Flexible Interoperability in a Federated Digital Library of Theses and Dissertations

Marcos André Gonçalves*, Robert K. France*, Edward A. Fox*,
Eberhard R. Hilf†, Michael Hohlfeld†, Kerstin Zimmermann†, Thomas Severiens†

*Department of Computer Science
Virginia Tech
Blacksburg, VA 24061, USA
Email:{mgoncalv,france,fox}@vt.edu

†Institute for Science Networking
University of Oldenburg
Oldenburg, Germany
Email:{hilf,severien}@uni-oldenburg.de

Abstract

Federated digital libraries are composed of autonomous, possibly heterogeneous information services distributed across the Internet. Federation provides users with a seamless, integrated view of the collected information. We are creating a federated system for the Networked Digital Library of Theses and Dissertations (NDLTD), an international consortium of universities, libraries, and other supporting institutions focused on electronic theses and dissertations (ETDs). The NDLTD allows its members minimal restrictions and maximal autonomy, so federating requires dealing flexibly with differences among ontologies, data formats, and finding aids involving several thousand ETDs in four formats and two languages.

Our solution involves adapting MARIAN, an object-oriented digital library system, to serve as mediation middleware for the federated NDLTD. Components of the solution include: 1) the use of several harvesting techniques; 2) an architecture based on object-oriented ontologies of searchers and metadata; 3) diversity within the harvested data joined to a single collection view for the user; and 4) an integrated framework for addressing such issues as data quality, information compression, and flexible search. The system can handle very large dynamic collections. It can add new sites and adapt to changes in existing sites. MARIAN's modular architecture and powerful and flexible data model work together to build an effective integrated solution within a simple uniform framework.

1. INTRODUCTION

Digital Libraries (DLs) are managed collections of information stored in digital formats, with associated services accessible over a network [Arm00]. DLs have emerged as an important research and application area, facilitated by advances in information technology, such as the World Wide Web (WWW). Networked or federated DLs are composed of autonomous, possibly heterogeneous information services, distributed across the Internet [Lag98, PF98]. The objective of federation is to provide users with a seamless, integrated view of heterogeneous and distributed sources of information. In this paper we focus on one of the most challenging problems in the field of federated DLs: interoperability [PCW+98]. The interoperability problem deals with heterogeneity, occurring in both information representation and services, and at four levels: system, structural, syntactic, and semantic.

An interesting example of a federated DL where heterogeneity is a major problem is the Networked Digital Library of Theses and Dissertations [Pha99] (NDLTD – see Figure 1). Theses and dissertations are prime research results and are thus extremely interesting to other groups working in the same field. They are a major source of new knowledge. NDLTD is an international federation of member universities and other institutional members focused on electronic theses and dissertations (ETDs), who wish to improve graduate education, promote access to scholarly research, increase sharing of knowledge, help universities build their information infrastructure, and extend the beneficial impact of DLs. Many libraries and universities run their own programs and services, but there also are consortia at the state (OhioLINK), regional (Catalunya), and national (Australia, Germany, India, Portugal, South Africa) levels. Currently, more than 100 members are participating in NDLTD. NDLTD has particular characteristics that must be taken into account to support interoperability across member systems:

1. Autonomy: (Groups of) universities manage services for their scholars.
2. Decentralization: Members are not (yet) asked to report collection updates/changes in metadata (i.e., data about data – generally classified as descriptive, structural, or for access management) to central coordinators.
3. Minimal interoperability: Each source must provide unique identifiers (URNs) and metadata records for all stored works, but need not (yet) support the same standards or protocols. [Bal97]
4. Heterogeneity: There is diversity in terms of language, metadata, protocols, repository technologies, character coding, nature of the data (structured, semi-structured and unstructured, along with multimedia) [ABS99], as well as user characteristics, preferences, and capabilities.

5. Massive amount of data and dynamism: NDLTD already has over 100 members and aims to support all those that will produce ETDs, ultimately millions. New members are constantly added and there is a continuing flow of new data as theses and dissertations are submitted.

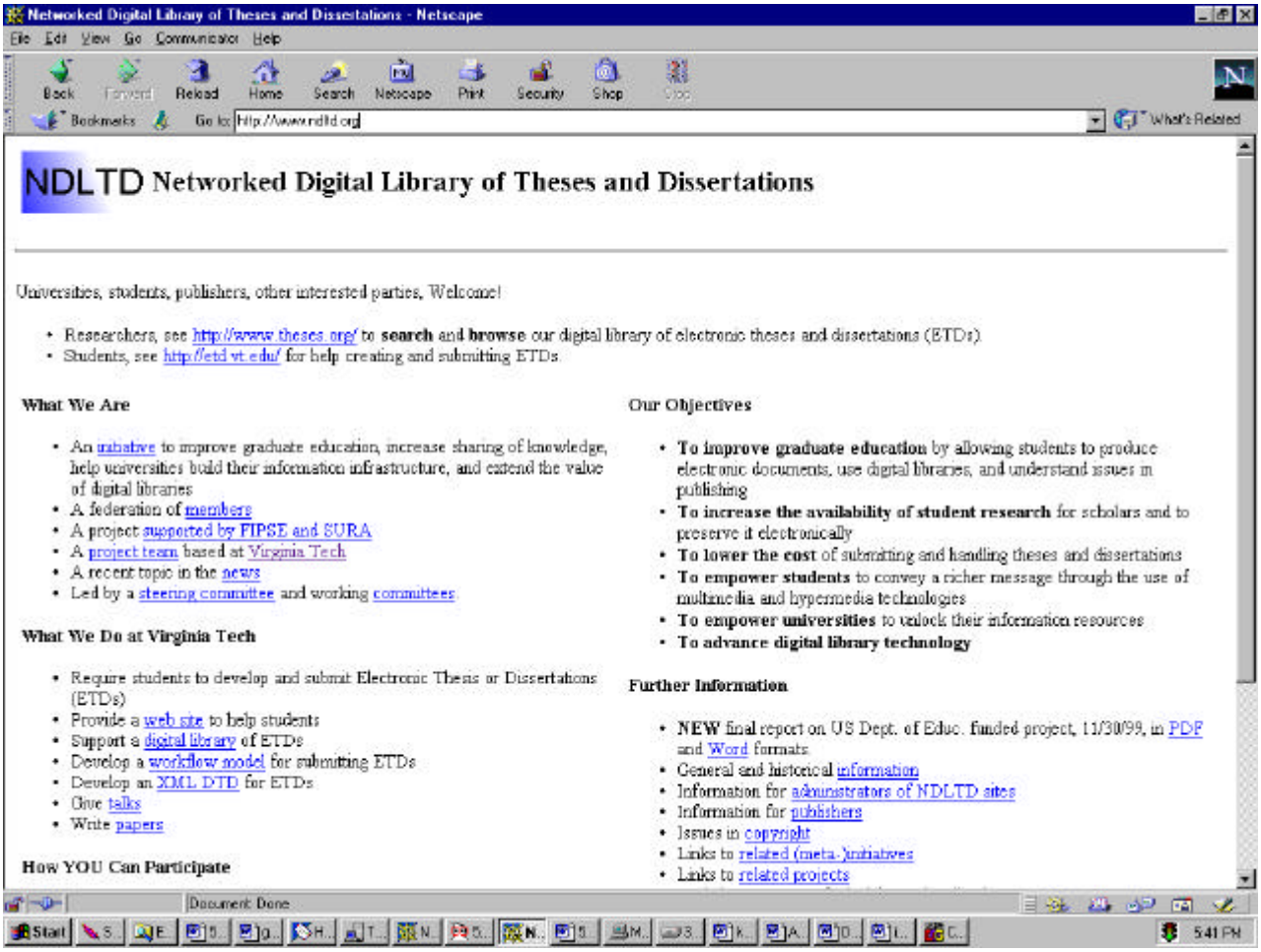

Figure 1. The NDLTD Homepage

Seamless interoperability involves reconciling heterogeneity and integrating information sources at several levels (e.g., collections, services) [Ada00]. The most common architecture to deal with that problem uses mediators and wrappers [Wie92]. Mediators export a common data model of each source's data and provide a common query interface. Wrappers overcome some barriers of heterogeneity and produce source-specific queries. Wrappers also translate results between source and mediator data models. Within the mediated approach there are two possible architectures to deal with the problem of system integration [Flo98], namely: 1) the warehousing or union archive approach; and 2) the federated search approach.

In the federated search solution, data remains at the sources and queries to the integrated system are decomposed at run time into queries to those sources. Data is not replicated and query results are guaranteed to be fresh. On the other hand, sophisticated query optimization and fusion techniques are required. Performance is also a drawback (see, e.g., [Pow00]). Such factors must be considered as network latency and availability, amount of data to be transferred, etc. The overall performance is bounded by the worst-case situation, e.g., the slowest remote site.

NDLTD efforts aim to solve those problems through an architecture based on harvesting mechanisms. The MARIAN DL system provides a common query interface and integration platform. In our union archive

implementation, information from NDLTD members can be periodically extracted from different sources using several harvesting approaches (including the Harvest™ system [BDH+95]and protocols related to the Open Archives Initiative (www.openarchives.org), Dienst [Lag98], and Z39.50 [Lyn97]), processed (e.g., generating indexes), merged with information from other sources, and then loaded into a centralized data store – the union archive. Queries are posed against the local data without further interaction with the original sources and modifications are filtered (for relevance, update-time) and used to update the union archive. Particularly, the use of the harvesting protocol of the Open Archives Initiative provides a partial solution for metadata interoperability problems and a simple but powerful mechanism to overcome heterogeneity barriers between NDLTD members.

This paper is organized as follows. Section 2 describes the MARIAN DL system. Sections 3 and 4 discuss the union archive and the unique characteristics of MARIAN supporting it. Finally, Sections 5 concludes the paper.

2. MARIAN DIGITAL LIBRARY SYTEM

MARIAN is an indexing, search, and retrieval system optimized for DLs [Fox+93, Fra+99]. It can support the large number of simultaneous sessions required for library catalogs (i.e., short sequences of often unrelated queries punctuated by browsing and quick examination of small documents). Thanks to National Library of Medicine funding, MARIAN is almost fully converted to Java – well over 150K lines of code.

The MARIAN data model is based on three main concepts: an *information network* of explicit nodes and links organized into a hierarchy of *classes* in an object oriented fashion where any collection of nodes or links can be *weighted* to represent how well they suit some description or fulfill some role. To illustrate the first and third concepts, we note that a collection of ETDs may include classes of *Document*, *Person*, and *Organization* nodes together with *HasAuthor* and *HasAffiliation* links.

Weights can represent such informal concepts as "importance," "uncertainty," and "goodness of fit". In MARIAN, weights are defined in comparison with other weights, through the construct of a *weighted set*: a set of objects whose relationship to some external proposition is encoded in their (decreasing) weight within the set. For example, weights occur in the weighted *match sets* of objects that satisfy a particular query. There also are weighted classes. Thus, *HasAuthor* might be a weighted link class, if links with different provenance had different amounts of authority. Weights have been successfully used in information retrieval systems, probabilistic reasoning systems, and fuzzy set theory. The MARIAN model extends them uniformly throughout the entire system.

Classes are familiar from object-oriented programming languages (e.g., C++, Java) and databases (e.g., O2). In the context of DLs, MARIAN defines classes of *information objects* that support methods, including calculating how well objects of that class match input descriptions. The classes form hierarchies, and behaviors and semantics are inherited from more general classes to more specific ones. Synthetic union classes are created as needed, e.g., the union class of ETDs is created from classes of structured documents in different formats and sources.

Networks have long been used in traditional library and computer representation systems. Through hypertext and the World-Wide Web, information networks have become commonplace. In recent research, networks have become the preferred representation for semi-structured data, like BibTex, HTML, or XML [ABS99], and for translating among different DL systems [Mel00]. MARIAN search modules ("searchers") are specialized for a universe where searching is distributed over a large graph of information objects.

3. HARVESTING APPROACHES

Any warehouse [Run00] approach must be based on two building blocks: 1) a mechanism to gather or harvest data from sources; and 2) some way of combining that data for use. This section covers harvesting approaches. Section 4 describes our architecture for combining harvested data, aimed to efficiently handle all current approaches.

Electronic theses and dissertations are large, sometimes in the form of several files. Many authors include material that would be difficult or impossible to include in printed publications: audio, images, video, simulations, and large collections of primary data. In response to this, a de facto standard has emerged at NDLTD sites of requiring a *title page*, presented in HTML, to serve both as directory to document files and as a convenient point for collecting and

publishing metadata. These metadata – title, abstract, committee members, subject descriptors, etc. – are created by the author, usually with faculty/staff oversight. At some sites additional metadata is generated by trained catalogers. We choose to harvest all such metadata – both controlled and uncontrolled – to create images of the sites in the union archive. We do not harvest the documents, choosing instead to leave them at the remote sites.

Many popular systems provide only a single mean of harvesting, concentrating for instance on HTML documents on the Web. Much of current work on federated DLs assumes a homogeneous structure or protocol. (e.g., Dienst [Lag98] and Z39.50 [Lyn97] – both supported by MARIAN). We have been working with two different paradigms for harvesting data from heterogeneous sites: the paradigm proposed by the Open Archives Initiative and the one used in the Harvest™ system. In addition, a variety of data has been harvested using ad-hoc source-oriented approaches. The three approaches differ mainly in the support they require from source archives.

3.1 The Open Archive Initiative

The Open Archives Initiative (OAI) is a project to address interoperability of archives and DLs. Originally, the two main support mechanisms, as described in the Santa Fe Convention [SL00, SKN+00], were a metadata model (OAMS) based on the Dublin Core metadata set, and a protocol (based on Dienst) for harvesting metadata. Our early work followed the original specifications, but present efforts are being revised in accord with what will be the official specifications, to be released in January 2001. The OAI framework promotes an effective partial solution for interoperability, but particular archives must agree on implementing the new protocol and on exporting XML-encoded (meta)data using the Dublin Core standard (as well as optionally in other formats, like the MARC standard format for library interchange). The initiative emphasizes the distinction between data providers and service providers. The former may be the manager of an e-print archive, acting on behalf of the authors submitting documents to the archive. The latter is a third party, creating end-user services based on data in archives. In our efforts concerning OAI, we act as both data and service providers. Our work of making MARIAN compliant with OAI is concentrated on three fronts: 1) making MARIAN serve as a harvester and a mediation middleware layer, able to deal with the heterogeneity of many specific sources and protocols, including OAI sources; 2) providing a search service whereby MARIAN, using a light-weight Java implementation; can work with open archives; and 3) developing an XML transport mechanism for MARC records so open archives can export them.

3.2 Harvest™ system

The Harvest™ system [BDH+95] corresponds to a set of integrated and customizable tools for harvesting information from diverse repositories and building topic-specific content indexes. The architecture of the system is based on two main components: *gatherers* and *brokers*. Gatherers collect and extract indexing data from repositories. They act as directed crawlers able to get information from topic-specific listed sources. Several parameters can be configured to provide better performance and guarantee data quality from the information gathered. Gatherers extract summaries of content from harvested sources into a specific proprietary format (SOIF).

Brokers provide the indexing and query interface to the gathered information [Zim00b]. They retrieve information from one or more Gatherers or other brokers and incrementally update indexes. Brokers also can filter or refine the information from other brokers. In reference to the OAI, brokers can be seen as indexing and searching services over data harvested by the gatherers. Unlike in the OAI, however, no metadata standard is forced. The gathering crawler both extracts existing document attributes and itself generates some meta-information, like timestamps and identifiers. External metadata standards (e.g., Dublin Core) can be followed. Some specific and useful entry points in the collection (e.g., the URL of a file that is a list of metadata records pointing to ETDs) must be provided to allow effective control of the crawler and guarantee quality of the generated data.

3.3 Other data sources

Since NDLTD will be harvesting metadata from hundreds of universities, it must accommodate existing sources while encouraging suitable standards (e.g, OAI, Z39.50, or Dienst) to make the process of adding new sites more efficient). Because Virginia Tech functions as coordinator of NDLTD efforts, we developed ad hoc ways to gather metadata in heterogeneous formats. A case in point is the Virginia Tech ETD collection, which provides us with

dumps of local thesis and dissertation metadata by exporting from an SQL database. The obvious drawback to ad hoc conversions is that they require development of specific solutions that are strongly dependent on the source.

Z39.50 is well suited to federated search. Also, with strong server support, Z39.50 can enable harvesting. MARIAN can import MARC records harvested through Z39.50 servers. MARIAN also can use the Dienst protocol to access data at repositories that do not support the harvesting protocol used in OAI.

4. THE UNION ARCHIVE

The NDLTD union catalog uses a mediation / wrapper architecture, modified to make it more extensible (Figure 2). We have prepared a special XML-based language (called 5SL) to describe DLs. This allows us to semi-automatically generate wrappers. Further, in our union archive approach, we use several object-oriented metadata document classes, to produce an intermediate MARIAN representation useful for stemming, parsing, and indexing.

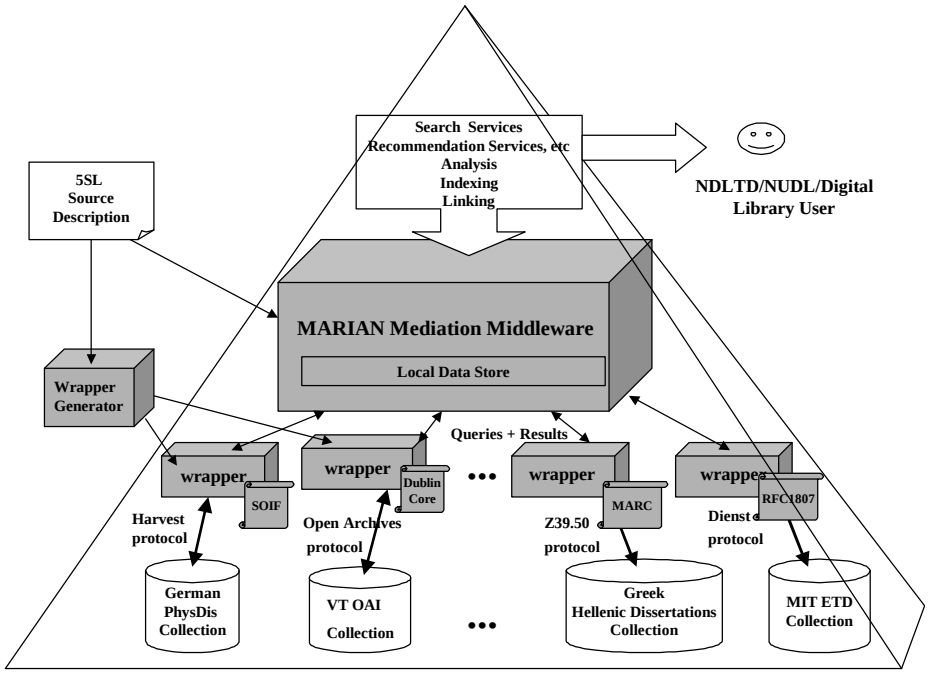

Figure 2. The NDLTD Union Archive Architecture

Among the participating institutions in NDLTD are diverse collections, especially regarding document format. NDLTD does not specify the format in which participant institutions maintain either documents or their metadata [Zim00a]. Though standards are emerging for both, we now must handle multiple formats. In the prototype union collection described here, we have harvested metadata in four formats. The German Physics collection of ETDs (see Section 4.2) comes to us in Harvest´s SOIF format, including both Dublin Core and uncontrolled attributes produced by their local crawler. The Virginia Tech ETD collection is available both as metadata created by authors, exported in the old Open Archives Metadata Standard (OAMS), and as records created by professional catalogers in MARC format. And the MIT ETD collection comes to us as RFC1807 records received via the Dienst protocol. Our approach is to work with the metadata as it arrives rather than trying to translate these different formats into a single format. We fuse search results from the several collections as we present a unified view to users. Advantages of this solution will be explored below.

The primary device for unification is the *union superclass*. Information objects in the MARIAN system are all part of a class hierarchy. All ETDs and all metadata formats are subclasses of *StructuredDocument.* We begin by defining a class for the local image of the document collection belonging to each institution. In a case like Virginia

Tech where we have disparate but overlapping sources of metadata, we create a class for each format and a general superclass to mediate among them. Finally, we define the synthetic superclass *ThesisDissertation* (Figure 3). Since all the subclasses involved are *StructuredDocuments*, any *ThesisDissertation* will be a *StructuredDocument* as well; mapping its structure onto the structure of the subclasses is the main part of its class definition.

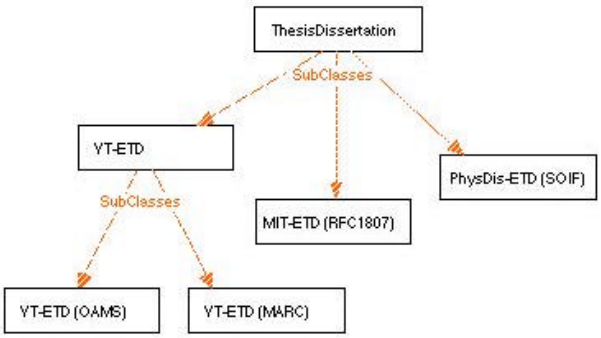

Figure 3: The ThesisDissertation class is a superclass of ETDs from separate institutions. Some institutions (in this prototype, Virginia Tech) may support separate versions of their collection that need to be merged before being reported.

Once we have harvested metadata from each remote collection and built local images for each, we can treat the local data with a unified set of text parsing, indexing and retrieval tools. Document (metadata) text fields such as *title*, *abstract*, or *bod*y are reduced to their individual terms using the same set of parsers, then matched to users' queries using the same search algorithms and ranking formula. This way we can ensure that the smallest atomic components, the text fields, will receive uniform treatment.

The next problem to address in combining collections is that these atomic text components serve different purposes in different collections; or in other words that the structure of documents is different in different collections. Different collections support different document attributes, and represent those attributes with different structures of data. Similar structures can be given different names by different collections. Structures with similar names may have very different semantics. Finally, the same purpose can be addressed by semantically different fields.

Even within collections, there may be differences in document structure. MARC records in the VT-ETD collection distinguish between personal and corporate authors, while the <*creator*> field of Dublin Core records may contain either. As another example, some documents from the PhysDis collection are represented with Dublin Core metadata, including *dc.subject*, while others describe the subject with lists of automatically extracted keywords.

4.1. Presenting collection views

Thus we see that documents in the harvested collections have multiple attributes – some redundant and some complementary. (This same situation can obtain within a single source collection; see the discussion of the PhysDis collection in Section 4.2.) This complexity is potentially confusing for users of the union collection. We address these issues first by studying a collection. Then, in our information network model, we create a *collection view* ontology, analogous to a database view. The view is composed of superclasses of entities and attributes from member ontologies, which are combined by the searchers for each superclass to create the presented view. Thus, each remote collection image in the union collection has a different structure (Figure 4). We use node and link classes to faithfully capture that structure, thus minimizing labor expended and information lost during harvesting.

For purposes of simplicity and familiarity, we have chosen to present a global view based on the Dublin Core model (Figure 5). Four attributes are presented to the user: title, author, subject, and description. In accordance with MARIAN's networked information model, the view ontology actually 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, and the *Subject* class covers a welter of possible treatments. This

view can be modified or extended as future usability requires. More importantly, the connections between the view and the underlying structure can be modified without affecting what the user sees.

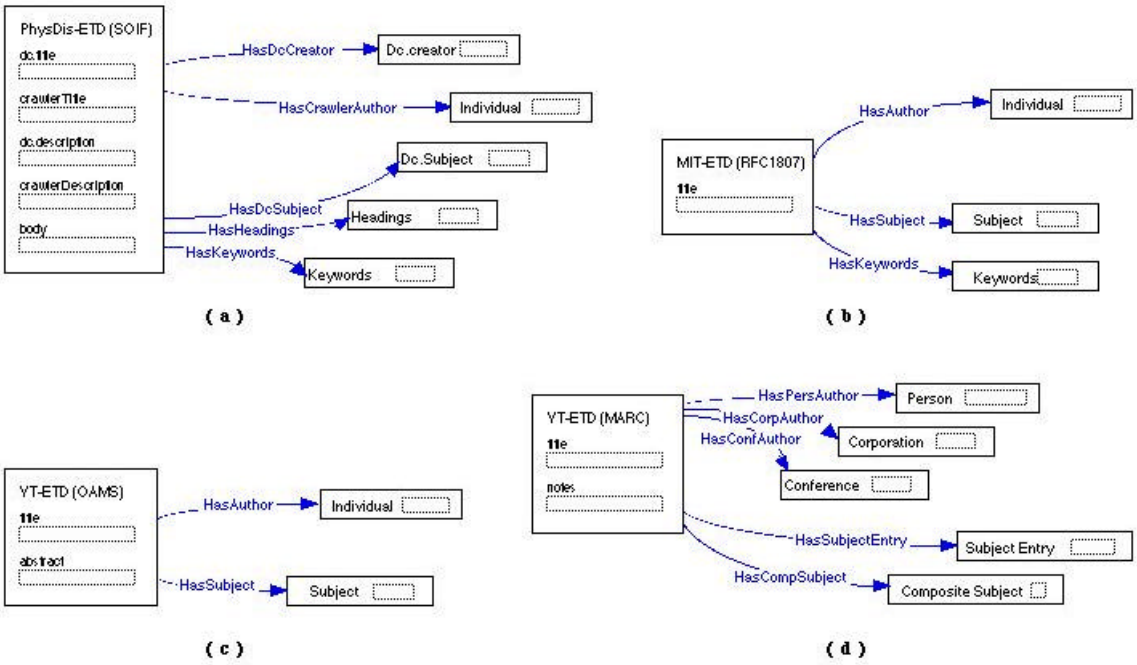

Figure 4: Images for (a) the SOIF PhysDis collection, (b) the MIT RFC1807 collection, (c) the VT OAMS collection, and (d) the VT MARC collection, all represented as class networks.

Figure 5: The synthetic view presented to users consists of the ThesisDissertation class, a class of Individuals linked to ThesisDissertations by an authorship relation, and a linked class of Subject descriptions.

4.2. Extended example: the PhysDis Collection

To illustrate the complexity of representing heterogeneous collections, we consider PhysDoc [Bor97, Hil94, Hil96a, Hil96b, SHZ+00,]. PhysDis (http://elfikom.physik.uni-oldenburg.de/dissonline/PhysDis/dis_europe.html, see Figure 6), a key part of PhysDoc, is a discipline-based service run by the University of Oldenburg, Germany, which collects dissertations on physics [Zim99]. There are 250 links to collections of these works, leading to 1818 datasets, including 250 online full text dissertations. TheO (http://www.iwi-iuk.org/dienste/TheO) is a search engine with broader coverage: all dissertations in Germany in all learned fields. PhysDis is part of the German project Dissertationen Online (in German at www.dissonline.de, in English at www.dissonline.org), a project of the IuK Initiative Information and Communication of the Learned Societies in Germany, supported by the German Science Foundation (Deutsche Forschungsgemeinschaft, DFG). Within Germany, the German National Library (Die Deutsche Bibliothek, DDB) serves as an archive, the university libraries as local repositories for postings on WWW, the departments as sources of key parts of the metadata, and the learned societies for the search engines and handling of related publications. According to German law theses have to be published. This can be done through microfiche, printed copies given to the university and from there to the national library, articles in professional

journals, or books handled by a publisher. In addition, online distribution is accepted now as a publication channel. The author owns the copyright, regardless of who distributes it.

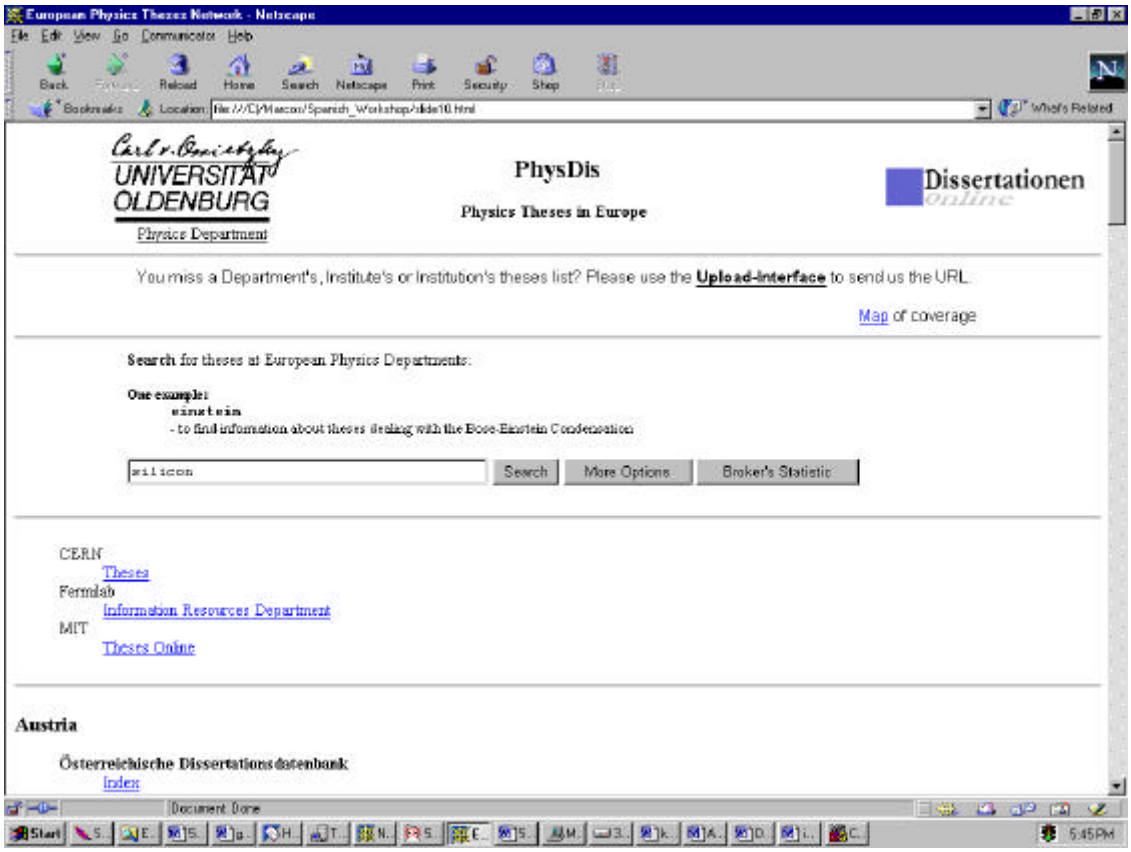

Figure 6. The PhysDis collection

Using the Harvest system, we have harvested 1256 documents from this collection. All of these include various fields in SOIF format, but no single field is present in every document. In addition, 166 of the documents contain controlled Dublin Core metadata, which is presented together with other uncontrolled metadata in the SOIF record. The two sorts of metadata are neither closely related nor mutually exclusive. All documents presenting Dublin Core metadata also present some form of uncontrolled metadata but there is no single uncontrolled attribute occurring in every Dublin Core document. The nearest is *keywords*, which occurs in 97% of documents with Dublin Core attributes while appearing in only 88% of the collection as a whole. There is little overlap between the values in Dublin Core and Harvest metadata attributes even where the attributes can be expected to have similar semantics, in spite of the fact that a small but significant fraction of documents have both Dublin Core and uncontrolled attributes. When we consider the values that appear in the attributes, the differences appear in significant ways. This can be attributed, at least in part, to imprecision on the part of the local crawler, which generates titles like "Dissertation" (58 occurrences), "Dissertationen" (52 occurrences), and "Archiv Publikationen URZ TU Chemnitz" (42 occurrences).

Of the document attributes returned by the Harvest software for the PhysDis collection, we have chosen to represent ten in our union collection: the Dublin Core attributes *dc.creator, dc.description. dc.subject* and dc.*title*, and the free attributes *author, body, description, headings, keywords*, and *title*. We regard the various sorts of titles and descriptions, as well as *body*, as attributes of the documents themselves, while regarding the names in *author* and *dc.creator,* as well as the descriptive strings in *dc.subject, headings* and *keywords* as first-class objects connected to the documents with links (Figure 4a).

We maintain all this structure and complexity in the local image of the PhysDis collection. At the same time we want to be able to present a simplified view to the user. We serve both goals, as well as the ultimate goal of providing a simple view of the complete global union collection of ETDs, by the judicious use of ontologies of superclasses representing collection views.

The PhysDis collection provides a good example of view presentation and use of weights to enhance data quality. Each text class in the view (Figure 7) corresponds to two or more classes in the underlying collection: a Dublin Core class and at least one uncontrolled class. The Dublin Core texts generally are of better quality than the uncontrolled texts. The superclass searchers capitalize on this by giving more weight to DC subclasses. In addition, the *Description* superclass depends more heavily on the PhysDis *body* attribute than on either DC or uncontrolled description attributes, because the *body* text tends to be a better representation of document content. All of these weights can be tuned as our experience with the union collection increases.

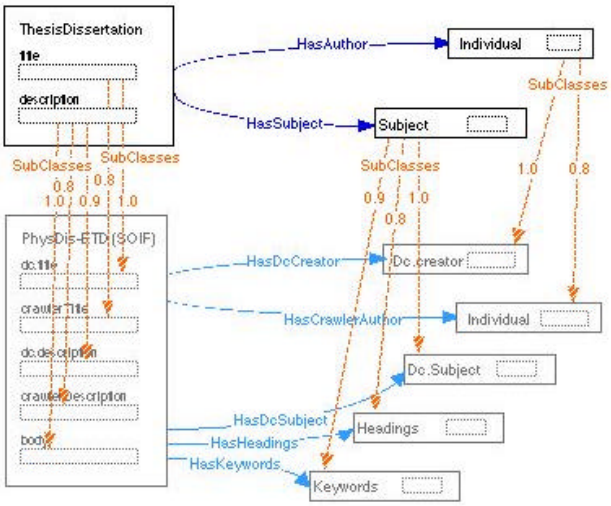

Figure 7: The collection view is abstracted from the PhysDis data to increase retrieval and usability

Similar connections are made between the view superclasses and the other ETD collections, with weights chosen to maximize reliability and effectiveness of retrieval. In addition, weights can be tuned between the classes, both to promote one subclass over another, as in the PhysDis example, and to mediate artifactual differences between the collections. An example of the latter would be normalizing similarity values computed for the various collections. Most of the weight-values functions used to measure similarity between a query and a document (text) are sensitive to collection size or the distribution of attributes or terms across the collection. We can impose adjustments at the superclass levels to mitigate these disparities.

5. Conclusions

We have developed MARIAN to provide seamless and effective support for heterogeneous distributed digital libraries. Its application to NDLTD demonstrates the value of our approach to the problem of DL interoperability.

References

[Arm00] Arms, W., *Digital Libraries*, Cambridge, MA: MIT Press, 1999

[ABS99] Abiteboul, S., Buneman, P. Suciu, D., *Data on the Web: from relations to semistructured data and XML*. San Francisco, CA: Morgan Kaufmann, 1999

[Ada00] Adam, N., Atluri, V., Adiwijaya, I., "Systems Integration in Digital Libraries", *Commun. ACM*, 43(6):64-72, 2000

[Bal97] Baldonado, M., et al., "The Stanford Digital Library metadata architecture." *Int'l Journal on Digital Libraries*, 1997. 1(2): 108-121

[Bor97] Borghoff, Uwe M., Eberhard R. Hilf, Remo Pareschi, Thomas Severiens, Heinrich Stamerjohanns, and Jutta Willamowski, "Agent-Based Document Retrieval for the European Physicists: A Project Overview", in Proc.

PAAM'97; Second International Conference and Exhibition, 21-23 April 1997, London, UK, http://www.physik.uni-oldenburg.de/documents/UOL-THEO3-97-3/

[BDH+95] Bowman, C. M., Danzig, P. B., Hardy, D. R., Manber, U., Schwartz, M. F., "The Harvest information discovery and access system", *Computer Networks and ISDN Systems*, 28(1–2) 119–126, 1995

[Fox+93] Fox, E.A., R.K. France, E. Sahle, A.M. Daoud, and B.E. Cline: "Development of a Modern OPAC: From REVTOLC to MARIAN. *Proc. of the 16th International ACM SIGIR Conference on R&D in Information Retrieval*, 1993, 248-259

[Fra+99] France, R.K., L.T. Nowell, E.A. Fox, R.A. Saad, and J. Zhao: "Use and usability in a digital library search system." Feb. 1, 1999, CoRR xxx.cs.DL/9902013, http://xxx.lanl.gov/abs/cs.DL/9902013

[Flo98] Florescu,D., Levy,A., Mendelzon,A. "Database techniques for the World-Wide Web: A Survey", *SIGMOD Record* 27(3):59-74,1998

[Hil94] Hilf, E.R., "Integrated Information Management in Physics", Online Proceedings of APS E-PRINT Workshop, 14 October 1994, Los Alamos, NM, http://publish.aps.org/EPRINT/KATHD/ebs.html

[Hil96a] Hilf, E.R., G. Rohen, T. Severiens, "Electronic Information Management in Physics", Proceedings Software-Entwicklung in der Chemie 10, ed. J. Gasteiger, pp. 89-96, published by GDCh Gesellschaft Deutscher Chemiker, 1996, http://www2.ccc.uni-erlangen.de/tagung/10_cic/hilf/index.html

[Hil96b] Hilf, E.R., B. Diekmann, H. Stamerjohanns, J. Curdes, "Integrated Information Management for Physics", in Proc. The Information Revolution: Impact on Science and Technology, J.-E. Dubois, N. Gershon (Eds.); Springer-Verlag Berlin Heidelberg (1996), p.189-196. http://www.physik.uni-oldenburg.de/documents/UOL-THEO3-95-2/codata7/

[Lag98] Lagoze, C., Fielding, D., Payette, S., "Making Digital Libraries Work: Collection, Services, Connectivity Regions, and Collection Views", *Proc. 3rd ACM Conference On Digital Libraries*. 1998, pp. 134-143

[Lyn97] Lynch, C., "The Z39.50 Information Retrieval Standard - Part I: A Strategic View of Its Past, Present and Future", *D-Lib Magazine*, April 1997, 3(4), http://www.dlib.org/dlib/april97/04lynch.html

[Mel00] Melnik, S., H. Garcia-Molina and A. Paepcke, "A mediation infrastructure for digital library services." *Proc. 5th ACM Conference on Digital Libraries* (San Antonio, June 2-7, 2000), pp.123-132

[PCW+98] Paepcke, A., Chang, C. K., Winograd, T., Garcia-Molina, H., "Interoperability for digital libraries worldwide." *Communications of the ACM* 41(4): 33–42, 1998

[Pha99] Phanouriou, C., Kipp, N. A., and Sornil, O., Mather, P., and Fox, E. A., "A Digital Library for Authors: Recent Progress of the Networked Digital Library of Theses and Dissertations", *Proc. 4th ACM Conference on Digital Libraries*, 1999, pp. 20-27

[Pow00] Powell, A.L. and J.C. French, "Growth and server availability of the NCSTRL digital library." *Proc. 5th ACM Conference on Digital Libraries* (San Antonio, June 2-7, 2000), pp. 264-265.

[PF98] Powell, J., Fox, E. A., "Multilingual Federated Searching Across Heterogenous Collections", *D-Lib Magazine*, September 1998, 4(9), http://www.dlib.org/dlib/september98/powell/09powell.html

[Run00] Rundensteiner, E., Koeller, A., and Zhang, X., "Maintaining Data Warehouses over Changing Information Sources", *Communications of the ACM*, 43(6): 57-62, 2000

[SHZ+00] Severiens, T., Hohlfeld, M., Zimmermann, K., Hilf, E.R., "PhysDoc: A Distributed Network of Physics Institutions Documents: Collecting, Indexing, and Searching High Quality Documents by Using Harvest", D-Lib Magazine, December 2000, 6(12), http://www.dlib.org/dlib/december00/severiens/12severiens.html

[SL00] Sompel, H., Lagoze, C., "The Santa Fe Convention of the Open Archives Initiative", *D-Lib Magazine*, Feb. 2000, 6(2)

[SKN+00] Van de Sompel, H., Krichel, T., Nelson, M.L., Hochstenbach, P., Lyapunov, V.M., Maly, K., Zubair, M.K., Liu, X., O'Connell, H., "The UPS Prototype project: exploring the obstacles in creating a cross-print archive end-user service", *D-Lib Magazine*, Feb. 2000, 6(2), http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html

[Wie92] Wiederhold, G., "Mediators in the Architecture of Future Information Systems", *IEEE Computer*, 25(3): 38-49, 1992

[Zim99] Zimmermann, K., "Dissertationen Online", Proc. CRISP II (Coop. Research Info. Systems in Physics II), Oldenburg, 11-13 Oct. 1999, http://elfikom.physik.uni-oldenburg.de/crisp2/talks/Zimmermann.pdf

[Zim00a] Zimmermann, K., "Interoperability of Retrieval", ETD/XML Workshop, Berlin, May 2000, http://www.dissonline.org/berlin/

[Zim00b] Zimmermann, K., Interoperability of Retrieval: Harvest as Search Engine, International Women's University (ifu), Hamburg, Aug. 2000, http://www.physik.uni-oldenburg.de/~kerstin/hamburg.ppt

Acknowledgements:

We thank DFG for support of the work at University of Oldenburg and NSF (through grant IIS-9986089) for their support of the work at Virginia Tech. We thank the many faculty, staff, and students who have assisted in prior related work, and our many partners involved in NDLTD, PhysDoc, Dissertationen Online, OAI, etc.