Technology and Research in a Global Networked University Digital Library (NUDL)

Marcos André Gonçalves
Ph.D. student in the Digital Library Research Laboratory at Virginia Tech. He works on aspects of digital library theory, models, architectures and interoperability issues.

Edward A. Fox
director of NDLTD, is a Professor of Computer Science at Virginia Tech

Resumo

As Bibliotecas Digitais (DLs) encerram sistemas extremamente complexos de informação que suportam a criação, gestão, distribuição e preservação de fontes complexas de informação, enquanto permitem uma interação eficaz e eficiente entre as diversas sociedades que se beneficiam do conteúdo e serviços da Biblioteca Digital. Neste trabalho, centralizamo-nos sobre nossa experiência defrontando desafios de construção, manutenção e desenvolvimento de uma Biblioteca Digital Universitária em Rede (www.nudl.org), uma extensão da Biblioteca Digital em Rede de Teses e Dissertações (www.ndltd.org). NUDL é uma iniciativa de âmbito mundial que visa a tornar a propriedade intelectual produzida em universidades mais acesssível, estimulando a colaboração internacional entre todas as disciplinas. Damos os detalhados aspectos tecnológicos de nossas soluções e atividades de pesquisa realizados para proporcionar sólidos e enriquecidos serviços para as comunidades servidas por esta iniciativa.

Palavras-chave

Bibliotecas digitais; Sistemas de informação; Biblioteca digital universitária.

Technology and Research in a Global Networked University Digital Library (NUDL)

Abstract

Digital Libraries (DLs) are extremely complex information systems that support the creation, management, distribution, and preservation of complex information resources, while allowing effective and efficient interaction among the several societies that benefit from DL content and services. In this paper, we focus on our experience facing challenges of building, maintaining, and developing the Networked University Digital Library (www.nudl.org), an extension of the Networked Digital Library of Theses and Dissertations (www.ndltd.org). NUDL is a worldwide initiative that addresses making the intellectual property produced in universities more accessible, stimulating international collaboration across all disciplines. We detail technological aspects of our solutions and research activities carried out to provide powerful and enriched services for the communities served by this initiative.

Keywords

Digital libraries; Information systems; University digital library.

1. INTRODUCTION

Digital libraries (DLs) are among the most complex information systems. The practical and research field of digital libraries is inherently multidisciplinary and involves a number of related fields including database management, distributed systems, hypertext, human-computer interaction, information retrieval, information science, and multimedia services.

The term "digital library" first became widely used in the early 1990s. NSF-supported workshops helped the growing number of interested parties to identify key areas of research (Fox 1993). By 1993 it was clear that a new "hot topic" had emerged (Fox 1993). Now there are books available to provide overviews (Lesk 1997; Arms 2000). Journals have covered the field with special issues (Fox and Lunin 1993; Fox, Akscyn et al. 1995; Fox and Marchionini 1998; Marchionini and Fox 1999; Fox and Marchionini 2001). There are yearly conferences in Europe (ECDL series), Asia (ICADL series), as well as broad international events (Fox and Marchionini 1996; Fox and Rowe 1999; Borgman and Fox 2001). Overview chapters (Fox and Sornil 1999) can be found in books on related fields like information retrieval as well as in encyclopedia on technology (Fox and Sornil 2000). There also are WWW sites with references, tutorial materials, and online quizzes (Fox 1998).

Digital libraries can be justified by their benefits and advantages as compared with their physical counterparts; e.g., 1) timely and wide availability of up-to-date, high quality multimedia resources which help to remove physical and conceptual barriers; 2) network connectivity and interactive technologies which extend social interactions by allowing the creation of rich virtual workplaces; 3) digital technology, allowing the building of advanced and innovative services, many of which have features not possible or extremely difficult or expensive to implement in traditional libraries. These include, for example, powerful and effective search engines; new methods of visualization and interaction with information (Rao, Pedersen et al. 1995); complex interconnections of information through hypertext and automatic bibliographical cross-reference (Giles, Bollacker et al. 1998); and automatic dissemination of information, recommendations, and personalization (Rieken, 2001).

An excellent example of an extremely complex DL is the Networked University Digital Library – NUDL (http://www.nudl.org). NUDL aims to make the intellectual property produced in universities more accessible, stimulating technology transfer, international collaboration, and knowledge sharing across all disciplines. NUDL builds upon the foundations of the Networked Digital Library of Theses and Dissertations – NDLTD (http://www.ndltd.org), an international initiative to improve graduate education, by allowing students to produce electronic documents (electronic theses and dissertations, ETDs), use and contribute to DLs, and understand issues in electronic publishing. NUDL expands these objectives to:

1) make the (currently often only locally available) work of scholars writing in diverse languages more widely used;

2) link with related efforts, like NCSTRL (http://www.ncstrl.org) and CoRR (http://xxx.lanl.gov/archive/cs/intro.html) for reports/preprints; NSDL (http://www.nsdl.nsf.org, formerly, SMETE-lib), for Science, Mathematics, Engineering, and Technology Education; and the Open Archives Initiative (http://www.openarchives.org, OAI), for interoperability; and

3) prepare the next generation of scholars as effective knowledge workers, so they will learn about DLs through personally submitting their own works.

Besides ETDs, other NUDL content includes courseware, demonstrations, personal publications, portfolio, etc.

In this paper, we focus on the challenges of launching, building, maintaining, and developing the Networked University Digital Library (NUDL). This paper is organized as follows. Section 2 describes NDLTD in terms of tools and artifacts developed for building local and institutionalized ETD DLs. Section 2 also details the technical infrastructure put in place to provide NDLTD global services for all members. Section 3 covers other facets of NUDL, including courseware and discipline-oriented efforts. Section 4 presents a number of NUDL related research activities underway at the Virginia Tech Digital Library Research Laboratory (VT-DLRL), regarding searching, DL theory, and personalization. Section 5 concludes the paper.

2. NDLTD – A GLOBAL DIGITAL LIBRARY OF ELECTRONIC THESES AND DISSERTATIONS

NDLTD, in contrast with many other digital libraries, is focused on graduate students and universities activities. NDLTD increases the sharing of knowledge and availability of student research for scholars, and allows preserving it electronically. It lowers costs of submitting and handling theses and dissertations, helps universities build their information infrastructure, and empowers them to unlock their information resources. It endows students with capabilities to convey a richer message through the use of multimedia and hypermedia technologies and provides a better grounding to the "information age". With over 120 members, varying from a range of individual universities, libraries, and consortia at the state (OhioLINK), regional (Catalunya), and national (Australia, Germany, Portugal) levels, NDLTD ultimately extends the value and use of digital libraries and contributes to the advance of that technology.

ETD activities in general, and NDLTD organization, progress, growth, and evolution in particular, are described in a sourcebook on ETDs (Fox, Moxley et al. 2002)[*]. A guide to help those interested in the ETD initiative, funded in part by UNESCO, is available online and will be translated into many languages (Moxley 2001). In a summary assessment (Fox, Gonçalves et al. 2002)[**], we evaluate NDLTD under numerous human, system, and society-centered criteria. Those include membership and collection growth, including considerations about international interest and support, access (information seeking activities and physical distribution); student learning and skills development; worldwide availability of ETDs; and qualitative and economic aspects (e.g., usability and economic impact). In the next sections, we concentrate on describing NDLTD technological and research activities.

2.1 Building an ETD Digital Library

Virginia Tech has developed a suite of tools in order to support NDLTD members in their ETD activities as well as the several societies served by them. The ETD database (*ETD-db*) is the key-supporting tool for all those activities. It consists of a series of web pages and PERL scripts that interact with a MySQL database. These scripts provide a standard interface for web users and researchers, ETD authors, graduate program staff, and librarians to

[*] FOX, E. A. The networked digital library of theses and dissertations: changes in the university community. Journal of Computing in Higher Education, 2002.

[**] FOX, E. A. The ETDs sourcebook: theses and dissertations in electronic age. New York : Marcel Dekker, 2002.

enter and manage the files and metadata related to a collection electronic theses and dissertations. ETD-db includes interfaces and basic functionality for fulltext and keyword-based searching as well as browsing of local ETDs (by department and author). This software and its corresponding documentation, in ever-improving versions, is packaged periodically and made available without charge to other NDLTD institutions (from http://scholar.vt.edu/ETD-db/developer/). Recently the software also has been upgraded in connection with the Open Archives Initiative (http://www.openarchives.org) to support the Open Archives Metadata Harvesting Protocol (Van_de_Sompel and Lagoze 2001), which will allow automatic harvesting of ETD metadata from new members using this version of the software. Patches to incorporate this functionality into earlier versions of the software have been released. Other OAI related tools also are available for those using alternative software packages (http://www.dlib.vt.edu/projects/OAI/).

ETD submissions are controlled by a *workflow system* that guides the ETD author through all the steps of this process. It involves uploading ETD files into an ETD database, providing metadata in a particular format for ETDs (see 2.1.1 below) and filling in an online survey for university feedback. With preservation concerns in mind, NDLTD has chosen to support and allow submissions in a small set of standard formats. This includes XML, SGML, PDF, and Latex as well as standard multimedia formats. There area several Document Types Definitions (DTDs) to choose from regarding XML and SGML, and an XML schema for ETDs is currently in development. Instructions for conversion from unsupported formats also are available. A number of training materials and workshops have been developed and presented in order to teach to students how to use work with these tools, as well as popular word processing and formatting software.

For the Graduate School at Virginia Tech, an *Approval Wizard* was developed in connection with ETD-db to help staff receive, annotate, and approve ETDs. Included with this package is the ability to update metadata fields (e.g., for correcting spelling errors and applying controlled vocabularies) and to send email messages to students whose ETD chapters need to be updated. Built-in is the capability of the Graduate School to revise availability restrictions on ETDs as students request such changes. The wizard also will automatically generate email messages to authors, committee chairs, and third-party archiving services (e.g., Proquest, formerly UMI).

2.1.1 Electronic Thesis and Dissertation Metadata Standard (ETD-MS). ETD-MS (Atkins, Fox et al. 2001) is a metadata standard developed (through years of community discussion) and promoted by NDLTD to describe theses and dissertations. ETD-MS is based on the Dublin Core Element Set (Dublin-Core-Community 1999) but includes additional elements specific to theses and dissertations. It incorporates special fields to represent information like advisor and committee members, degree and level of education associated with the work, institution granting the degree and area of study of the intellectual content of the document. It also is designed to handle metadata in many languages, including metadata regarding a single work that is recorded in different languages. The ETD-MS standard provides detailed guidelines on mapping information about an ETD to metadata elements. ETD-MS already is supported as an output format for the Open Archives interface to the Virginia Tech ETD collection and a reversible mapping to/from MARC-21 also is available. NDLTD strongly encourages use of ETD-MS.

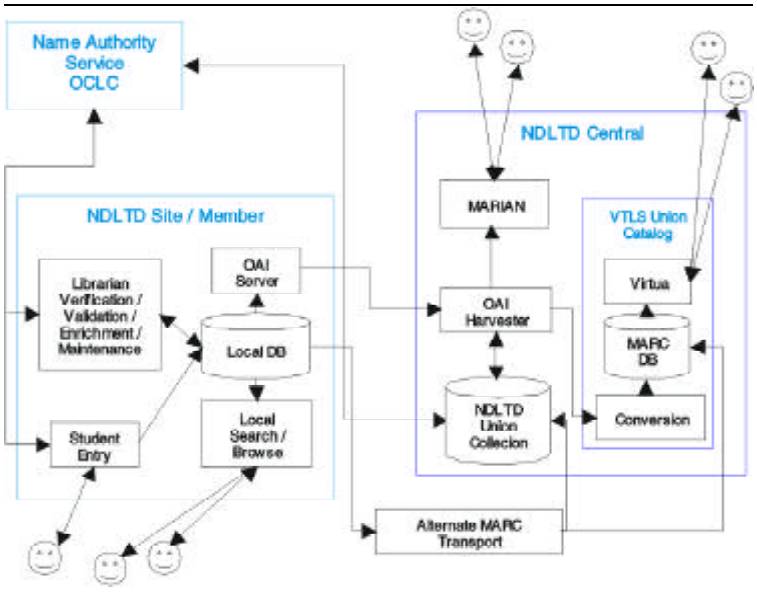
2.2 NDLTD as a Federation of DLs

NDLTD is a federation of digital libraries focused on efforts related to ETDs. One of the main objectives of the federation is to provide transparent and integrated services that span NDLTD members' ETD collections, while keeping with the general desire at each site to maintain their individual collections. Searching and browsing constitute some of these basic services. One of the first projects aimed towards these objectives was the federated search system (Powell and Fox 1998) (available at http://www.theses.org). The system multicasts queries to NDLTD member sites by consulting site configuration files to determine how to transliterate search arguments for the various federated systems. Results for each site are visited one by one. The results are not merged largely because of the complexity of merging search results without knowledge of the underlying presentation and ranking algorithms. The system also suffers from high network latency and uncertain availability of servers.

To overcome those problems, we have adopted a new solution based on a "union catalog" architecture, using the Open Archives Initiative's Metadata Harvesting Protocol to gather metadata in specific metadata formats and then make it accessible at mirrored portals. This solution has the advantage of a single database upon which many services can be built but requires a more active and collaborative role from the NDLTD members. Current portals to provide a web interface to the ETD Union Catalog include the Virtua system by VTLS Inc.,

working in a production mode (www.vtls.com/ndltd), and the MARIAN digital library system (Fox, France et al. 1993; Gonçalves, France et al. 2000; Gonçalves, France et al. 2001; Gonçalves, France et al. 2001) for the investigation of new research trends in ETD services.

FIGURE1
Architecture of OAI-based Union Collection



VTLS Inc. (www.vtls.com) is an established developer of software to manage library collections, both digital and non-digital. Virtua ILS, their flagship product, is an integrated library automation system specifically designed to cater to the differing needs of librarians in different contexts. Virtua is especially suited to the needs of NDLTD because it is inherently a distributed system and adheres to emerging standards for encoding of metadata. All metadata is stored in Unicode and this makes it much easier to deal with the non-ASCII character sets used by a growing number of NDLTD member sites in non-English-speaking countries. This extends Virtua's search capabilities to every language that can be represented in Unicode, thus providing users with multilingual search and retrieval services. The Virtua interface also has multilingual capabilities. There are currently versions of the user interface in over a dozen languages, with planned support for all languages used by NDLTD members. The coupling of multilingual information retrieval with multilingual interfaces has the desirable effect of providing a complete and consistent digital library for users who speak languages other than English.

Since the OAI protocol is simple and flexible, there is much leeway in designing distributed systems like the NDLTD Union Catalog. In exploiting this ability, the Union Catalog uses a two-tier approach to separate the merging of collections from the provision of high-level user services through our several portals. This is illustrated in Figure 1.

Each of the participating member archives exports data using the OAI protocol. This data is then harvested from each site into a central Merged Collection and republished as a single collection through an Open Archives interface. The Virtua and MARIAN systems in turn harvests data from this Merged Collection to provide higher-level user services. This separation between merging of collections and provision of services has the advantage that the merged collection can act as a local cache for use by production services like Virtua as well as research projects like MARIAN (Goncalves, France et al. 2000). Such a local cache reduces the network load on ETD archives and also simplifies the problem of data management for service providers like Virtua. In addition, this tiered architecture supports more natural integration and serves as a proof-of-concept for NDLTD members, which are inherently federations rather than single institutions, e.g., the Brazilian Digital Library.

3. OTHER NUDL EFFORTS

In addition to supporting graduate students, this project also supports other university related activities and materials as well as specialized services for different disciplines and communities. Two such involved communities are those of Computing and Physics.

3.1 Computing

The first efforts of NUDL to collect educational and curricular resources produced by teachers and professors around the globe were carried out in the field of Computing and are materialized as the Computer Science Teaching Center (www.cstc.org and (Fox 1998; Fox 1998; Fox 1998).

The CSTC is a digital library of reviewed resources for teaching computer science (Knox, Dale et al. 1999). By setting up a digital library to facilitate submission of curricular material (in similar fashion to NDLTD's supporting of thesis submissions), NUDL aims to help achieve the critical mass required to attract the attention of teachers. By encouraging small submissions, in

particular of knowledge modules (Fox and Kieffer 1995), and by instituting a system of peer review to ensure quality, we hope to turn CSTC into an example of how to develop an NSF-funded Science, Mathematics, Engineering and Technology Education (SMETE) DL (Arms 1999), i.e., NSDL (Zia 2001).

Originally, we began by soliciting three types of materials for CSTC. First, we seek laboratory resources that can be used in lab sessions in (primarily lower level) courses on computing, or for self-study. Second, we seek both visualizations of important concepts in computing, and tools to help prepare such visualizations. Third, we seek interactive multimedia resources that teach about important aspects of computing. Each type requires knowledgeable editorial control controlled by CSTC editors. Other editors have joined the effort, so a wide variety of types of materials are being collecting, including syllabi and lecture notes.

Contributors log onto the CSTC site, fill in online forms to describe their resource, and upload their works – for test, review, and eventual archiving. Thus, they provide the metadata called for by the Dublin Core Metadata Initiative (Dublin-Core-Community 1999) as well as the digital object or objects to be shared with other teachers or learners (IEEE_WG12 2000). Special interface screens have been devised to support people in the roles of contributor, user (e.g., a teacher discovering a useful resource), reviewer, and editor. These have been refined for several years, and will continue to be enhanced to suit the needs of the project and community.

CSTC also serves the ACM Journal of Educational Resources in Computing (JERIC (Cassel and Fox 2000)). The highest quality works submitted to CSTC can be reviewed for this archival journal that is part of the ACM Digital Library (www.acm.org/dl). Dedicated courseware developers thus can gain some of the recognition that is due, but rarely awarded.

Extending the work on CSTC and JERIC is a broader initiative, under the rubric of the NSDL, to develop the Computing and Information Technology Interactive Digital Educational Library (CITIDEL, see www.citidel.org). Led by Virginia Tech, a consortium also including the College of New Jersey, Hofstra, Penn State, and Villanova, aims to stimulate production and use of educational resources in the area. This collection project in NSDL will produce a portal, in both English and Spanish, aimed to support a broad range of users (faculty, teachers, college students, K-12 students, practitioners, etc.). The content will range from that suitable for JERIC, to that in other parts of CSTC, to resources in special collections (e.g., on software architecture, or computer organization), to likely resources harvested by niche search engines that have emerged in connection with the CiteSeer/ResearchIndex effort (Lawrence, Giles et al. 1999).

3.2 Physics

In the context of a NSF-DFG funded joint NUDL project, one of our partners at the University of Oldenburg, in Germany, has set up a set of services for the Physics community (and for some adjacent fields). Collectively called *PhysNet,* they offer online services that enable a physicist to keep in touch with the worldwide physics community and to receive all information he or she may need. One service is *PhysDoc* (Hilf 2000; Severiens, Hohlfeld et al. 2000). This consists of a Harvest™-based online information broker- and gatherer-network, which harvests information from the local web-servers of professional physics institutions worldwide (mostly in Europe and USA so far). *PhysDoc* focuses on scientific information posted by the individual scientist at his local server, such as documents, publications, reports, publication lists, and lists of links to documents. *PhysDis* (Hilf 2000) is an analogous service but specifically for university theses, with their dual requirements of examination work and publication. Other services include information about physics departments, jobs, calls for conferences, and educational resources.

Despite being extremely useful, the low quality and/or limited coverage of some of these services, mainly due to the occasional low relevance of the gathered data, and the technological limitations of the software, have prevented its widespread use. NUDL efforts, w.r.t. this community, are concentrated on: 1) improving the quality of services already supported by moving towards an OAI-based architecture, enrichment of metadata, and the use of ontologies to add-value to the collections; and 2) building of new services specially tailored to the specific needs and interests of that community. This project will serve as a foundation, setting procedures and standards, regarding similar activities for other communities.

4. RESEARCH ACTIVITIES

One of the main objectives of NUDL and NDLTD is to advance the state of the art in digital library technology. This aim has been carried by many research projects in the Virginia Tech Digital Library Research Lab. We describe some of those activities below.

4.1 Effective and Efficient Searching: The MARIAN System

MARIAN is a digital library system designed and built to store, search over, and retrieve large numbers of potentially complex digital objects. Originally planned for online library catalogs, it has been used successfully for collections of varying sizes and structures, and has been enhanced to support digital library and semantic web-like (Decker, Melnik et al. 2000; SemanticWebActivity 2001) applications.

The MARIAN data model combines three powerful concepts. First, structure and relationships in MARIAN collections are captured in the form of an *information network* of explicit nodes and links. Similar graph-based models have proven effective in representing semi-structured data and Web documents (Abiteboul, Buneman et al. 1999), and for translating among different DL systems (Melnik, Garcia-Molina et al. 2000).

FIGURE 2
ETD-MS semantic network representation as understood by the MARIAN system.

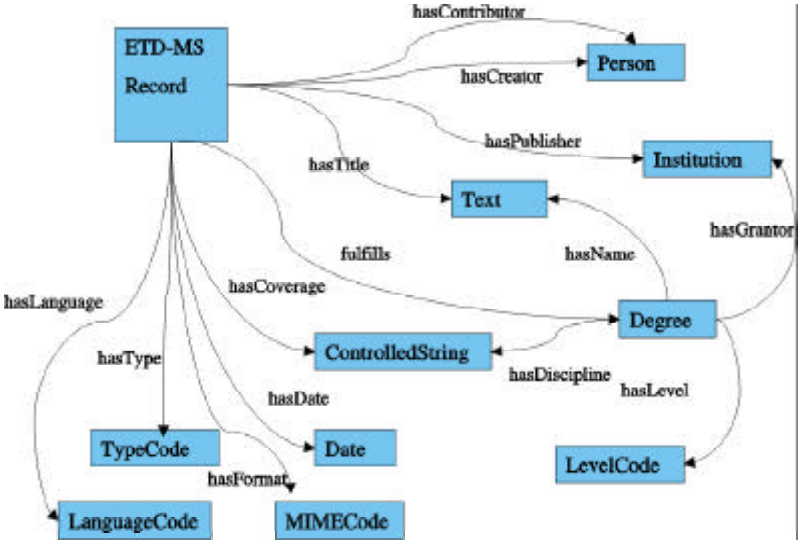


Figure 2 shows an example of MARIAN's semantic network representation for the ETD-MS metadata standard.

Second, MARIAN expands this model by insisting that the nodes and links of a collection graph be members of object-oriented *classes.* Classes are an organizing method similar to link labels in semi-structured graphs, but are strictly more powerful because they form a full lattice of subsets and can support inheritance of behavior. Furthermore, since nodes in the collection graph are instances of *information object* classes, they can support complex behaviors. In particular, they can support approximate matching of the sort pioneered in information retrieval (IR) systems.

Third, nodes or links can be *weighted* to represent how well they suit some description or fulfill some role. MARIAN is specialized for a universe where searching is distributed over a large graph of information objects. The output of a search operation is a *weighted set* of objects whose relationship to some external proposition is encoded in their (decreasing) weight within the set. Weights are used in IR, probabilistic reasoning systems, and fuzzy set theory. Our model grounds them firmly in a framework of weighted set operations (France 1995) and extends them throughout the entire MARIAN system.

The use of object-oriented data and process abstractions in MARIAN helps to achieve physical and logical independence – common and useful concepts in the database field, often neglected in IR. Most current IR systems emphasize the physical level of term indexes and weight metrics, making it difficult to integrate systems at a conceptual level (Fuhr 1999). The flexibility of the MARIAN data model allows it to be used for object-oriented or semi-structured databases, knowledge representation, or IR. Its power comes from the smooth combination of a number of successful concepts from such fields as programming languages and artificial intelligence (Fuhr and Rolleke 1997).

MARIAN has been used in a number of different digital libraries projects including the full metadata collection of the Virginia Tech Library, a digital library of mixed metadata and fulltext documents, a directory of health-care organizations from the National Library of Medicine, and the NDLTD Union Catalog. The Union collection in particular highlights how MARIAN can smoothly combine divergent document formats and semantics into a single unified collection view. The flexibility of the MARIAN data model also has provided an integrated framework for addressing such questions as data quality, flexible and efficient search, and scalability.

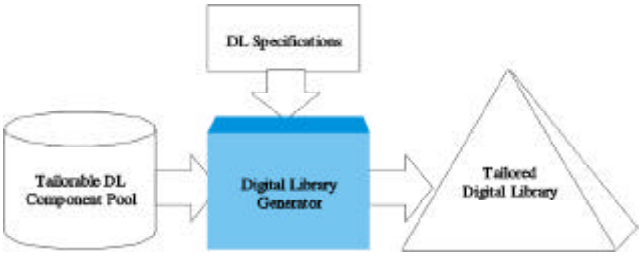
Various new research trends are being explored in MARIAN. To address variations in quality of service in the current network infrastructure and initial impedance of OAI acceptance, we are investigating a new hybrid architecture that integrates local searches on the union collection with federated search of selected sites. It can produce integrated results while improving freshness of information beyond that found with the union architecture. Problems to be solved include how to manage both approaches and how to combine or *fuse* results while maintaining efficiency and effectiveness. We are investigating approaches for solving those problems, including extension of our network model. These extensions involve, with the help of a group of the Federal University of Minas Gerais (UFMG), incorporation of *belief network models* (Ribeiro-Neto and Muntz 1996). These would allow combining results from different sources of information and improving the quality of delivered ranked results through the incorporation of additional evidential information (e.g., past queries and citations).

We also are investigating several ways to customize and cluster the union collection to better attend to user needs, for example, by exploring techniques to "slice-and-dice" the union collection content in new ways in order to provide better searching and browsing services. Finally, we are investigating how to use our VT-PetaPlex-1 system (a parallel machine with 2.5 terabytes of storage, 100 Pentium processors with 64M RAM, and high speed connectivity) (Akscyn 1998) as a storage system for MARIAN along with parallel information retrieval techniques to address issues of scalability and performance. Thus we can address the needs of NUDL, NDLTD, and of other heterogeneous federated collections.

4.2 Understanding, Generating and Tailoring DLs: 5S and 5SL

DL research at the DLRL is not restricted to directly solving practical DL problems. We also have explored theoretical issues, motivated in part by such problems. The experience gained along the years with many DL projects, including NDLTD (www.ndltd.org), NUDL, ENVISION (Heath, Hix et al. 1995), among many others, has allowed us to develop the 5S theory for digital libraries, of Streams, Structures, Spaces, Scenarios, and Societies (Gonçalves, Fox et al. 2001). 5S helps to understand the many facets of DLs and fulfills a much urgent and necessary gap induced by the absence of any formal theory for the field (Licklider 1965; Ranganathan 1965).

FIGURE 3
Digital Library generation with 5SL



In more detail: Streams are sequences of abstract items used to describe DL static and dynamic content. Structures can be defined as labeled directed graphs, which impose organization in the digital library. Spaces are sets of abstract items and operations on those sets that obey certain rules. Scenarios consist of sequences of events or actions that modify states of a computation in order to accomplish a functional requirement, therefore describing DL services. Societies comprehend entities and the relationships between and among them inside the context of the DL. Thus, the 5S theory provides a comprehensive and unified view of the many aspects involved in a digital library. We have used 5S to formally define digital library concepts, to analyze a DL taxonomy, and to semi-formally describe NDLTD and the Open Archives Metadata Harvesting Protocol. As such, 5S also serves as a tool for DL requirement analysis (Fox 1999).

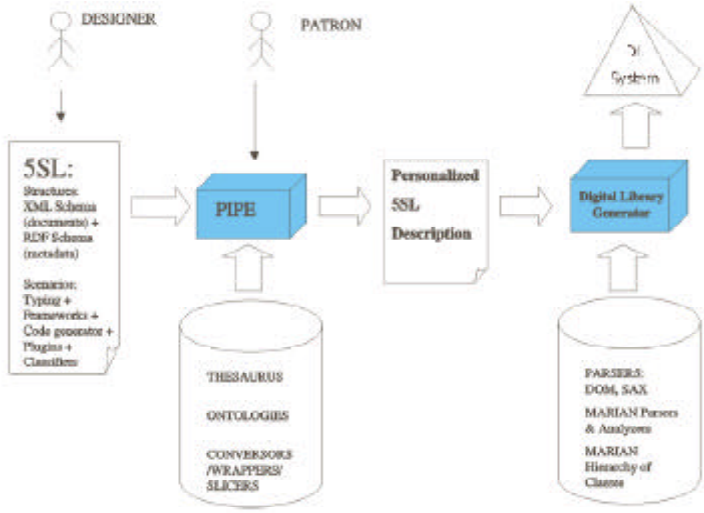
Based on the last observation, on the empirical evidence of the difficulty of building DLs, and on the lack of specific DL patterns, models, methodologies, formalisms, and languages to help builders to create and tailor DL system to different DL community needs, we have been designing 5SL, a declarative language for conceptual DL design. 5SL is a high-level, domain-specific language, which allows specifying a number of DL features that often are considered in isolation. The 5SL language is based on XML syntax and, with the help of a pool of components, can be fed to software generators for automatically producing the implementation of a DL (Figure 3). We are currently finishing a DL generator for the MARIAN system.

4.3 Other Advanced Services: Personalization

Digital libraries can be enormous information warehouses, with huge amounts of data, encompassing many kinds of multimedia formats. Societies of users/patrons are frequently inundated with massive quantities of information and are rarely provided with effective tools

that allow them to customize DL services and content for their own interests. Personalization (Resnick and Varian 1997; Rieken 2000), a possible approach to the problem, involves techniques and mechanisms to reduce this information overload and tailor DL systems for particular user communities with specific interests. However, due to the inherent complexity of DLs, incorporating personalization as a basic DL capability is an extremely hard task. Personalization can be carried out in three basic ways (Perugini, Gonçalves et al. 2002)*: 1) recommending services; 2) restructuring of information architectures to attend personal preferences, needs, or requirements; and 3) identification and exploitation of social networks which contribute to bring people together based on common characteristics.

FIGURA 4
The Integrated DL Personalization Framework



In the context of NDLTD, we have been playing with the two first kinds of personalization themes. Recommendations of ETDs were first tried with a experimental software extending the SIFT package (Yan and Garcia-Molina 1999) from Stanford University and, to provide filtering and routing services based on stored user profiles, for those who wish to be notified whenever a relevant ETD arrives. More recently, a class project has explored a similar architecture using OAI and MARIAN. This work will be concluded when MARIAN conversion to Java is completed, by early 2002. Finally, a third service based on a componentized OAI-based DL (Suleman and Fox 2001) is being currently developed (http://purl.org/net/etdunion).

Solutions for the second kind of personalization involve the construction of a framework that enables the automatic building of personalized digital libraries targeted for different communities of users. Our framework relies on two major building blocks (Gonçalves Zafer et al. 2001): 1) 5SL and 2) PIPE (Ramakrishnan 2000), a methodology for personalization that supports the building of automatic personalized views of the DL without enumerating explicit restructurings or interaction sequences. PIPE provides a systematic conceptual methodology to study the design, implementation, and evaluation of personalization systems. PIPE models personalization by the programmatic notion of *partial evaluation* (Jones 1996).

Partial evaluation is a technique used to automatically tailor programs, given incomplete information about their input. The input to a partial evaluator is a program and some static information about its arguments. The output is a specialized version of the program (typically in the same language) that uses the static information to "precompile" as many operations as possible.

Our Integrated Personalization Architecture is shown in Figure 4. It is an extension of the original architecture of Figure 3. 5SL encodes the original description of the DL as provided by the DL designer. In order to personalize it with PIPE, 5SL descriptions for structures and scenarios are modeled as programs, which abstracts the underlying schema and flow of information in the DL. Here we take advantage of the more stable, structured and managed nature of the DL and of our language. 5SL XML syntax makes easier to reuse external components able to transform those representations in a programmatic one, amenable to be used by PIPE. The 5SL programs are partially evaluated with respect to user input and the output is used to recreate a personalized 5SL description from the specialized program. This personalized description is then utilized to create personalized DL matching user interests.

* Perugini, S., M. A. Gonçalves, et al. (2002). "A Blueprint for Studying Recommendation and Personalization."

We have been experimenting with this architecture to personalize browsing of DL content through classification schemes, described with 5SL (represented with a special RDF syntax). Some input from the user (e.g., statement of interest in Digital Libraries) allows the personalization mechanism to isolate the user interests in just a small part of the whole classification system and corresponding DL content. An experiment was performed with the Library of Congress subject heading Hierarchy, the Virginia Tech Library catalog, and the MARIAN system supporting the link between both, with moderate success. Problems of scalability related to the size and deepness of the LC classification system size still need to be solved. Future work will concentrate on that and on personalizing navigation through discipline–based classification schemes like the ACM classification system for Computing and PACS for Physics.

5. CONCLUSIONS

Digital libraries are complex information systems. In this paper we have shown how to deal with that complexity, by way of example, with regard to various aspects of the Networked University Digital Library (NUDL). NUDL has evolved out of the Networked Digital Library of Theses and Dissertations, which continues to expand rapidly as a worldwide federation that aims to enhance graduate education and research. NUDL now encompasses not only electronic theses and dissertations, but also reports, eprints, and courseware. In addition to providing general coverage, it supports disciplinary focus, such as on computing or physics.

With regard to computing, our focus has been on courseware, to support teaching and learning. The Computer Science Teaching Center (CSTC), ACM Journal of Educational Resources in Computing (JERIC), and Computing and Information Technology Interactive Digital Educational Library (CITIDEL) are all efforts in this direction for the field of computing. Similarly, PhysNet, which includes services like PhysDoc and PhysDis, is focused on supporting physics.

Services in NUDL build upon traditional work with searching and browsing. Of particular interest are scalability and personalized services. Further, in order to collect distributed content, NUDL employs harvesting, such as with the OAI-MHP. This leads to union collections or archives, and alternative services supported thereby, such as VTLS's Virtua, and the MARIAN research system.

All of these efforts are based on a broad R&D effort, ranging from work on the 5S theoretical framework to deployment of the 100-node PetaPlex superstorage system. 5S shows particular promise through its language, 5SL, which may allow specification and description of DLs, so specialized DLs can be generated rapidly for new applications. A different type of tailoring comes from the PIPE approach to personalization. At the heart of much of this work is the MARIAN system, a research DL, through which many studies will proceed. Thus, in conclusion, while DLs are complex, they can be managed, through integrated projects (e.g., NUDL) and through advanced technology (e.g., 5SL, MARIAN, OAI, PIPE, …).

REFERENCES

ABITEBOUL, S. *et al. Data on the web*: *from relations to semistructured data and XML*. San Francisco, CA : Morgan Kaufmann, 1999.

AKSCYN, R. M. The PetaPlex Project, status briefing for National Security Agency. [S. l. : s. n.], 1998.

ARMS, W. (1999). THE NSF SCIENCE, MATHEMATICS, ENGINEERING, AND TECHNOLOGY EDUCATION LIBRARY WORKSHOP, 1999. [S. l.] : National Science Foundation, Division of Undergraduate Education, 1988.

ARMS, W. Y. *Digital libraries*. Cambridge, MA : MIT, 2000.

ATKINS, A., E. et al. ETD-ms: an interoperability metadata standard for electronic theses and dissertations. Blacksburg, VA: NDLTD, 2001.

BORGMAN, C.; FOX, E. A. THE FIRST JOINT CONFERENCE ON DIGITAL LIBRARIES, 2001, Roanoke, VA. ***Proceedings***... New York : ACM, 2001.

CASSEL, L.; FOX, E. A. ACM. *Journal of Education Resources in Computing*, 2000.

DECKER, S. The Semantic web: the roles of XML and RDF. *IEEE Internet Computing*, v. 4, n. 5, 2000.

DUBLIN-CORE-COMMUNITY. Dublin Core metadata initiative the Dublin Core: a simple content description model for electronic resources. Dublin, Ohio : OCLC, 1999.

FOX, E.; FRANCE, R. *et al.* Development of a modern OPAC. *In*: REVTOLC to MARIAN. ANNUAL INT'L ACM SIGIR CONFERENCE ON R&D IN INFORMATION RETRIEVAL, 16th , 1993, Pittsburgh. ***Proceedings***... [S. l.] : ACM, 1993, p. 248-259.

FOX, E. A. Digital libraries (hot topics). *IEEE Computer*, v. 26, n. 11, 79-81, 1993.

FOX, E. A. Sourcebook on digital libraries: report. Blacksburg, VA : The National Science Foundation. Dept. of Computer Science, 1993.

FOX, E. A. Digital libraries: the networked digital library of theses and dissertations and the Computer Science Teaching Center. *In*: COMPUTER SCIENCE WORKSHOP, 1998, Puebla, Mexico. [S. l.] : CONACyT, NSF, 1988.

FOX, E. A. Digital library courseware. Blacksburg, VA : Virginia Tech Department of Computer Science, 1998.

FOX, E. A. *Helping learners through digital libraries*: the networked digital library of theses and dissertations and the Computer Science Teaching Center. Amherst, MA : Department of Computer Science, Univ. of Mass., 1998.

FOX, E. A. Improving education through the networked digital library of theses and dissertations and the Computer Science Teaching Center. *In*: RUSSIAN-AMERICAN DIGITAL LIBRARIES WORKSHOP, 1988, Moscow. [S. l. : s. n.], 1998.

FOX, E. A. THE 5S FRAMEWORK FOR DIGITAL LIBRARIES AND TWO CASE STUDIES: NDLTD and CSTC, 1999, Taipei, Taiwan. ***Proceedings***... [s. l. : s. n.],1999.

FOX, E. A. Guest editors' introduction to digital libraries. *Communications of the ACM*, 6v. 38, n. 4, p. 22-28, 1995.

FOX, E. A.; KIEFFER, L. Multimedia curricula, courses and knowledge modules. *ACM_Computing* Surveys, v. 27, n. 4 : 549-551, 1995.

FOX, E. A.; LUNIN, L. Introduction and overview to perspectives on digital libraries: guest editor's introduction to special issue. *Jounal of the American Society of Information* Science, v. 44, n. 8, p. 441-443, 1993.

FOX, E. A.; MARCHIONINI, G. FIRST ACM INTERNATIONAL CONFERENCE ON DIGITAL LIBRARIES,1996, Bethesda, MD. ***Proceedings***... New York : ACM, 1996.

FOX, E. A.; MARCHIONINI, G. Toward a worldwide digital library: guest editors' introduction to special section on digital libraries: global scope, unlimited access. *Communications of the ACM*, v. 41, n. 4, p. 28-32, 1998.

FOX, E. A.; MARCHIONINI, G. Digital libraries: guest editor's introduction. *Communications of the ACM*, v. 44, n. 5, 2001.

FOX, E. A.; ROWE, N. ACM CONFERENCE ON DIGITAL LIBRARIES, 4th , 1999, BERKELEY, CA. ***Proceedings***... Berkeley : ACM, 1999.

FOX, E. A.; SORNIL , O. *Digital libraries*: modern information retrieval. Harlow : England, ACM, 1999, Addison-Wesley-Longman, 1999, p. 415-432.

FOX, E. A.; SORNIL, O. Digital libraries. *In*: ENCYCLOPEDIA OF COMPUTER SCIENCE, 4th ed. London : Nature Publishing Group, p. 576-581, 2000.

FRANCE, R. K. *Weights and measures*: an axiomatic approach to similarity computations. Blacksburg : Virginia Tech, 1995.

FUHR, N. Towards data abstraction in networked information Systems. *Information Processing and Management*, v. 35, n. 2, p. 101-119, 1999.

FUHR, N.; ROLLEKE, T. (1997).A probabilistic relational algebra for the integration of information retrieval and database systems. *ACM Transactions on Information* Systems, v. 15, n. 1, p. 32-66.

GILES, C. L. (1998). CiteSeer: an automatic citation indexing system.*In*: ACM CONFERENCE ON DIGITAL LIBRARIES, 3. rd., 1998, Pittsburgh. ***Proceedings***... [S. l.] : ACM, 1998. p. 89-98.

GONÇALVES, M. A. *Streams, structures, spaces, scenarions, societies (5S):* a formal model for digital libraries. Blacksburg, VA : Virginia Tech Dept. of Computer Science, 2001.

GONÇALVES, M. A. MARIAN: flexible interoperability for federated digital libraries. *In*: EUROPEAN CONFERENCE ON RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, 5th , 2001, Darmstadt, Germany. ***Proceedings***... [S. l. : s. n.], 2001.

GONCALVES, M. A. MARIAN: searching and querying across heterogeneous federated digital libraries. *In*: DELOS NETWORK OF EXCELLENCE WORKSHOP ON INFORMATION SEEKING, SEARCHING AND QUERYING IN DIGITAL LIBRARIES, 1rs., 2000. Zurich, Switzerland, DELOS. ***Proceedings***... [S. l. : s. n.], 2000.

GONÇALVES, M. A. Flexible interoperability in a federated digital library of theses and dissertations. *In*: WORLD CONFERENCE ON OPEN LEARNING AND DISTANCE EDUCATION, THE FUTURE OF LEARNING - LEARNING FOR THE FUTURE: SHAPING THE TRANSITION, 20th , 2001. Düsseldorf, Germany. ***Proceedings***... [S. l.] : ICDE, 2001.

GONÇALVES, M. A. Modeling and building personalized digital libraries with PIPE and 5SL. *In*: JOINT DELOS-NSF WORKSHOP ON PERSONALIZATION AND

RECOMMENDER SYSTEMS IN DIGITAL LIBRARIES, 2001, Dublin, Ireland. [S. l], 2001.

HEATH, L. Envision: a user-centered database from the computer science literature. *Communications of the ACM*, v. 38, n. 4, p. 52-53, 1995.

HILF, E. R. PhysDis: physics theses in Europe. [S. l. : s. n.], 2000.

HILF, E. R. PhysDoc: physics documents worldwide. [S. l. : s. n.], 2000.

IEEE_WG12. Draft standard for learning object metadata, learning technology standardization committee of the IEEE. [S. l.] :, 2000.

Jones, N. D. An introduction to partial evaluation. *ACM Computing* Surveys, v. 28, n. 3, p. 480-506, 1996.

KNOX, D. The peer review process of teaching materials. *SIGCSE Bulletin Inroads* , v.

31, n. 4, p. 87-100, 1999.

LAWRENCE, S. Digital libraries and autonomous citation indexing. *IEEE Compute*, v. 32, n. 6, p. 67-71, 1999.

LESK, M. *Practical digital libraries*: books, bytes and bucks. San Francisco : Morgan Kaufmann, 1997.

LICKLIDER, J. C. R. *Libraries of the future*. Cambridge, MA : MIT, 1965.

MARCHIONINI, G.; FOX E. A Progress toward digital libraries: augmentation through integration. *Information Processing and* Management, v. 35, n. 3, p. 219-225, 1999.

MELNIK, S., H. (2000). A mediation infrastructure for digital library services. *In*: ACM CONFERENCE ON DIGITAL LIBRARIES, 5th , 2000, San Antonio, TX. ***Proceedings***... New York, 2000, p.123-132.

Moxley, J. (2001). ETD Guide, USF.

PERUGINI, S., M. A. Gonçalves, et al. (2002). "A Blueprint for Studying Recommendation and Personalization." (in preparation).

POWELL, J; E. FOX.Multilingual federated searching across heterogeneous collections.

D-Lib Magazine, v. 4, n.8, 1998.

RAMAKRISHNAN, N. PIPE: web personalization by partial evaluation. *IEEE Internet Computing, v.* 4, n. 6, p. 21-31, 2000.

RANGANATHAN, S. R. A *Descriptive account of colon classication*. Bangalore, Sarada : Ranganathan Endowment for Library Science, 1965.

RAO, R., J. O. Rich interaction in the digital library. *Communications of the ACM*, v. 38, n. 4, p. 29-39,1995.

RESNICK, P.; VARIAN, H. R. Recommender systems: introduction to the special section. *Communications of the ACM*, v. 40, n. 3, p. 56-58, 1997.

RIBEIRO-NETO, B.; MUNTZ , R. A belief network model for IR. *In*: ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 19th, 1996. ***Proceedings***... [S. l.], 1996. p. 253-260.

RIEKEN, D. Personalized views of personalization. *Communications of the ACM*, v. 43, n. 8, p. 27-28, 2000.

SEMANTIC web activity. [S. l. : s. n.], 2001.

SEVERIENS, T. PhysDoc: a distributed network of physics institutions documents: collecting, indexing, and searching high quality documents by using harvest *D-Lib Magazine*, v. 6, n. 12, 2000.

SULEMAN, H; FOX, E. A. A framework for building open digital libraries. *D-Lib Magazine*, v. 7, n. 10, 2000.

SOMPEL, H. van_den; LAGOZE, C.The open archives initiative protocol for metadata harvesting: protocol version 1.0, document version 2001. Ithaca, NY : Cornell University, 2001.

YAN, T. W.; GARCIA-MOLINA, H. The SIFT information dissemination system. *ACM Transactions on Database* , v. 24, n. 4, p. 529-565, 1999.

ZIA, L. L.The national science, mathematics, engineering, and technology education digital library program. *CACM*, v. 44, n. 5, 2001.

ACKNOWLEDGEMENTS

We acknowledge support by many groups, including ACM, CONACyT, DFG, FIPSE (P116B61190), NLM, OCLC, NSF (DUE-9752190, DUE-0121679, IIS-9986089, IIS-0080748, IIS-0086227), UNESCO, and VTLS. The first author also is supported by CAPES, 1702/98-0. A large number of people also have assisted with the efforts described above. We hereby give thanks to all of those, especially: Anthony Atkins, Lillian Cassel, Vinod Chachra, John Eaton, Robert France, Eberhard Hilf, John Impagliazzo, Deborah Knox, JAN Lee, Paul Mather, Gail McMillan, Ryan Richardson, Hussein Suleman, Jun Wang, Ye Zhou, and Royce Zia.

Artigo recebido em 23/11/2001.