
ARTICLES

D-Lib Magazine
September 2001

Volume 7 Number 9

ISSN 1082-9873

Networked Digital Library of Theses and Dissertations**Bridging the Gaps for Global Access - Part 2: Services and Research**

[Hussein Suleman](#), [Anthony Atkins](#), [Marcos A. Gonçalves](#), [Robert K. France](#), [Edward A. Fox](#)
([hussein](#), [anthony.atkins](#), [mgoncalv](#), [france](#), [fox](#)) @vt.edu
Virginia Tech

[Vinod Chachra](#), [Murray Crowder](#)
([chachra](#), [crowderm](#)) @vtls.com
VTLIS Inc.

[Jeff Young](#)
[jyoung@oclc.org](#)
OCLC

Abstract

The Networked Digital Library of Theses and Dissertations (NDLTD) is a collaborative effort of universities around the world to promote creating, archiving, distributing and accessing Electronic Theses and Dissertations (ETDs). Since its inception in 1996, over a hundred universities have joined the initiative, underscoring the importance institutions place on training their graduates in the emerging forms of digital publishing and information access. The outreach and training mission of NDLTD is an ongoing project so in this article we report on the current status of membership and support activities. Recent research has focused on creating a union database that will provide a means to search and retrieve ETDs from the combined collections of NDLTD member institutions. The Virtua system developed by VTLIS will serve as the heart of this union database. In order to bridge the gap between the existing distributed institutional archives and a unified collection of ETDs, we have developed a metadata standard especially suited to ETDs - this is then used by partner sites to export their freely-available metadata using the Metadata Harvesting Protocol of the Open Archives Initiative. We also link name authority information into the metadata records to support unique identification of authors and others associated with the works. Additional research efforts include advanced search mechanisms, semantic interoperability, the design and development of multi- and cross-lingual search systems, and software modules that support the development of higher-level services to aid researchers in seeking relevant ETDs.

The Union Catalog Project**Motivation**

Simply by virtue of being called the "Networked" Digital Library of Theses and Dissertations, NDLTD immediately conjures up images of an interconnected system of digital archives. With this aim in mind, many recent efforts have attempted to supplement increasing membership with more advanced services, such as searching and browsing that span multiple collections of ETDs. At Virginia Tech, the first of such projects was the Federated Search system [[Powell and Fox, 1998](#)]. This system distributes a query to multiple sites and then gathers the result pages into a cache for browsing. The results are not merged largely due to the complexity of merging search results without knowledge of the underlying ranking algorithms. The system also suffers from high network latency and uncertain availability of servers. An alternative solution is typified by the OCLC WorldCat project [[OCLC, 2001](#)] that collects bibliographic data from libraries all over the world into the OCLC Online Union Catalog. Libraries may then acquire extracts from this catalog corresponding to theses and dissertations. This solution has the advantage of a single database upon which many services may be based but requires subscription to OCLC's services and obscures the differences between ETDs and their traditional paper counterparts. In response to this need for a focused and accessible catalog with a low barrier to participation, NDLTD has adopted a solution that uses the Open Archives Initiative's Metadata Harvesting Protocol [[Lagoze and Van de Sompel, 2001](#)] to gather metadata in the ETDMS format and then to make it accessible at a central portal. This central portal is maintained by VTLS, using their Virtua system [[VTLS, 2001a](#)] to provide a web interface to the ETD Union Catalog.

The Virtua NDLTD Portal

VTLS Inc. [[VTLS, 2001b](#)] 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.

An instance of Virtua has been developed by VTLS to serve as the central portal for the NDLTD Union Catalog (see Figure 1). This portal provides users with a simple and intuitive interface to search and browse through the merged collection of theses and dissertations. After potentially relevant items are discovered, a user can follow the links provided to go directly to the items in their source archives.

Figure 1. Virtua-based NDLTD Portal

a. Main entry page, in English, Spanish or Korean

The main entry page features a blue header with the text "NETWORKED DIGITAL LIBRARY OF THESES AND DISSERTATIONS" and a search type dropdown menu set to "English". Below the header, there are navigation links: "New Session | Save Session | Cart | History | Help". The main content area is titled "To begin a search..." and includes instructions: "Enter your search terms. Select a search type. Choose a target database. Click the Search button." A search form contains a text input field, a dropdown menu set to "Author", and a button labeled "NDLTD Union Catalog" with a "Search" button next to it. A tooltip window is overlaid on the page, providing instructions in English, Spanish, and Korean. The English text reads: "Welcome to NDLTD Union Catalog! We are pleased to announce that the new version of the NDLTD Union Catalog is now available. This version includes a new search interface, a new search type, and a new search button. The new search interface is designed to be more user-friendly and easier to use. The new search type is 'Browse Search', which allows you to search for a specific author, title, or subject. The new search button is 'Search', which is located in the toolbar. To begin your search, click on 'Browse Search' in the toolbar, enter your search terms, and click on 'Search'." The Spanish text reads: "Bienvenido al catálogo por autor de tesis, tema, título y número indicativo bibliotecario. Someta clic en 'Browse Search' en el 'toolbar' gris para iniciar su búsqueda." The Korean text reads: "NDLTD Union Catalog에 오신 것을 환영합니다. 이 유니온캐탈로그는 전세계의 여러 학회기관이 제공한 것으로서, 대학원 과정 교육자료의 풍부한 자료원으로 활용될 수 있습니다. 멀티미디어와 하이미디어의 기술 및 전세계의 진보된 디지털도서관 기술을 이용함으로써, 우리는 학생들의 학문적 연구의 역량을 높여주고, 그들이 훌륭한 논문을 쓸 수 있기를 희망합니다. 학회논문의 저자, 주제, 제목, 접근번호 목록으로 검색할 경우에는, 검색 도구상자에서 'Browse Search' 단추를 클릭하여 검색을 시작하십시오. 학회논문 목록에서 정확하게 일치하는 단어나 어구로 검색할 경우에는, 검색 버튼을 클릭해서 저자, 제목, 주제 등과 같은 키워드 항목을 지정하고, 회색 도구상자에서 'Keyword Search' 단추를 클릭하여 검색을 시작하십시오. 전문적 검색의 경우에는, 논리연산 명령어를 사용할 수 있습니다. 'Expert Search' 단추를 클릭하여 전문적 검색을 시작하고, 데이터 입력상자 밑에 있는 검색 도움말의 지시를 따르십시오." Below the tooltip, there is a section for "Expert search" instructions.

b. Search results, showing entries in multiple languages

The search results page shows a table of results for the search "Author:(ΕΒνικό)". The table has two columns: "Hit Count" and "Scan Term". The results are as follows:

Hit Count	Scan Term
1	Zhuang, Hong.
1	Εθνικό Μετσόβιο Πολυτεχνείο. Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών.
1	ΜΑΣΤΡΟΣ, ΓΙΑΝΝΗΣ Γ.
1	ΟΙΚΟΝΟΜΟΥ, ΚΩΝΣΤΑΝΤΙΝΟΣ Γ.
2	ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ.
1	ΠΑΠΑΓΙΑΝΝΗ, ΔΙΚΑΤΕΡΙΝΗ Α.
1	Πανεπιστήμιο Θράκης. Σχολή Πολυτεχνική. Τμήμα Πολιτικών Μηχανικών.
1	Πανεπιστήμιο Θράκης. Τμήμα Ανεξάρτητα. Τμήμα Ιατρικής.
1	가민현
1	강미영

Figure 1. Virtua-based NDLTD Portal

Part of Virtua's appeal is the high degree of customization of its Chameleon Web Gateway, appropriately named for its ability to superimpose multiple "skins" on a user portal for different communities. Besides being thus tailored to the typical information-seeking behavior of researchers, the interface also has multilingual capabilities. There are currently versions of the user interface in Korean and Spanish, 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.

Virtua supports multiple modes of data entry, converting all input data into the standard USMARC format that is familiar to many librarians and archivists. Some of the data currently loaded into Virtua was acquired in batches from source archives in different countries (including Greece, Korea, Portugal, and the USA) to ensure demonstrable multilingual support at the initial launch of the portal. In parallel with the development of this portal, mechanisms have been put into place to support fully automated importing of metadata using protocols such as that developed by the Open Archives Initiative.

Open Archives as an Interoperability Framework

In order to gather metadata into the central Union Catalog, NDLTD has adopted the Open

Archives Initiative's (OAI) interoperability framework. This includes using the Protocol for Metadata Harvesting as well as defining a metadata format targeted at the community of ETD archives.

The OAI protocol is a request-response protocol layered over HTTP that allows one computer to collect metadata incrementally over time from another computer - this is commonly known as harvesting. The requests have a minimal number of parameters and correspond to HTTP GET or POST operations.

The responses are XML documents whose structures are defined precisely using the XML Schema Description [Fallside, 2001] language, with protocol tags designed to support the operation of the protocol and container tags that encapsulate individual record or archive-level metadata. Multiple metadata formats can be supported for each item in an archive's collection, with Dublin Core (DC) [DCMI, 1999] being mandatory. For NDLTD, ETDMS is recommended as a community standard.

The following HTTP request and XML response (see Figure 2) are a typical use of the protocol to retrieve the "oai_etdms" metadata corresponding to a single record identified by "oai:VTETD:etd-520112859651791".



Figure 2. HTTP request and XML response for GetRecord

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 the Virtua interface. This is illustrated in Figure 3.

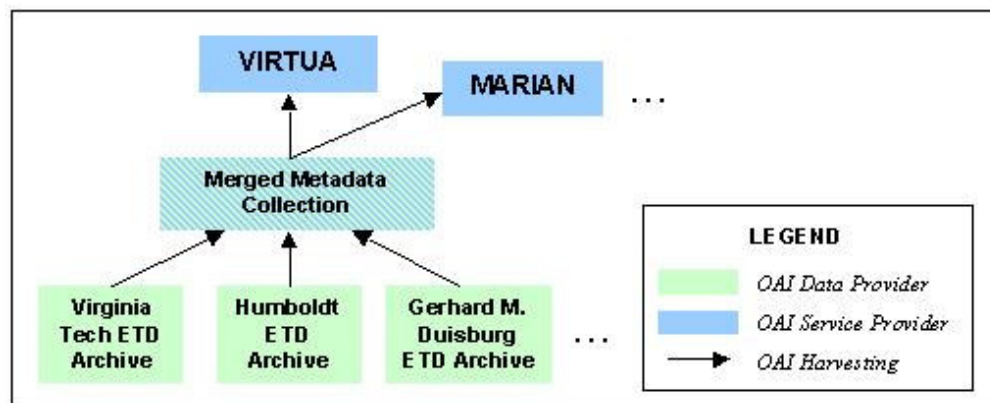


Figure 3. Architecture of OAI-based Union Collection

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 system 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 [Gonçalves, 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 who are inherently federations rather than single institutions, e.g., OhioLINK.

Current NDLTD Research Efforts

NDLTD has a strong commitment to advance the state of the art in electronic publishing and digital library technologies and services, in connection with its core activities supporting graduate education and sharing of knowledge. Much of the current research into providing global services for NDLTD is taking place in the context of the MARIAN digital library system [Fox, et al., 1993; Gonçalves, et al., 2000; Gonçalves, et al., 2001], developed at the Virginia Tech Digital Library Research Laboratory, and its interoperability with a number of other digital library systems. MARIAN supports flexible and extensible search over networks of digital information objects, which may include documents, metadata records, or digital surrogates for people and organizations. Digital information objects and connecting links are organized into object-oriented classes, each supporting indexing, retrieval, and presentation methods. New classes of information objects can be added to a MARIAN digital library by producing new or modified code implementing these functions for the new class manager.

Semantic Interoperability

Adoption of the OAI framework as a primary harvesting mechanism for ETD metadata can help overcome interoperability problems at the system, syntactic, and structural levels [Ouksel and Sheth, 1999]. However, several ETD members use underlying technologies like Z39.50 [Lynch, 1997], which are difficult to adapt for conformance to the OAI standards. Moreover, the OAI framework itself opens up the possibility of a single community utilizing several heterogeneous metadata standards. While the adoption of ETDMS as an official NDLTD standard is encouraged with the hope that someday all NDLTD members will use it, it is realistic to assume that it will take considerable time and

effort to achieve widespread adoption. Therefore, in the interim, semantic heterogeneity is a problem that has to be dealt with.

Semantic heterogeneity is solved in NDLTD by exploiting two MARIAN mechanisms: 1) semantically "tuned" but functionally equivalent searchers; and 2) a collection view ontology. Nodes in the MARIAN information network can be simple atomic or scalar objects, as in the semi-structured model [Abiteboul, 2000], but they also can be complex information objects. Information objects support methods proper to their classes, and all information objects in MARIAN support the method of approximate match to a query. For instance, MARIAN treats title text as a special sort of natural language sequence, with various rules for capitalization, punctuation, and sentence formation, but treats names of persons as sequences of atomic strings. Matching methods vary from class to class but all have the same functional profile: given an object description of the appropriate type, they calculate how closely they match the description and return that value as a weight. Class managers draw on these methods to provide class-level search functions that, given an object description, return a weighted set of objects in the class that matches the description. MARIAN already has a library of matching functions and searchers for a number of common information object classes, a sample of which are shown in Figure 4.

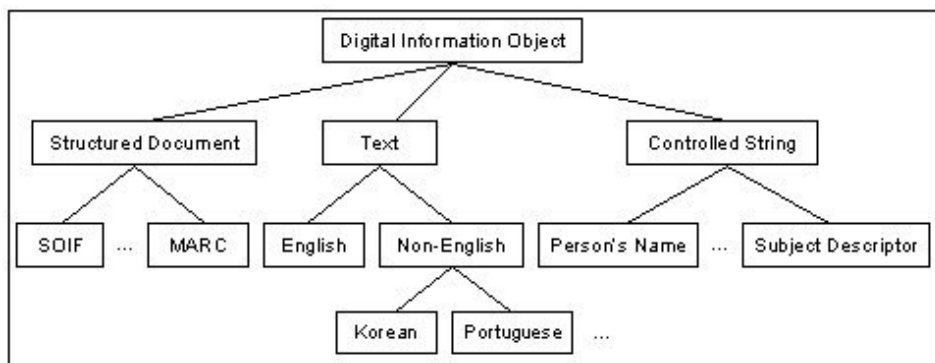


Figure 4. Part of the hierarchy of classes used in MARIAN

Thus the first step in bringing a new document collection into semantic interoperability is to choose appropriate matching functions and searchers for the different objects in the collection. Since class managers and searchers are object-oriented, specialized versions often can be created easily through inheritance. For truly different information objects, new matching functions sometimes need to be defined, but even in this case stock searcher algorithms often can be reused. All that is necessary is to provide methods that follow the API of generating a weight from an object description and, thence, a weighted set of objects.

We have mapped the ETDMS standard into a MARIAN information network model, thus providing a stable common view of the union collection to the outside world. A subset of the ETDMS model is presented in Figure 5; to keep things simple we show only the attributes title, creator, subject, and description. The model consists of three classes of objects, ThesisDissertation, Individual, and Subject, together with HasAuthor and HasSubject links. The Individual class subsumes both persons and corporate individuals, while the Subject class covers diverse treatments. Mappings between this model and the underlying structures can be modified seamlessly.

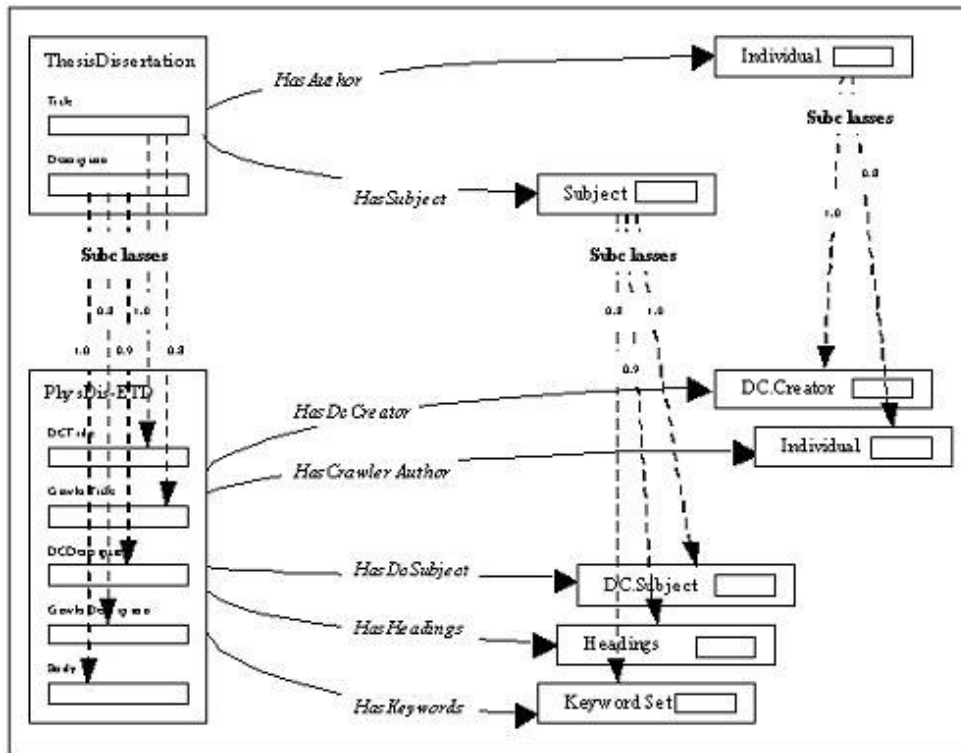


Figure 5. A collection view is derived from the PhysDis data to increase retrieval and usability.

In the particular case of the German PhysDis collection [Severiens, et al., 2000] whose treatment is shown in Figure 5, all mappings make use of the weighted superclass construction. This construction asserts that all members of some specific class are also members of some ETDMS class, but that the extent to which they count as class members is different for different subclasses. In the case of PhysDis subject descriptions, subclass relationships are weighted to reflect the authority of the description. Weights also can be used to address data quality issues. These uses interact; while the simple construct of synthetic superclasses with weighted subclasses cannot handle every situation, we have found it strikingly effective.

Multilingual and Cross-lingual Searching

Supporting a global community means supporting multiple languages. This leads to two research topics: 1) supporting documents and queries in several natural languages simultaneously (multilingual retrieval) and 2) making it possible for queries in one language to retrieve related documents in other languages (cross-lingual retrieval) [Oard, 1997]. In the MARIAN NDLTD union collection, both topics are being investigated in the context of the MARIAN class hierarchy.

Any document (part) made up of natural language can be considered to be an object of the class Text. The MARIAN Text class manager is responsible for all types of text; individual texts are stored by the manager for one or another subclass. These include class managers for text in different natural languages (including English, Spanish, German, and Korean) as well as managers for sublanguages and for personal names, which are presumed to have no linguistic structure. Each natural language class manager can recognize words in the language, generally removing inflectional and morphological affixes. Treatment of sentence structure and layout conventions also may vary among different languages. Text class managers generate indexing structures from the stream of terms discovered in text. They

also make use of weight-valued matching functions to calculate how well any given text matches a free-text query.

Most ETDs are made up of (structures of) text in a particular language. During document analysis, component pieces of text are extracted from the ETD or its metadata and passed to the appropriate class manager in the union collection. Structural information describing the place and function of the text component in the document is also stored in the form of patterns of links. These two sorts of information are used together during retrieval to calculate the overall similarity of a document to a structured query. If the language of a query component is known, it can be relayed to the class manager for that language. On the other hand, in the common case where the language of the query is not known, query components are sent to the Text (super)class manager, which broadcasts the query to all its subclasses. Matches in any language are then returned to the superclass, where they are ranked by closeness to the query, and the combined set becomes the result set for the query component.

Cross-lingual retrieval works within the same class hierarchy. Cross-lingual queries are sent to the Text class manager, which again queries its various subclasses. In this case, however, the Text class manager first translates the query into each subclass language before it passes it on. Thus the (sub)sets returned and combined are composed of matches to the concepts in the query rather than the strings. Current research on translation is based on the work of Akira Maeda and the NAIST Multilingual group [[Maeda, 1998](#)], using co-occurrence statistics to scale and combine different translations.

Other Research efforts

The NDLTD community has explored several other research trends. Experimental software has been developed to add annotation capabilities to ETDs; this service was selected as the most important to add, based on focus groups, to determine the most popular use scenarios [[Miller, 1999](#)]. There also is experimental software extending the SIFT package [[Yan and Garcia-Molina, 1999](#)] from Stanford University and a prototype in the MARIAN system, to provide filtering and routing services based on stored user profiles, for those who wish to be notified whenever an interesting ETD arrives. As time proceeds, our work in interoperability with other digital library software like Greenstone [[Witten, et al., 2001](#)], Phronesis [[Garza-Salazar, 2001](#)], and Emerge [[Futrelle, et al., 2001](#)] may allow us to support other universities that choose to use those packages to provide access services for their local ETDs. Another major research trend of NDLTD deals with user interfaces and information visualization. There are multiple graphical user interfaces that relate to our various software components, including the ENVISION interface [[Heath, et al., 1995](#)]. In addition, ongoing experimentation is investigating how the library metaphor can be extended to virtual reality environments, specifically the CAVE (CAVE Automatic Virtual Environment) [[Das Neves and Fox, 2000](#)].

References

Abiteboul, Serge, Peter Buneman, and Dan Suciu. 2000. *Data on the Web - From Relations to Semistructured Data and XML*. Morgan Kaufmann Publishers, San Francisco, California.

Das Neves, Fernando A. and Edward A. Fox. 2000. A study of user behavior in an immersive virtual environment for digital libraries, in *Proceedings of the ACM Digital Libraries Conference*, pg. 103-111, San Antonio, Texas.

DCMI. 1999. *Dublin Core Metadata Element Set, Version 1.1: Reference Description*. Available from <<http://www.dublincore.org/documents/dces>>.

Fallside, David. 2001. *XML Schema Parts 0,1 and 2*. Available from <http://www.w3.org/XML/Schema#dev>.

Fox, Edward A., Robert K. France, Eskinder Sahle, Amjad M. Daoud, and Ben E. Cline. 1993. Development of a Modern OPAC: From REVTOLC to MARIAN, in *Proceedings of the 16th International ACM SIGIR Conference*, pg. 248-259.

Futrelle, Joe, Su-Shing Chen, and Kevin C. Chang. 2001. NBDL: A CIS Framework for NSDL, in *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '2001)*, pg. 124-125, Roanoke, Virginia, 24-28 June 2001.

Garza-Salazar, David A., 2001. *Phronesis*. Available from <http://copernico.mty.itesm.mx/~tempo/Proyectos/>.

Gonçalves, Marcos A., Robert K. France, Edward A. Fox, and Tamas E. Doszkocs. 2000. MARIAN Searching and Querying of Heterogeneous Federated Digital Libraries, in *Proceedings of First DELOS workshop on Information Seeking, Searching and Querying in Digital Libraries*, Zurich, Switzerland, 11-12 December 2000. http://www.ercim.org/publication/ws-proceedings/DelNoe01/11_Fox.pdf

Gonçalves, Marcos A., Robert K. France, Edward A. Fox. 2001. MARIAN: Flexible Interoperability for Federated Digital Libraries, in *Proceedings of the Fifth European Conference on Research and Advanced Technology for Digital Libraries (ECDL2001)*, Darmstadt, Germany, September 4-9 2001, pp. 173-186.

Heath, Lenwood S., Deborah Hix, Lucy T. Nowell, William C. Wake, Guillermo A. Averbach, Eric Labow, Scott A. Guyer, Dennis J. Brueni, Robert K. France, Kaushal Dalal, and Edward A. Fox. 1995. Envision: A User-Centered Database of Computer Science Literature. *Communications of the ACM*, vol. 38, no. 4, pg. 52-53.

Lagoze, Carl and Herbert Van de Sompel. 2001. *The Open Archives Initiative Protocol for Metadata Harvesting*. Open Archives Initiative. January 2001. Available from <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

Lynch, Clifford. 1997. The Z39.50 Information Retrieval Standard - Part I: A Strategic View of Its Past, Present and Future", *D-Lib Magazine*, April 1997. Available at <http://www.dlib.org/dlib/april97/04lynch.html>.

Maeda, Akira, Myriam Dartois, Takehisa Fujita, Tetsuo Sakaguchi, Shigeo Sugimoto, and Koichi Tabata. 1998. Viewing Multilingual Documents on Your Local Web Browser. *Communications of the ACM*, Vol. 41, No. 4, pp. 64-65, April 1998.

Miller, Todd. 1999. *Annotation system for a collection of ETDs*. Available from <http://www.ndltd.org/ndltd-sc/990416/annsystem.pdf>.

Oard, Douglas W.. 1997. Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries. *D-Lib Magazine*, December 1997. Available at <http://www.dlib.org/dlib/december97/oard/12oard.html>.

OCLC. 2001. *WorldCat*. Available from <http://www.oclc.org/oclc/menu/colpro.htm>.

Ouksel, Aris M. and Amit P. Sheth. 1999. Semantic Interoperability in Global Information Systems: A Brief Introduction to the Research Area. *SIGMOD Record*, vol. 28, no. 1, pg. 5-12.

Powell, James and Edward A. Fox. 1998. Multilingual Federated Searching Across

Heterogeneous Collections. *D-Lib Magazine*, September 1998. Available from <http://www.dlib.org/dlib/september98/powell/09powell.html>.

Severiens, Thomas, M. Hohlfeld, K. Zimmermann, and Eberhard R. Hilf. 2000. PhysDoc - A Distributed Network of Physics Institutions Documents: Collecting, Indexing, and Searching High Quality Documents by using Harvest. *D-Lib Magazine*, Vol. 6, No. 12, December 2000. Available from <http://www.dlib.org/dlib/december00/severiens/12severiens.html>.

VTLS. 2001a. Virtua ILS. Available from <http://www.vtls.com/products/virtua>.

VTLS. 2001b. VTLS Home Page. Available from <http://www.vtls.com>.

Witten, Ian H., Stefan J. Boddie, David Bainbridge, and Rodger J. McNab. 2000. Greenstone: a comprehensive open-source digital library software system, in *Proceedings of the Fifth ACM Digital Libraries Conference*, pg. 113-121.

Yan, Tak W. and Hector Garcia-Molina. 1999. The SIFT Information Dissemination System. *ACM Transactions on Database Systems*, vol. 24, no. 4, pg. 529-565.

Copyright 2001 Hussein Suleman, Anthony Atkins, Marcos A. Gonçalves, Robert K. France, Edward A. Fox, Vinod Chachra, Murray Crowder, and Jeff Young

[Top](#) | [Contents](#)
[Search](#) | [Author Index](#) | [Title Index](#) | [Back Issues](#)
[Previous Article](#) | [Next Article](#)
[Home](#) | [E-mail the Editor](#)

[D-Lib Magazine Access Terms and Conditions](#)

DOI: [10.1045/september2001-suleman-pt2](https://doi.org/10.1045/september2001-suleman-pt2)