MARIAN Searching and Querying across Heterogeneous Federated Digital Libraries

Marcos André Gonçalves, Robert K. France, Edward A. Fox[1], and Tamas E. Doszkocs[2]

Abstract: We explore the complex problem of providing searching services across interoperable heterogeneous federated digital library systems with rich structure and content. We discuss system issues and architectural support in MARIAN, now being extended in the context of implementing a global digital library of electronic theses and dissertations. We consider challenges faced, progress, and future plans.

1. Introduction

Many digital libraries (DLs) provide distributed and autonomous federated information services (e.g., [1]). They often are heterogeneous, since different DLs have evolved from a variety of earlier systems. When considering information seeking, searching, and querying in the context of such complex, distributed and heterogeneous environments, interoperability becomes a major problem [2].

An interesting example of a federated DL requiring interoperability is the Networked Digital Library of Theses and Dissertations (NDLTD, see www.ndltd.org) [3-5], an international federation of universities, libraries, and other supporting institutions interested in worldwide access to electronic theses and dissertations (ETDs). NDLTD faces significant challenges including autonomy of management; decentralization; heterogeneity in all levels (metadata, protocols, repository technologies, language, character coding, nature of the data, user characteristics, preferences, and capabilities); and massive amounts of often highly dynamic data. Furthermore, from the point of view of a collection, NDLTD has unique characteristics: multilingual content, large book-size documents accompanied by multimedia files; availability of full-content in several disparate formats (XML, PDF, etc.) along with large numbers of bibliographic references; rich sets of metadata with different ranges of quality; and diversity in range and scope of user interests.

2. The NDLTD Federated System

To replace our simple federated search system that has been running since late 1997 (see www.theses.org) [6], we have worked since early 2000 to develop a testbed search system that addresses many of the abovementioned challenges. This new system uses a mediation approach as well as an information warehouse / union archive architecture so we can mix federated search and harvesting. It is being developed in the context of the NSF-funded Networked University Digital Library (NUDL, including ETDs and physics information in collaboration with the DFG-funded effort at U. Oldenburg, Germany [7]), and is implemented using our MARIAN system [8-10], an object-oriented information retrieval and DL system developed at Virginia Tech's Digital Library Research Laboratory, with partial support by the National Library of Medicine.

The architecture of the system is presented in Figure 1. The MARIAN Mediation Middleware provides a common query interface. Wrappers overcome barriers of heterogeneity. The XML-based specification language 5SL is used to describe capabilities of remote collections and to feed several data structures inside the wrappers. In our union archive based implementation, information from NDLTD members can be periodically extracted from different sources using several harvesting approaches, processed, merged with information from other sources, and then loaded into a centralized data store – the union archive. Documents are harvested via a number of protocols. The prototype system uses both the Harvest™ package [11] as well as protocols such as Open Archives [12, 13], Dienst [14], and Z39.50 [15]. In particular the Open Archives protocol provides a partial solution for metadata interoperability problems and a simple but powerful harvesting mechanism to overcome heterogeneity barriers between NDLTD members. Processing includes exposing structure in the documents, recognizing terms in text, and generating indexes. The harvested documents are filtered (e.g., for relevance,

[1] Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA, {mgoncalv, france, fox}@vt.edu
[2] National Library of Medicine, doszkocs@tamas.nlm.nih.gov

update-time) and used to update the union archive. With the harvested data stored locally, queries can be posed against the local data without further interaction with the original sources.

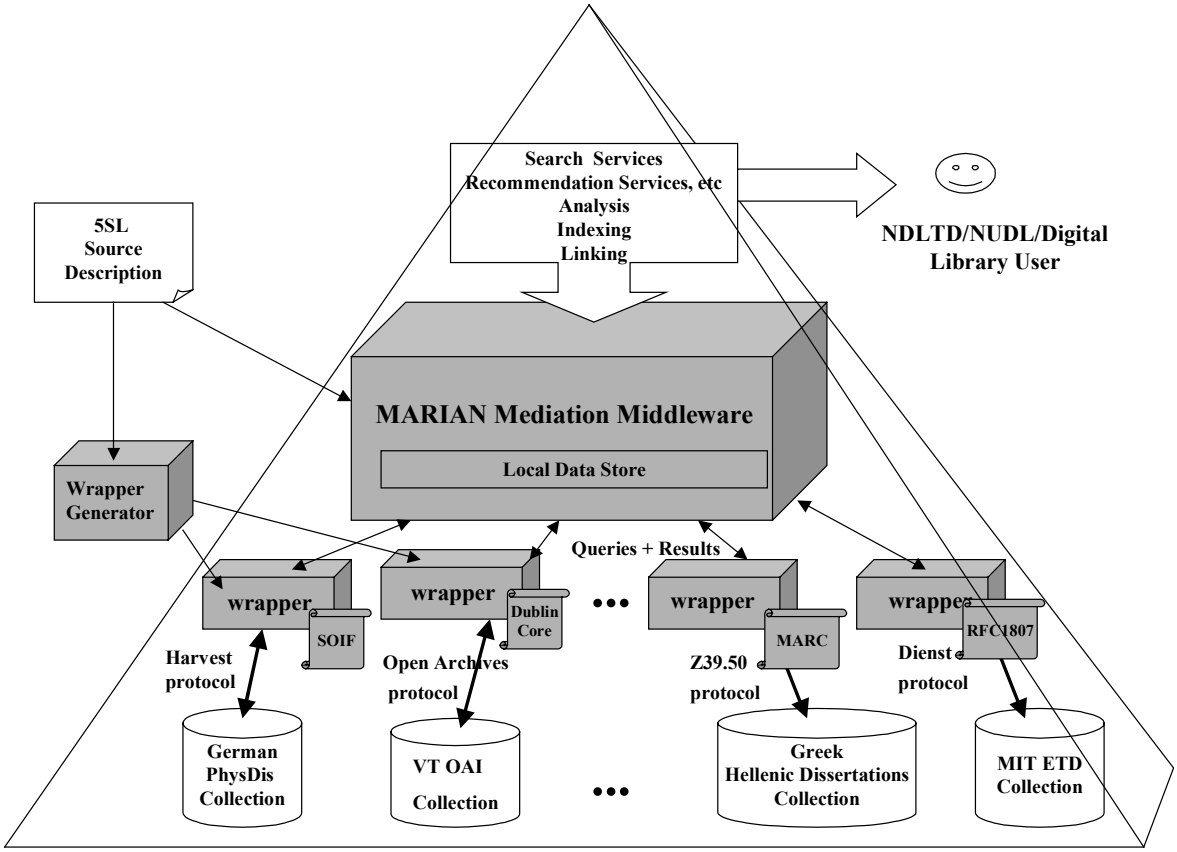

Figure 1. The NDLTD Union Archive Architecture

MARIAN's Mediation Middleware provides a powerful and flexible set of information retrieval functionalities. Digital library objects in MARIAN are represented in an object-oriented model. This approach encourages search methods to be tailored to each class of object and inherited by closely related classes. For instance, structured text documents like theses and dissertations make use of stock MARIAN classes for *StructuredDocument*, *PersonalName*, and *Text*. Particular sorts of documents, like XML theses or Dublin Core [16] metadata records, are specified to have particular structure but inherit general search methods from the parent class. Similarly, the *Text* class defines a particular search method based on an approximate match between a set of query terms and the underlying sequence of terms found in the text, but is specialized into subclasses for different languages. (In the current version of the MARIAN NDLTD union prototype, those languages are English and Non-English European. The class for the latter understands sentence structure but treats all words as un-analyzable strings. Work is currently under way to add Korean, Japanese, and Spanish text classes.) The resulting hierarchy of searchable *DigitalInformationObjects* (Fig. 2) can be added to at any point.

Information object classes can be constructed to search complex objects by implementing class methods that manipulate the objects in idiosyncratic ways. This is the approach used in the Wrapper classes when they function as federated search modules: complex queries are accepted by a Wrapper class, translated into one or more queries in the communication protocol, evaluated against the remote collection, and assembled into a set of matching objects. We are currently exploring this approach to interoperability with local collections as well, including MARIAN classes for Greenstone [17], Phronesis [18, 19], and Emerge collections. Another approach, however, and one that we believe provides for better flexibility and performance, is to expose the structure within the local store as a network of objects and links.

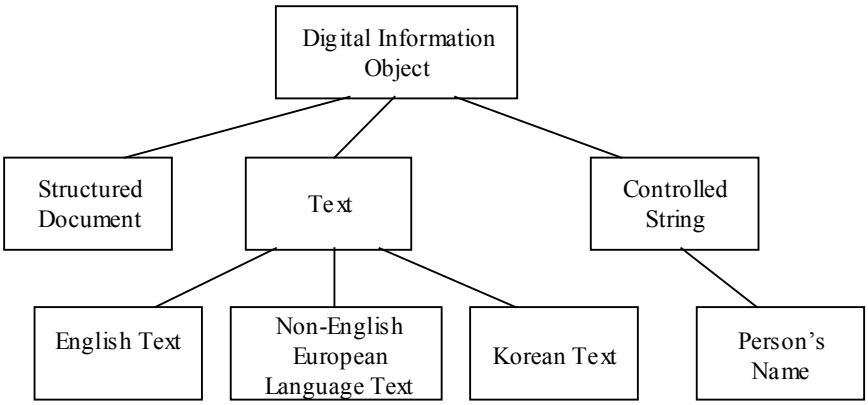

Figure 2. Part of the hierarchy of classes used in MARIAN systems to date

Network representations are commonly used in describing document structure and markup [20]. They have also proven useful in database representation of semi-structured data [21], a category which certainly includes XML and HTML documents. MARIAN is constructed to search through networks of information objects, providing approximate matches to objects within a context of labeled and directed links. For instance, Figure 3 shows how several aspects of document metadata may be usefully represented as links to independent object classes. Stock MARIAN searchers are available for both normal and weighted links, and the basic class *DigitalInformationObject* includes default methods for combining the evidence contained in link contexts to provide an overall figure of merit for a structured query. Each searcher family is extensible to other object classes, ontologies, and models of information.

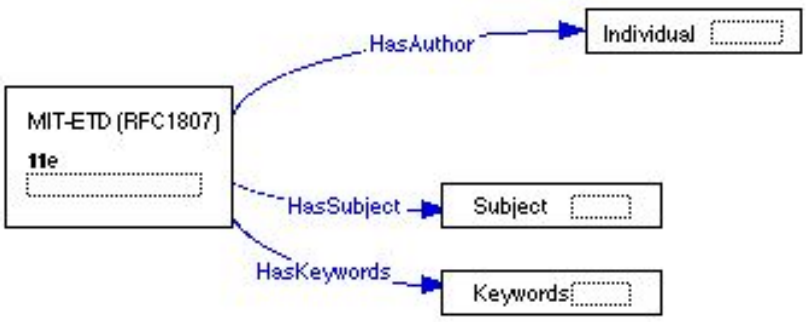

Figure 3. Relevant document structure (here of an ETD in RFC1807 format) is exposed as links to further objects.

Lazy evaluation is another important design principle. All searchers in the MARIAN community do only the work required to return as many elements as are requested. By design and construction, the first results provided by any searcher are those with the highest weight. This minimizes delays resulting from the highly skewed distributions found in large text collections. MARIAN's class-based search engines are built on a single formal model and an Application Program Interface (API) that can be used in the collection infrastructure of a wide range of DL systems.

The MARIAN middleware provides solutions for other inherent problems of federated digital libraries. We address semantic interoperability resolution through sets of object-oriented ontologies of search modules and metadata and through a *collection view* mechanism comparable to database view techniques. Briefly, a collection view is a set of synthetic classes that follow the same model and API as the instantiated classes in the local collection. Unlike the local collection classes, however, the classes making up the view have no members of their own. Instead, they function as weighted superclasses of multiple underlying classes. Mapping provisions allow us to define the structure and content of the view classes in terms of structures in the underlying collection (Fig. 4). In this way, we can provide simpler views to more complex collections, and a unified view to diverse underlying collections. In particular, view classes in the NDLTD prototype enable us to present users

with a single view of the disparate structures in the local images of harvested collections, while still preserving the structure of each harvested collection image. This means that a user can search either the union or the individual collections. What is more, when searching an individual collection, the user can see either the full structure of the collection as it was harvested or the simple universal view.

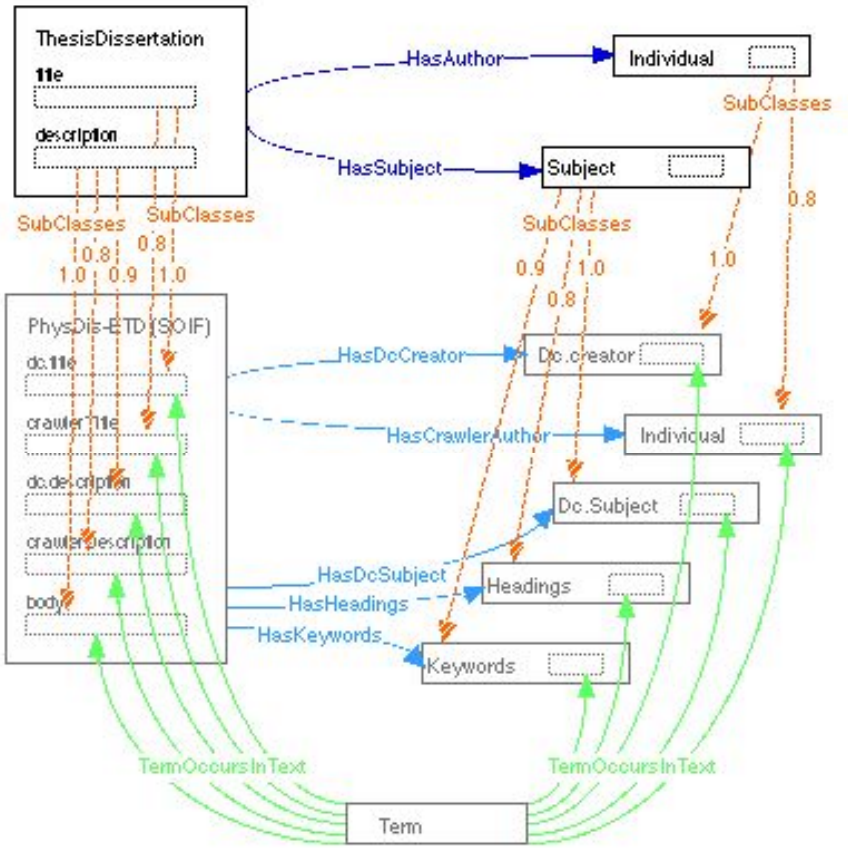

Figure 4. A collection view is composed of view classes (the links and objects in the upper level) defined as weighted superclasses of instantiated classes, one set of which is shown in the lower level.

View classes also help us address the need for data quality. Since the mapping between a view class (structure) and an underlying class (structure) is based on weighted set operations, it is a simple matter to represent that one underlying structure is to be treated as less reliable than another. In Figure 4, for instance, the underlying collection includes both metadata constructed on a Dublin Core model (e.g., *dc.title* and *HasDcCreator)* and information gathered from HTML pages by a crawler. Our evaluation of the data indicated that the DC metadata was generally more reliable. Because of this, we have given it a higher weight in the view. A similar mechanism can be used to smooth statistical variations in the index data caused by, e.g., differences in average text length, and variations in searcher performance. Finally, the middleware architecture addresses such scalability concerns as information compression, the indexing of structured documents, and flexible search.

3. Future Work

To address variations in quality of service in the current network infrastructure, 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.

The network model was chosen in part because it can serve as a carrier not just for structural information, but for other forms as well. For instance, semantic networks and associative networks can be modeled using

unweighted and weighted classes of links. Semantic network classes would replace the current structural search functions with inferential functions, and would allow us to integrate static metathesauri such as UMLS, and NLP techniques, with our existing approximate retrieval functions. Associative network functions (which are not very different from our current matching functions) would enable on-the-fly semantic association networks for ad-hoc queries. Finally, incorporation of *belief network models* [22, 23] 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 are extending our source description module to incorporate a formal *digital library description language* based on the 5S (streams, structures, spaces, scenarios, and societies) framework for DLs [24]. This language will provide better interoperability through a richer representation model and therefore richer query models as well as easy creation of new DL systems and collections. We are also investigating several ways to customize and cluster the union collection to better attend 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) [25] as a storage system for MARIAN along with parallel information retrieval techniques to address issues of scalability and performance. Thus we may meet the needs of NDLTD and of other heterogeneous federated collections.

References

[1] C. Lagoze, "NCSTRL: Networked Computer Science Technical Reference Library", Cornell University, 1999. http://www.ncstrl.org

[2] A. Paepcke, C.-C. K. Chang, H. Garcia-Molina, and T. Winograd, "Interoperability for Digital Libraries Worldwide," *Communications of the ACM*, vol. 41, pp. 33-43, 1998.

[3] C. Phanouriou, N. Kipp, O. Sornil, P. Mather, and E. A. Fox, "A Digital Library for Authors: Recent Progress of the Networked Digital Library of Theses and Dissertations," in *Proceedings of the Fourth ACM Conference on Digital Libraries (DL '99)*. Berkeley, CA: ACM, 1999, pp. 20-27.

[4] E. Fox, "Networked Digital Library of Theses and Dissertations", in *Nature Web Matters*, August 12, 1999. http://helix.nature.com/webmatters/library/library.html

[5] E. A. Fox, "Networked Digital Library of Theses and Dissertations," in *Proceedings DLW15*. Japan: ULIS, 1999. http://www.ndltd.org/pubs/dlw15.doc

[6] J. Powell and E. Fox, "Multilingual Federated Searching Across Heterogeneous Collections," *D-Lib Magazine*, vol. 4, 1998. http://www.dlib.org/dlib/september98/powell/09powell.html

[7] E. R. Hilf, "PhysDis: Physics Theses in Europe", part of *Dissertationen Online*. Home page, 2000. http://elfikom.physik.uni-oldenburg.de/dissonline/PhysDis/dis_europe.html

[8] E. Fox, R. France, E. Sahle, A. Daoud, and B. Cline, "Development of a Modern OPAC: From REVTOLC to MARIAN," in *Proc. 16th Annual Int'l ACM SIGIR Conf. on R&D in Information Retrieval, SIGIR '93*. Pittsburgh: ACM Press, 1993, pp. 248-259.

[9] F. Can, E. Fox, C. Snavely, and R. France, "Incremental Clustering for Very Large Document Databases: Initial MARIAN Experience," *Information Systems*, vol. 84, pp. 101-114, 1995.

[10] R. K. France, "MARIAN Digital Library Information System" (home page), 2000. http://www.dlib.vt.edu/products/marian.html

[11] C. M. Bowman, P. B. Danzig, D. R. Hardy, U. Manber, and M. F. Schwartz, "The Harvest information discovery and access system," *Computer Networks and ISDN Systems*, vol. 28, pp. 119-126, 1995.

[12] H. Van de Sompel, "Open Archives Initiative". WWW site home page. OAI Group, 2000. http://www.openarchives.org

[13] H. Van de Sompel and C. Lagoze, "The Santa Fe Convention of the Open Archives Initiative," *D-Lib Magazine*, vol. 6, 2000. http://www.dlib.org/dlib/february02vandesompel-oai/02vandesompel-oai.html

[14] C. Lagoze and J. R. Davis, "Dienst: An Architecture for Distributed Document Libraries," *Communications of the ACM*, vol. 38, pp. 47, 1995.

[15] International Standard Maintenance Agency_Z39.50, "International Standard Maintenance Agency Z39.50": The Library of Congress Network Development and MARC Standards Office, 2000. http://lcweb.loc.gov/z3950/agency/

[16] S. Weibel, "The State of the Dublin Core Metadata Initiative: April 1999," *D-Lib Magazine*, vol. 5, 1999. http://www.dlib.org/dlib/april99/04weibel.html

[17] I. H. Witten, R. J. McNab, S. J. Boddie, and D. Bainbridge, "Greenstone: A Comprehensive Open-Source Digital Library Software System," in *Proceedings of the Fifth ACM Conference on Digital Libraries: DL '00, June 2-7, 2000, San Antonio, TX*: New York, 2000, pp. 113-121.

[18] D. Garza-Salazar, M. Sordia-Salinas, and Y. Martinez-Trevino, “The Phronesis System: A Practical and Efficient Tool for the Creation of Distributed Digital Libraries on the Internet,” ITESM-Campus Monterrey, Monterrey Technical Report, 1999. http://copernico.mty.itesm.mx/~tempo/Projects/report/

[19] D. Garza-Salazar, “Phronesis Project Web Site”: ITESM-Campus Monterrey, 2000. http://copernico.mty.itesm.mx/~tempo/Proyectos

[20] D. Brickley and R. V. Guha, “Resource Description Framework (RDF) Schema Specification 1.0”: W3C, 2000. http://www.w3.org/TR/rdf-schema/

[21] S. Abiteboul, P. Buneman, and D. Suciu, *Data on the Web: From Relations to Semistructured data and XML*. San Francisco: Morgan Kaufmann, 2000.

[22] B. Ribeiro-Neto and R. Muntz, “A Belief Network Model for IR,” *Proceedings of the 19th ACM-SIGIR conference on research and development in information retrieval*, pp. 253-260, 1996.

[23] I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura, and N. Ziviani, “Link-based and Content-based Evidential Information in a Belief Network Model,” presented at Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information retrieval, Athens, Greece, 2000.

[24] N. Kipp, M. A. Gonçalves, E. A. Fox, and L. T. Watson, “Streams, Structures, Spaces, Scenarions, Societies (5S): A Formal Model for Digital Libraries,” Virginia Tech, in preparation, 2000.

[25] R. M. Akscyn, “The PetaPlex Project,” Status Briefing for National Security Agency, March 9, 1998, 1998. http://ks.com/pages/ksi100.html