packaged files are also accepted.



Información actual de: Nohema Angélica Castellanos Ramírez

Datos generales

Tesis profesional que, en cumplimiento parcial para obtener el título de Maestría en Ciencias con Especialidad en Ingeniería en Sistemas Computacionales

presenta
Nohema Angélica Castellanos Ramírez

Dr. José Alfredo Sánchez Huitrón -Vocal y Director
Dr. Gerardo Ayala San Martín -Presidente
Dr. Santos Gerardo Lázzeri Menéndez -Secretario

Fecha: Cholula, Puebla, México

Archivos

archivo	comentarios	actualizar

Derechos Reservados © 1999-2003 Universidad de las Américas-Puebla
7-4-2003
tesis@mail.udlap.mx

Información actual de: Nohema Angélica Castellanos Ramírez

Datos generales

Espacios personales genéricos en bibliotecas digitales

Tesis profesional que, en cumplimiento parcial para obtener el título de Maestría en Ciencias con Especialidad en Ingeniería en Sistemas Computacionales

presenta
Nohema Angélica Castellanos Ramírez

Dr. José Alfredo Sánchez Huitrón -Vocal y Director
Dr. Gerardo Ayala San Martín -Presidente
Dr. Santos Gerardo Lázzeri Menéndez -Secretario

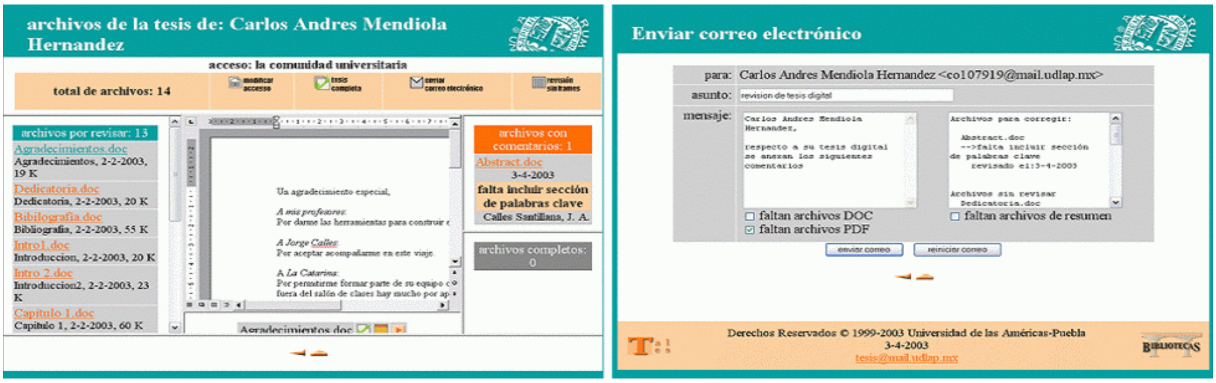
Fecha: Cholula, Puebla, México a 12 de Mayo de 2003

Archivos

archivo	comentarios	actualizar
capitulo1.doc	7-4-2003 sin revisar	☒
capitulo2.doc	7-4-2003 sin revisar	☒
capitulo3.doc	7-4-2003 sin revisar	☒
resumen.doc	7-4-2003 sin revisar	☒
indice.doc	7-4-2003 sin revisar	☒

Figure 2: Authors' interfaces.

Once an author has uploaded a thesis, the advisor will receive a notification. When advisors enter the SYSTEM; a list of studens and the number of files that have been uploaded will be displayed. Typically, at this point the student and the advisor have agreed on a final version of the document after several revisions (which can be done either electronically or on paper); the advisor's task is to verify whether the submitted version corresponds to the most recent version of the document. The interface used by advisors is shown in Figure 3. The advisor may review each of the received files, add comments or mark the document as complete and specify the desired access type. An e-mail message is sent to the author upon completion of the review process. On the left window, the files are separated into three sections: *pending revision*, *with comments* and *complete*. On the right window, an interface is presented that allows advisors to send an e-mail message to authors, including all comments on the submission. Advisors may add or remove comments as desired.



archivos de la tesis de: Carlos Andres Mendiola Hernandez

acceso: la comunidad universitaria

total de archivos: 14

archivos por revisar: 13

Agradecimientos.doc
Agradecimientos, 2-2-2003, 19 K
Dedicatoria.doc
Dedicatoria, 2-2-2003, 20 K
Bibilografia.doc
Bibliografia, 2-2-2003, 55 K
Intro1.doc
Introduccion, 2-2-2003, 20 K
Intro 2.doc
Introduccion2, 2-2-2003, 23 K
Capitulo 1.doc
Capitulo 1, 2-2-2003, 60 K

Un agradecimiento especial,

A mis profesores:
Por darme las herramientas para construir e

A Jorge Calles:
Por aceptar acompañarme en este viaje.

A *La Catarina:*
Por permitirme formar parte de su equipo c
fuera del salón de clases hay mucho por ap

Agradecimientos.doc

archivos con comentarios: 1

Abstract.doc
3-4-2003
falta incluir sección de palabras clave
Calles Santillana, J. A.

archivos completos: 0

Enviar correo electrónico

para: Carlos Andres Mendiola Hernandez <co107919@mail.udlap.mx>
asunto: revision de tesis digital
mensaje:

```
Carlos Andres Mendiola
Hernandez,

respecto a su tesis digital
se anexan los siguientes
comentarios
```

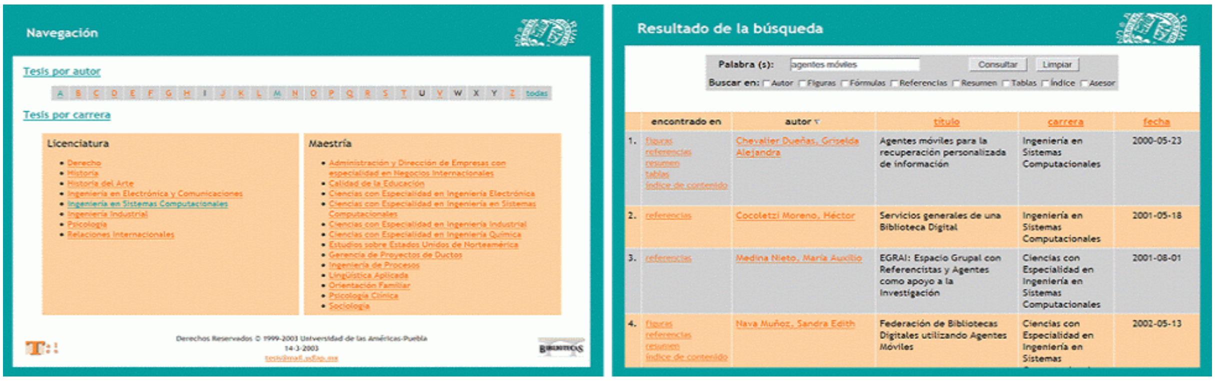
```
Archivos para corregir:

  Abstract.doc
  -->falta incluir sección
de palabras clave
    revisado el:3-4-2003


Archivos sin revisar
  Dedicatoria.doc
```

□ faltan archivos DOC
☑ faltan archivos PDF
□ faltan archivos de resumen

Derechos Reservados © 1999-2003 Universidad de las Américas-Puebla
3-4-2003
tesis@mail.udlap.mx

Figure 3: Advisors' interfaces.

Figure 4 shows the main browsing interface for the results of a search process. The browsing interface is divided into three sections. The top section shows initials linked to authors' last names and their theses. The second section lists theses from undergraduate majors whereas the third section lists theses produced by graduate students. Search results may be ordered by author, title, major or year. Since search is possible practically on any portion of the thesis (figures, formulae, references, etc.), the first column indicates where in the document terms were found and links that can be followed on to that section.



Navegación

Tesis por autor

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z todas

Tesis por carrera

Licenciatura

- Derecho
- Historia
- Historia del Arte
- Ingeniería en Electrónica y Comunicaciones
- Ingeniería en Sistemas Computacionales
- Ingeniería Industrial
- Psicología
- Relaciones Internacionales

Maestría

- Administración y Dirección de Empresas con especialidad en Negocios Internacionales
- Calidad de la Educación
- Ciencias con Especialidad en Ingeniería Electrónica
- Ciencias con Especialidad en Ingeniería en Sistemas Computacionales
- Ciencias con Especialidad en Ingeniería Industrial
- Ciencias con Especialidad en Ingeniería Química
- Estudios sobre Estados Unidos de Norteamérica
- Gerencia de Proyectos de Ductos
- Ingeniería de Procesos
- Lingüística Aplicada
- Orientación Familiar
- Psicología Clínica
- Sociología

Derechos Reservados © 1999-2003 Universidad de las Américas-Puebla
14-3-2003
tesis@mail.udlap.mx

Resultado de la búsqueda

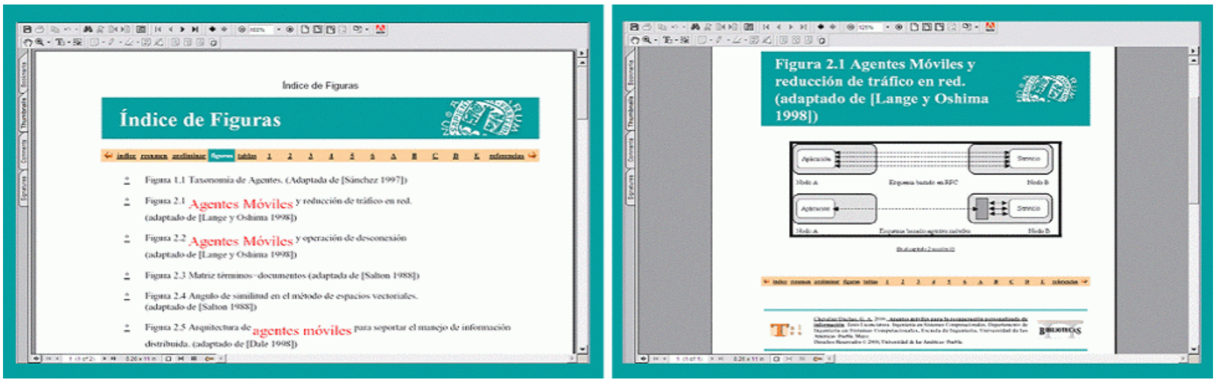
Palabra (s): agentes móviles Consultar Limpiar

Buscar en: □ Autor □ Figuras □ Fórmulas □ Referencias □ Resumen □ Tablas □ Índice □ Asesor

	encontrado en	autor	titulo	carrera	fecha
1.	figuras referencias resumen tablas índice de contenido	Chevalier Dueñas, Griselda Alejandra	Agentes móviles para la recuperación personalizada de información	Ingeniería en Sistemas Computacionales	2000-05-23
2.	referencias	Cocoletzi Moreno, Héctor	Servicios generales de una Biblioteca Digital	Ingeniería en Sistemas Computacionales	2001-05-18
3.	referencias	Medina Nieto, María Auxilio	EGRAI: Espacio Grupal con Referencistas y Agentes como apoyo a la Investigación	Ciencias con Especialidad en Ingeniería en Sistemas Computacionales	2001-08-01
4.	figuras referencias resumen índice de contenido	Nava Muñoz, Sandra Edith	Federación de Bibliotecas Digitales utilizando Agentes Móviles	Ciencias con Especialidad en Ingeniería en Sistemas	2002-05-13

Figure 4: Browsing and searching in Tales.

When a thesis is selected from the list, the corresponding section will be displayed as a PDF file highlighting query terms contained in the document. This is exemplified in Figure 5. Except for the table of contents and the cover page, which are displayed in HTML, all other sections are displayed using the PDF files. For the generated PDFs, a navigation bar and next/back buttons allow users to move to any section of the document.



Índice de Figuras

Índice de Figuras

índice resumen preliminar figuras tablas 1 2 3 4 5 6 A B C D E referencias

- Figura 1.1 Taxonomía de Agentes. (Adaptada de [Sánchez 1997])
- Figura 2.1 Agentes Móviles y reducción de tráfico en red. (adaptado de [Lange y Oshima 1998])
- Figura 2.2 Agentes Móviles y operación de desconexión (adaptado de [Lange y Oshima 1998])
- Figura 2.3 Matriz términos–documentos (adaptada de [Salton 1988])
- Figura 2.4 Angulo de similitud en el método de espacios vectoriales. (adaptado de [Salton 1988])
- Figura 2.5 Arquitectura de agentes móviles para soportar el manejo de información distribuida. (adaptado de [Dale 1998])

Figura 2.1 Agentes Móviles y reducción de tráfico en red. (adaptado de [Lange y Oshima 1998])

índice resumen preliminar figuras tablas 1 2 3 4 5 6 A B C D E referencias

Figure 5: Viewing a digital thesis in Tales.

We have implemented federation mechanisms that allow users to retrieve documents from other collections in response to queries posed in Tales. Specifically, we have implemented mobile agents that traverse the network and gather documents from two partner institutions (technical reports from another institution in Mexico and theses from Virginia Tech in the US) and present them to the user in a transparent fashion [Sánchez et al. 2002]. A server has been built that is compliant with the OAI Protocol for Metadata Harvesting (OAI-PMH) and allows for external clients to query Tales.

As for IR facilities, we have successfully tested our Hermes server and IR model library [Maldonado-Naude and Sánchez 2003] using Tales, in such a way that ranked results can be obtained using the vector space, extended boolean and latent semantic indexing models.

Work in progress

We expect our collection to include about 600 theses by the end our Summer Semester. At present, theses contain mostly text and images but we are encouraging the use of media such as audio, video and three-dimensional models. We are also exploring alternatives for data description and identification and we now have a parallel development using XML. We plan to complete the integration of Tales and our IR facilities this year. Also, other services for digital collections are in progress.

We have developed a tool named *Poseidon* [Sanchez and Flores 2002], which makes it possible to review and annotate digital documents. The advisor and other committee make comments on the thesis using this tool. Students review annotations and update the document.

Once the thesis is completed, it is automatically integrated to the collection. The integration of this tool into Tales is also planned for this year.

One recurring concern among Tales users is plagiarism. For this reason, we have started the development of tools for detecting similarity between documents also using our IR server. This tool considers comparing internal documents (from our collection) as well as external thesis and web pages.

Considering the growing volume of images that are comprised by Tales, we have started a project for content-based queries on images in our theses collection.

Conclusions

Tales is a collaborative environment that is facilitating the construction of a repository of digital thesis. Though a stable state of Tales should be reached in about a year after the approval of a digital thesis requirement, it is clear that it is contributing to changing work practices and opening new possibilities for expressing and disseminating knowledge. Still at the early stages of the collection, we are getting around 140 visitors per day at *http://biblio.udlap.mx/tesis*. Statistics show that that most of our visitors come from Latin America. We believe Tales will contribute to make knowledge more accessible to wider user communities and will be helpful as a platform for research in open issues in digital libraries.

Acknowledgements

This work is partly supported by the Mexican National Council of Science and Technology (Conacyt Projects No. 35804-A and G33009-A).

Bibliography

1. SÁNCHEZ, J. A., FLORES, L. A.: Provisions for collaborative revision and annotation of digital documents Poster at the ACM 2002; Conference on Computer-Supported Cooperative Work, ACM Press 2002
2. FOX, EATON, J. L., MCMILLAN, G., KIPP, N. A., MATHER, P., PHANOURIOU, C.: Networked Digital Library of Theses and Dissertations, www.ndltd.org/
3. KIPP, N. A., FOX, E., EATON, J.L., MCMILLAN, G., ARCE, J.E.: Document Type Definition for Electronic Theses and Dissertations, etd.vt.edu/etd-ml/dtdetds.htm
4. SÁNCHEZ, J. A.: HCI and CSCW in the context of digital libraries.; Proc. Conference on Human Factors in Computing Systems, ACM Press 2001, 63-64
5. MALDONADO-NAUDE, F., SÁNCHEZ, J. A.: Using Hermes-F: Experiences with a framework for developing information retrieval applications.; Proc. Encuentro de Ciencias de la Computación, IEEE comp., Soc. Press 2003
6. SÁNCHEZ, J. A., ARIAS, J. A.: Fourth-phase digital libraries: Pacing, linking, annotating and citing in multimedia collections.; Proc.of the Joint Conference on Digital Libraries, ACM Press/IEEEE, Comp. Soc. Press 2003
7. SÁNCHEZ, J. A., NAVA MUÑOZ, S., FERNÁNDEZ RAMÍREZ, L., CHEVALIER DUEÑAS, G.: Distributed information retrieval from web-accessible digital libraries using mobile agents. Upgrade, Special Issue on Information Retrieval and the Web 3, 2002, 2, www.upgrade-cepis.org
8. ICT: Publications of the Laboratory of Interactive and Cooperative Technologies., 2003, ict.udlap.mx/pubs