

# MARIAN: Flexible Interoperability for Federated Digital Libraries

Marcos André Gonçalves, Robert K. France and Edward A. Fox

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

**Abstract.** Federated digital libraries are composed of distributed, autonomous, and often heterogeneous information services but provide users with a transparent, integrated view of collected information. In this paper we discuss 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). Federation requires dealing flexibly with differences among systems, ontologies, and data formats while respecting information sources' autonomy. Our solution involves adapting the object-oriented digital library system MARIAN to serve as mediation middleware for the federated NDLTD collection. Components of the solution include: 1) the use and integration of several harvesting techniques; 2) an architecture based on object-oriented ontologies of search modules and metadata; 3) reconciliation of diversity within the harvested data joined to a single collection view for the user; and 4) an integrated framework for addressing such questions as data quality, flexible and efficient search, and scalability.

## Introduction

Networked or federated digital libraries are composed of autonomous, possibly heterogeneous information services, distributed across the Internet [14, 10]. The goal of federation is to provide users with a transparent, integrated view of such sources of information. Challenges faced include interoperability amongst different digital library systems/protocols [22], resource discovery (selection of the best sites to be searched) [12], and issues in data fusion (merging of results into a unique ranked list). In this paper we focus on the interoperability problem, one of the most challenging in the field of digital libraries. Heterogeneity occurs in both information representation and services, and must be addressed at four basic levels: system, structural, syntactic, and semantic [20].

One federated digital library where heterogeneity is a major problem is the Networked Digital Library of Theses and Dissertations [23], an international federation of

universities, libraries, and other supporting institutions focused on efforts related to electronic theses and dissertations (ETDs). NDLTD has particular characteristics that complicate interoperability across member systems:

1. **Autonomy:** Members manage most services for their scholars.
2. **Decentralization:** Members are not (yet) asked to report either collection updates or changes in their metadata to central coordinators.
3. **Minimal interoperability:** Each source must provide unique URNs and metadata records for all stored works, but need not (yet) support the same standards or protocols.
4. **Heterogeneity:** Members differ in language, metadata, protocols, repository technologies, character coding, nature of data (structured, semi-structured and unstructured, multimedia), user characteristics, preferences, and capabilities.
5. **Massive amount of data and dynamism:** NDLTD already has over 100 members and eventually aims to support all those that will produce ETDs. New members are constantly added and there is a continuing flow of new data as theses and dissertations are submitted.

Due to the primary-source nature of ETD collections, the site selection process that is found in other systems (identifying a small number of candidate databases to search) is not always important here. For example, a query asking for new results in mathematics could retrieve information from almost every member university.

## 1. Federated Systems: Remote Search vs. Local Union

Transparent interoperability involves reconciling heterogeneity and integrating information sources at several levels [2]. A common architecture to deal with this problem uses mediators and wrappers [30]. Mediators export a common data model 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 architecture there are two possible approaches to system integration [8]: 1) the union archive and 2) federated search.

In the union archive approach [26], information is periodically extracted from each source, processed, 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. The main advantage of this approach is that adequate performance can be guaranteed at query time. On the other hand, union archives cannot guarantee delivery of the most current information to users. Concerns about data quality and consistency also must be addressed.

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 is guaranteed to be fresh at query time. On the other hand, more sophisticated query optimization and fusion techniques are required. Performance is also a drawback (see, e.g., [25]). Such factors must be considered as network latency and availability, amount of data to be transferred, etc. Overall performance is bounded by the worst-case situation.

In this paper we present MARIAN, an object-oriented digital library system, and demonstrate how we have used its modular architecture, flexible data model, and powerful search mechanism to create a federated system for NDLTD while addressing the problems described above. Due to poor and inconsistent network connectivities in the global NDLTD, variability in server load and administration, and the complexity of query translations in such a heterogeneous environment, we have chosen a union archive architecture for our integrated system. Components of our solution include: 1) the use and integration of several harvesting techniques; 2) a mediated union archive collection based on object-oriented ontologies of search modules and metadata; 3) a *collection view* mechanism for network representations comparable to database views; and 4) an integrated framework for addressing such questions as data quality, flexible and efficient search, and scalability. We use the unique characteristics of our system to build a common integrated solution for interoperability inside a unified framework.

## 2. The MARIAN Digital Library System

MARIAN is a search system for digital libraries [5, 7]. Originally designed for 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 [27] 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 [1], and for translating among different DL systems [17]. 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. 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 [6] 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 oft 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 [11]. The flexibility of MARIAN's 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 and programming languages or artificial intelligence [9].

### 3. Harvesting Approaches

Any union archive approach includes: 1) mechanisms to gather or harvest data from the sources, and 2) some way of combining gathered data for use. This section covers harvesting approaches; Section 4 describes our architecture for combining harvested data.

Electronic theses and dissertations (ETDs) are large, sometimes archived in the form of several files. Many authors include multimedia material that would be difficult or impossible to include in printed publications. In response to this, a *de facto* standard has emerged at NDLTD sites of requiring a structured *title page* to serve both as directory to document files and as a convenient point for collecting and publishing metadata. Title page metadata are created by the author, usually with minimal oversight. At some sites additional metadata are added by trained catalogers. We choose to harvest all metadata – both controlled and uncontrolled – to create images of the sites in the union archive.

Much current work on federated DLs assumes a homogeneous structure or protocol (e.g., Dienst [14] or Z39.50 [16]) or a single means of harvesting (e.g., of HTML documents on the Web). In contrast, we work with several paradigms for harvesting data from heterogeneous sites, including the paradigms of the Open Archives Initiative and Harvest™. In addition, a variety of data has been harvested using ad-hoc source-oriented approaches. The three approaches differ in the support they require from source archives.

The Open Archives Initiative (OAI) [15] is a multi-institutional project to address interoperability of archives and digital libraries by defining simple protocols for the exchange of metadata. The current OAI technical infrastructure is expressed by the Metadata Harvesting Protocol, which defines mechanisms for archives to expose and export their metadata. The OAI framework promotes an effective partial solution for interoperability, but particular archives must agree to implement the protocol and to

export their metadata in a supported standard, which creates impedance to the solution. OAI emphasizes the distinction between data providers and service providers. The former manages a resource such as an e-print archive, acting on behalf of the authors who submit documents. The latter is a third party, creating end-user services based on data in archives.

The Harvest<sup>TM</sup> system [3] is a set of integrated 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 act as directed crawlers that collect and extract indexing and meta-information from repositories extracting summaries of content into a specific proprietary format (SOIF). Brokers provide the indexing and the query interface to the gathered information. They retrieve information from one or more Gatherers or other Brokers and incrementally update indexes. Although no metadata standard is enforced, external metadata standards (e.g., Dublin Core) can be incorporated.

We have faced situations where we cannot use any of these approaches, but where specific ways to gather data from sources exist. For example, in sources that use the Dienst protocol, specific combinations of services allow harvesting their data. The obvious drawback to ad hoc conversions is that they require development of specific solutions that are strongly dependent on the source.

#### 4. The NDLTD Union Archive

In the prototype union collection described here, we have harvested metadata from four sources, each with its own formats. Table 1 summarizes the characteristics of each.

Collection	Harvesting Protocol	No. of records	Metadata format
PhysDis-ETD	Harvest	1256	SOIF – All DublinCore – 166
VT-ETD	OAI Z39.50	2427	ETD-MS – All MARC – All
MIT-ETD	Dienst	5000	RFC1807 – All

**Table 1.** Collections in the prototype union archive and their characteristics.

Just as there are many differences among institutions participating in NDLTD, there also are differences among the collections, especially regarding document format and access protocols. NDLTD did not specify standard formats or access protocols for documents or metadata. Although the adoption of standards is encouraged for NDLTD, it will be some time until a complete standardization takes place. Consequently our current union collection must cope with a multiplicity of formats, systems, protocols, etc.

Also, different collections support different document attributes and represent those attributes with different structures of data. Similar structures can be given different names

by different sources, and structures with similar names may have very different semantics. For example, MARC records in the VT-ETD collection make a strong distinction between personal and corporate authors, while the *dc.creator* field of Dublin Core records may contain either. Again, some documents from the PhysDis collection are represented with Dublin Core metadata, including *dc.subject*, while others describe subjects with lists of automatically extracted keywords.

Thus, heterogeneity has several dimensions and induces four levels of interoperability concerns [20]: 1) system: which involves for example differences in harvesting protocols; 2) syntactic: including machine-dependable aspects of data representation; 3) structural: involving representational heterogeneity; and 4) semantic: with all the complexities related to meanings, significations, uncertainty, etc. In the following, we present how we use the unique characteristics of MARIAN to build an interoperability architecture that attacks each level of interoperability.

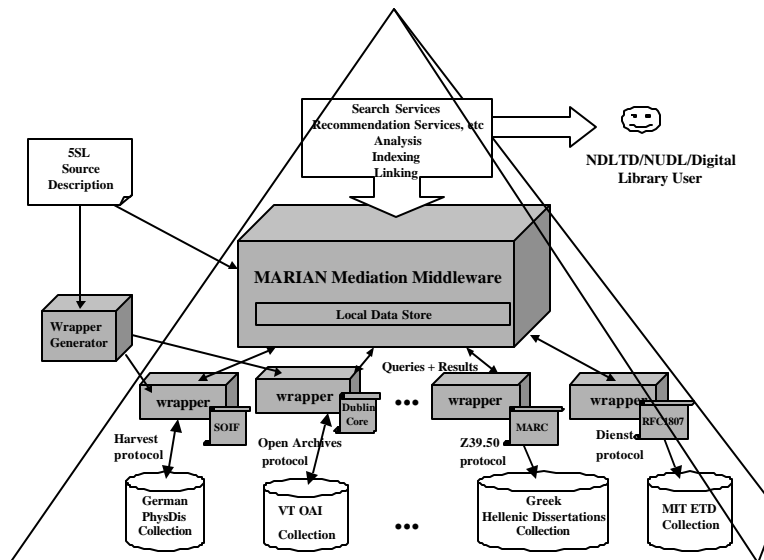


Fig. 1. The NDLTD Union Archive Architecture

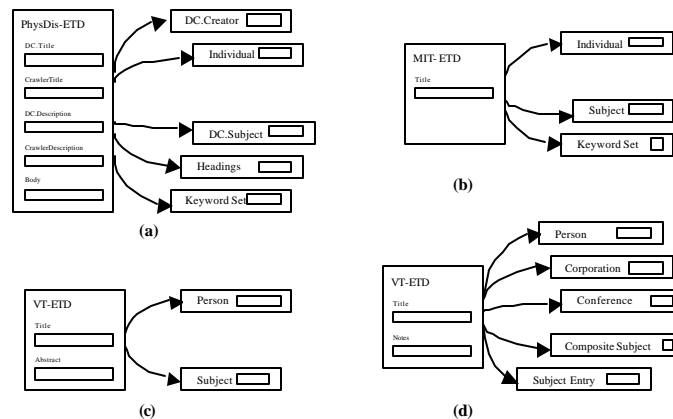
#### 4.1 MARIAN's Interoperability Architecture

The architecture of our system is presented in Fig. 1. The MARIAN Mediation Middleware provides the tools for structural and semantic interoperability. System and syntactic differences are addressed by wrapping sources with special software modules. Our SSL language for declarative specification of digital libraries is used to describe

capabilities of remote collections and their internal document structures. This information feeds data structures inside the mediator and allows semi-automatic generation of wrappers for harvested sources. Extended value-added services like searching, browsing, recommendation, personalization, and visualization are built on the top of the middleware.

## 4.2 System and Syntactic Interoperability

The harvesting process itself serves as a device to suppress some differences in source systems such as indexes and formatting, and helps towards systemic and syntactic homogenization. For example, textual information in different languages with different encodings can be locally homogenized to some standard like Unicode or UTF-8. 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 *body* 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.



**Figure 2.** Images for (a) the SOIF PhysDis collection, (b) the RFC1807 MIT collection, (c) the ETD-MS VT collection, and (d) the MARC VT collection, all represented as class networks. Upward-curving links are (subclasses of) *HasAuthor* links; downward links, *HasSubject* links.

## 4.3 Structural Interoperability

Mediators map different representations of heterogeneous data sources to a common data model. Like many current approaches for data integration/mediation [4, 19], we use MARIAN's network representation to overcome structural heterogeneity, capturing the

structure of the remote collections as faithfully as possible. Figure 2 represents the document structures in each of our experimental collections.

In contrast to other network approaches, MARIAN's nodes and links are associated with object-oriented classes, which give us three major advantages. First, instead of using a single global searcher for the entire network, nodes and links are partitioned among class managers for a marked decrease in search complexity. Second, indexing and search are regarded as functional aspects of the classes, and thus can capitalize on regularities of the class. Third, the hierarchy of classes and search mechanisms provide a basis for the next phase of resolution of semantic interoperability.

#### 4.4 Semantic Interoperability

Semantic heterogeneity is solved by exploiting two further 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, but also they 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 person's names 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 match the description. MARIAN already has stock matching functions and searchers for a number of common information object classes, a sample of which are shown in Fig.3.

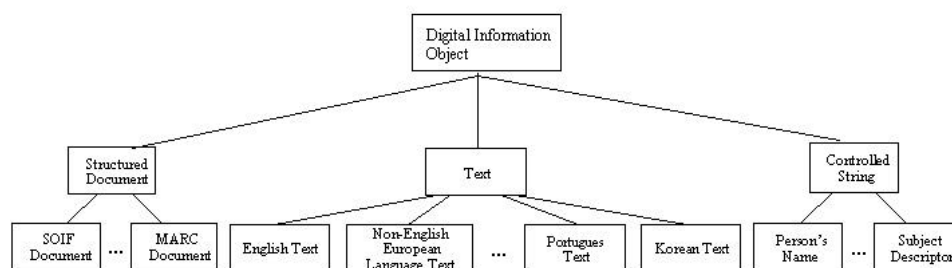
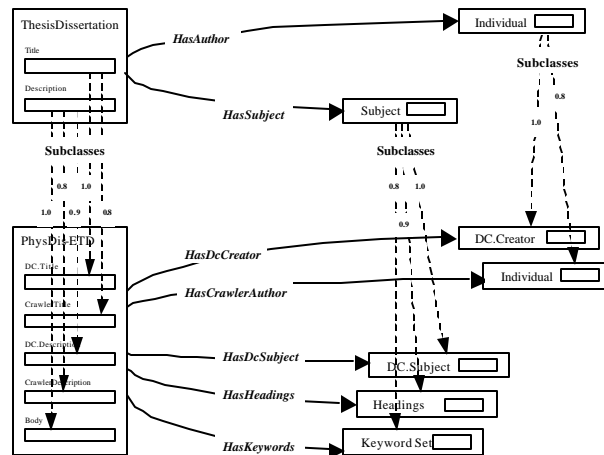


Figure 3. Part of the hierarchy of classes used in MARIAN.

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 can often be easily created through inheritance. For truly different information objects new matching functions sometimes need to be defined, but even in this case stock searcher algorithms can often be reused. All that is necessary is to provide methods that follow the API of taking an object description to a weight or weighted set of objects.

Once a local image has been defined for an NDLTD member collection a view can be constructed. This involves defining a mapping to the member collection classes from a supported view. Such a mapping may use any combination of linking, inheritance and weighting. In the remainder of this paper we concentrate on one mechanism that has proven powerful and useful in NDLTD: the creation of synthetic weighted superclasses.

In NDLTD we are fortunate that a complementary interoperability effort [<http://www.ndltd.org/standards/metadata/>] has developed a metadata standard for electronic theses and dissertations (ETD-MS). Mapped into an information network model, this standard provides a stable view to the outside world for the union collection. A subset of the ETD-MS view is presented in Figure 4; to keep things simple we show only the attributes *title*, *creator*, *subject*, and *description*. The view ontology 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 the view and the underlying structures can be modified seamlessly.

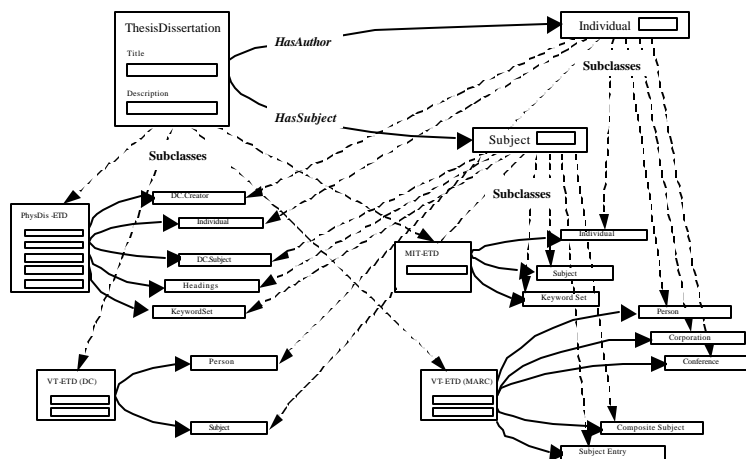


**Figure 4.** A collection view is abstracted from the PhysDis data to increase retrieval and usability.

In the case of the ETD-MS view of the PhysDis collection shown in Figure 4, all mappings make use of the weighted superclass construction. This construction asserts that all members of some specific class also are members of some view 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. In the next section we discuss the use of weights to address data quality issues. These uses interact, and the simple construct of synthetic superclasses with weighted subclasses cannot handle every situation, but we have found it strikingly effective.

#### 4.5. Combining Heterogeneous Collections and Merging Ontologies

A direct approach for combining collection images into a union collection is depicted in Figure 5. This solution involves a fair amount of redundancy. For instance, several source images include classes for *Individual* and *Subject*. Such redundancy raises a design issue: what should be the ontology for the overall union collection?



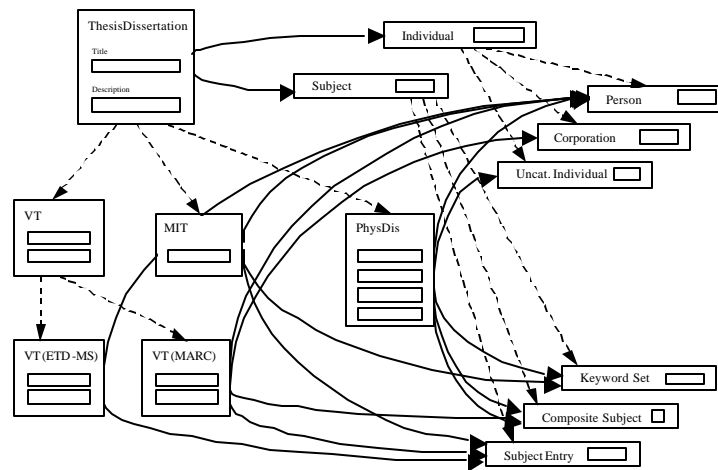
**Figure 5.** A direct approach for the union collection relates the views and images.

Figure 5 embodies one extreme where all the images are completely separate and only subclass-superclass relationships tie the object classes together. This approach has the disadvantage of data duplication: the same object (e.g., the subject heading “Computer Engineering”) appears in several classes. Such redundancy wastes storage space in the class managers, and increases retrieval time when multiple classes are searched.

At the other extreme, we could immediately force all harvested data into our collection view by processing all incoming documents into structures with a single title and a single description field, all types of individuals into a single class, and all types of subjects and keyword lists into a class of subject strings. This would have the disadvantage of forcing

us to combine fields as unlike as the PhysDis *Body* and *dc.description* fields into a single text, with corresponding losses to indexing specificity. It also would mean losing the information that sometimes we *do* know when an individual is a person, or when a subject heading comes from a controlled vocabulary.

Most importantly, however, pre-processing incoming data into the collection view ontology would mean giving up the ability to adjust to changing circumstances. Once our image of a remote collection has been cooked, we can no longer reconstruct it in its raw state. On the other hand, the more original structure we retain, the better we can react to changes in the original collection, to addition of new collections, and especially to changes in semantic requirements.



**Figure 6.** A more sensitive approach to the union catalog allows semantically similar object classes to overlap.

Between these two extremes lies a third alternative: merge image classes when these have sufficiently similar semantics, but keep classes separate when the semantics are different. Figure 6 shows this approach for the four images in the NDLTD union collection. The document hierarchy is as before. The *Individual* class has been analyzed into classes of (human) *Persons* and *Corporations*. SOIF, RFC1807 and Dublin Core author fields and MARC x00 fields, all of which require the name of a human person, are mapped to *HasAuthor* links to the *Person* class, while the MARC x10 fields produce links to the *Corporation* class. An *UncategorizedIndividual* class provides an image for those formats that make no such distinction, like the uncontrolled *author* field of the PhysDis collection. A similar breakdown of the *Subject* superclass into individual subject entries, composite strings with multiple entries, and sets of keywords provides image classes for all types of subject fields in the union collection. This approach simplifies the union collection

ontology, with corresponding benefits in administration time and effort. It also saves string storage and retrieval time and adds functionality.

## 5. Solution Analysis

Combining weights, networks, and class structures enables us to both respect the data as it is harvested and provide simplified virtual collection views to users. It also makes it easy to change either the collection ontology or the underlying data without changing the view presented to the user, or to change the view presented to the user without restructuring the underlying representation or data. Moreover, it provides a unified framework to enhance retrieval effectiveness in the union archive system by providing the flexibility to use different configurations and priorities on the same underlying data. In this section, we consider some properties we consider essential in any solution for interoperability.

### 5.1 Data Quality Issues

Data quality issues arise when one wants to correct anomalies occurring inside the integrated union archive to improve effectiveness of services. Examples of anomalies include errors in data, imprecision, multiplicity of representations, etc. In our architecture, we use MARIAN's weighting scheme as a way to mitigate data quality discrepancies.

The PhysDis collection provides a good example of the use of weights to enhance data quality. Each text class in the view corresponds to two or three classes in the underlying collection: a Dublin Core class and at least one uncontrolled class (Fig. 5). Our observations of the data indicate that the Dublin Core texts 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 we have observed that *Body* text tends to be a better representation of document content. All of these weights can be tuned with the increasing of experience with the union collection and with empirical experiments.

### 5.2 Efficiency

As seen, each MARIAN class manager functions as a searcher for objects in that class. All the searchers needed for the union archive are of five standard types: superclass searchers, text and structured document searchers, and weighted and absolute link searchers. The searchers are designed for optimal efficiency using three rubrics:

1. *Use all available information about the inputs*: searching can be a costly operation. Under certain circumstances, however, we can use information about the incoming sets to achieve better performance
2. *Capitalize on the power-law distribution*: Our observations of links in text and among digital library objects indicate that like small-world networks [29] they follow power-law distributions. Whenever possible, the MARIAN searchers are designed to run most efficiently when their inputs follow this distribution.
3. *Be lazy*: All searchers in the MARIAN community are designed to do only the work required to return as many elements as are requested. By design and construction, the first elements developed by any searcher are those with the highest weight. Lazy evaluation has its greatest pay-off in simple searches authored by human users, few of whom are interested in digesting more than a few dozen objects.

MARIAN searchers have been used for collections of up to a million objects and tens of millions of links, most noticeably in a “shadow” of the Virginia Tech academic library [7]. Response times for simple queries on collections of this size are comparable to other Net searchers, and remain acceptable for more complex queries. Research is currently under way to measure performance on collections of hundreds of millions of objects, and to verify the power-law model and its implications for searcher efficiency.

### 5.3 Scalability

Scalability, i.e., the ability to transparently and effectively grow a system, is a major concern in any platform based on integration of external sources. As important as large capacity are scalable query processing and the ability to incorporate new sources.

Just as analyzed objects are represented in MARIAN by graphs, queries are represented by relaxations of graphs and are processed following their structure. Elements of a MARIAN query are distributed to their governing class managers. Each class manager disassembles the portion of the query it receives; any parts it cannot handle are passed to others. Thus, an author / title search over the entire collection begins at the *Thesis-Dissertation* class. The title portion of the query is handled locally, but the author portion is passed to the link class manager for *HasAuthor* links, which passes the operation of finding matching people or corporations to the *Individual* class manager. Since query processing is distributed, it avoids evaluation bottlenecks and easily can be extended.

Easy incorporation of new sources is achieved by a mechanism for semi-automatically generating wrappers. Harvesting itself tremendously facilitates creating wrappers and communication with the mediator. In contrast to wrappers from other DL interoperation projects, which are *query processing oriented*, our wrappers are oriented towards *schema* and *ontological equivalence*. We have been using 5SL, a domain-specific language with a formal basis [13] designed for automatic generation of digital libraries, for describing external sources' capabilities for harvesting purposes. We describe external sources by

their metadata structures and the correspondent harvesting protocols as scenarios (abstract sequences of events) that occur between the harvester/wrappers and the sources. Template wrappers can be configured (e.g., for periodicity of harvesting, additional mirrored repositories, or maximum number of harvesters acting at the same time). The wrappers currently in use in the union catalog, while originally crafted by hand, have since been successfully generated automatically from 5SL descriptions.

## References

- [1] Abiteboul, S., Buneman, P. Suci, D., *Data on the Web: from relations to semistructured data and XML*. Morgan Kaufmann, 1999
- [2] Adam, N., Atluri, V., Adiwijaya, I., "Systems Integration in Digital Libraries", *Communications of the ACM*, **43**(6), 2000, pp. 64-72
- [3] 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), 1995, pp. 119-126
- [4] Fernandez, M. F., Florescu, D., Levy, A. Y., Suci, D. "Declarative Specification of Web Sites with Strudel". *VLDB Journal* 9(1): 38-55 (2000)
- [5] 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. 16<sup>th</sup> Int. ACM SIGIR Conf.*, 1993: pp. 248-259
- [6] France, R.K. "Weights and Measures: an Axiomatic Approach to Similarity Computations". Internal report, Virginia Tech, 1995; <http://www.dlib.vt.edu/reports/WeightsMeasures.pdf>
- [7] France, R.K., L.T. Nowell, E.A. Fox, R.A. Saad, and J. Zhao: "Use and usability in a digital library search system." CoRR cs.DL/9902013:
- [8] Florescu, D., Levy, A., Mendelzon, A. "Database techniques for the World-Wide Web: A Survey", *SIGMOD Record*. **27**(3) 1998, pp. 59-74
- [9] Fuhr, N., Rolleke, T., "A Probabilistic Relational Algebra for the Integration of Information retrieval and Database Systems", *ACM Transactions on Information Systems*, Vol. 15, No. 1, January, 1997, Pg. 32-66.
- [10] Fuhr, N. "A Decision-Theoretic Approach to Database Selection in Networked IR". *ACM Transactions on Information Systems* 17(3): 229-249 (1999)
- [11] Fuhr, N., "Towards Data Abstraction in Networked Information Retrieval Systems", *Information Processing and Management* 35(2): 101-119 (1999)
- [12] Gravano, L., Garcia-Molina, H., "Merging Ranks from Heterogeneous Internet Sources", *Proc. of the 23<sup>rd</sup> International Conference on Very Large Databases, 1997*, pp. 196-205
- [13] Gonçalves, M.A., Kipp, N.A., Fox, E.A., Watson, L.T., "Streams, Structures, Spaces, Scenarios and Societies(5S): A Formal Model for Digital Libraries", Tech. Rep., Virginia Tech, 2001.
- [14] Lagoze, C., Fielding, D., Payette, S., "Making Digital Libraries Work: Collection, Services, Connectivity Regions, and Collection Views", *Proc. 3<sup>rd</sup> ACM Digital Libraries*. 1998, pp.134-143
- [15] Lagoze, C., Sompel, H. V., "The Open Archives Initiative", Proc. of the First ACM-IEEE The Joint Conference on Digital Libraries, Roanoke, Virginia, 2001.
- [16] Lynch, C., "The Z39.50 Information Retrieval Standard - Part I: A Strategic View of Its Past, Present and Future", *D-Lib Magazine*, April 1997.
- [17] Melnik, S., H. Garcia-Molina and A. Paepcke, "A Mediation infrastructure for digital library services" *Proc. 5<sup>th</sup> ACM Digital Libraries, San Antonio, 2000* pp.123-132.
- [19] McBrien, P., Pouloussilis, A., "Automatic Migration and Wrapping of Database Applications - A Schema Transformation Approach". ER 1999: 96-113
- [20] Ouksel, A. M., Sheth, A. P., "Semantic Interoperability in Global Information Systems: A Brief Introduction to the Research Area" *SIGMOD Record* 28(1):5-12 1999
- [22] Paepcke, A., Chang, C. K., Winograd, T., Garcia-Molina, H., "Interoperability for digital libraries worldwide." *Communications of the ACM* **41**(4), 1998, pp. 33-42.
- [23] Phanouriou, C., Kipp, N. A., Sornil, O., Mather, P., Fox, E. A., "A Digital Library for Authors: Recent Progress of the NDLTD", *Proc. 4<sup>th</sup> ACM Digital Libraries*, 1999, pp. 20-27
- [25] Powell, A.L. and J.C. French, "Growth and server availability of the NCSTRL digital library." *Proc. 5<sup>th</sup> ACM Conf. On Digital Libraries (San Antonio, June 2-7, 2000)* pp. 264-265.
- [26] Rundensteiner, E., Koeller, A., and Zhang, X., "Maintaining Data Warehouses over Changing Information Sources", *Communications of the ACM*, **43**(6), 2000, pp. 57-62
- [27] Semantic Web Activity; <http://www.w3.org/2001/sw/>
- [29] Watts, D. J., "Small Worlds: The Dynamics of Networks between Order and Randomness", Princeton Univ. Press, 1999.
- [30] Wiederhold, G., "Mediators in the Architecture of Future Information Systems", *IEEE Computer*, **25**(3), 1992, pg. 38-49.