COMMENTARY

D-Lib Magazine
November 2005

Volume 11 Number 11

ISSN 1082-9873

What Is a Digital Library Anymore, Anyway?

Beyond Search and Access in the NSDL

Carl Lagoze[i], Dean B. Krafft[i], Sandy Payette[i], Susan Jesuroga[ii]

[i]Computing and Information Science, Cornell University, Ithaca, NY
{lagoze, dean, payette}@cs.cornell.edu

[ii]UCAR-NSDL, Boulder, CO
jesuroga@ucar.edu

Based on a variety of calibrations[1], we are now in the adolescence of digital libraries. Like any adolescence, there is reason for optimism and concern.

The optimism comes from the successes resulting from a decade of research, development, and deployment. Any list of these is inevitably incomplete, but certainly includes Google[2], the Handle System® [3], Dublin Core[4], OAI-PMH[5] (Open Archives Initiative Protocol for Metadata Harvesting), OpenURL [40], arXiv[6], Dspace [53], and LOCKSS [50]. These and other accomplishments, in combination with the general explosion of the Web itself over the last decade and a half[7], approach the digital library vision of "Universal Access to Human Knowledge" articulated in the 2001 report of the President's Information Technology Advisory Committee [45].

The concern originates in part from the vexing problems that were addressed in early digital library workshops [8, 34] for which practical solutions have yet to be fully deployed. A few examples are illustrative. While Handles and DOIs[8] have been successfully deployed in the library and publishing communities, the goal of ubiquitous, persistent identifiers remains unmet. Widespread acceptance of Dublin Core and OAI-PMH appears to address initial goals for interoperable resource description. However, problems with the quality of metadata [58] compromise the utility of the standards. Federated identity middleware such as Shibboleth[9] begins to address authorization and authentication issues, but the underlying public key infrastructure that was seen as "essential to the emergence of digital libraries" [34] remains undeveloped. Despite efforts of the W3C's Semantic Web initiative [13], the holy grail of semantic interoperability [42] remains elusive. Finally, with increasing amounts of rich

information born in digital form and stored in institutional repositories, we still lack standard, scalable techniques for fully preserving that information.

These technical snags exist in a broader and perhaps more worrisome institutional dilemma, which has been characterized by many[10] as the "googlization" of digital libraries and information in general. Like any neologism, googlization is used in a variety of ways. Here, it refers to the troublesome misconception that Google represents the apotheosis of digital information and that the remaining problems in this domain have either been solved or will be solved by Google (or perhaps by Yahoo!, MSN, etc.). Informal discussions with colleagues in the digital library research community indicate that googlization has infected funding agencies, both public and private. While the absence of a well-funded digital library research program within the National Science Foundation can be attributed to a number of causes, the notion that "Google has solved the problem" is contributory.

Google's accomplishments are indeed striking. But this "End of History"[11] myopia appears to be the result of confusion about "what is a (digital) library?". Perhaps driven by early misguided utopian visions like Al Gore's "schoolchild in Carthage, Tennessee" [9] comment, there seems to be a belief that a digital library is just about search ("can I find it?") and access ("can I get it?"). These functions are indeed essential (and remain challenging), but they are just part of an information environment. Traditional libraries are much more than well-organized warehouses of books, maps, serials, etc. In their full expression, they are places where people meet to access, share, and exchange knowledge. The resources they select and services they offer should reflect the character of the communities they serve [31].

As suggested by Borgman [14-16], digital libraries should match and indeed dramatically extend traditional libraries. As such, they should be much more than search engine portals. Like any library they should feature a high degree of *selection* of resources that meet criteria relevant to their mission, and they should provide *services*, including search, that facilitate use of the resources by their target community. But, freed of the constraints of physical space and media, digital libraries can be more adaptive and reflective of the communities they serve. They should be *collaborative*, allowing users to contribute knowledge to the library, either actively through annotations, reviews, and the like, or passively through their patterns of resource use. In addition, they should be *contextual*, expressing the expanding web of inter-relationships and layers of knowledge that extend among selected primary resources. In this manner, the core of the digital library should be an evolving information base, weaving together professional selection and the "wisdom of crowds" [54].

This expanded view of a digital library requires new thinking about the information models on which they are based. The legacy of the union catalog in the traditional library and the sometimes myopic emphasis on search in the digital library has led to widespread use of an information model built around a metadata repository. Although many digital libraries are implemented differently, we often find that at their core they collect, index, and provide queries over a catalog of metadata records. As we describe

later, this simple catalog model breaks down in the face of a more expansive view of a digital library.

This paper describes an information model for digital libraries that intentionally moves "beyond search and access", without ignoring those basic functions, and facilitates the creation of collaborative and contextual knowledge environments. This model is an *information network overlay* that represents a digital library as a graph of typed nodes, corresponding to the information units (documents, data, services, agents) within the library, and semantic edges representing the contextual relationships among those units. The information model integrates local and distributed information integrated with web services, allowing the creation of rich documents (e.g., learning objects, publications for e-science, etc.). It expresses the complex relationships among information objects, agents, services, and meta-information (such as ontologies), and thereby represents information resources in context, rather than as the result of stand-alone web access. It facilitates collaborative activities, closing the loop between users as consumers and users as contributors.

We describe how this data model is implemented in the Fedora open source repository software [27]. Fedora is particularly well-suited for this task, due to its unique combination of a flexible object model, web service integration, fine granularity access management, and incorporation of semantic web expressivity.

The backdrop for this work is the National Science Digital Library (NSDL) project [61]. The demands of the NSDL, both in scale and requirement set, mandate such a sophisticated approach. Its original vision, articulated by Frank Wattenberg, stated:

> *"In many respects, NSDL could move well beyond the image of a library. In addition to providing timely and wide access to up-to-date, high quality resources for SMET education, NSDL could exploit the connectivity provided by the Internet and the potential of interactive technologies to create a rich, asynchronous workplace -- a seminar room, a reading room, and laboratory for sharing and building knowledge. Consequently, it could provide a framework for facilitating the work of people in different settings through a diverse and powerful set of resources."* [59]

We believe that this expanded vision of a digital library is not unique to the NSDL. Although all information seeking and sharing communities need to find needles in haystacks [28] – a niche met by Google and its competitors – they also need functionality "beyond search and access", where digital libraries "create a rich, asynchronous workplace" in which information is shared, aggregated, manipulated, and refined.

1 Building a digital library with a metadata repository: NSDL Phase I

Readers of *D-Lib Magazine* and the digital library community are probably quite

familiar with the NSDL project. Thus, this section provides only brief background for the work described in the remainder of the paper. Those seeking more information are encouraged to read earlier papers [6, 7, 25, 61] and scan through the "about NSDL" page at <http://nsdl.org/about>.

The notion of an NSDL developed from a 1998 NSF-sponsored workshop [3]. This workshop addressed concerns about the state of science, technology, engineering, and mathematics (STEM) education in the U.S. and articulated the opportunities to use the Internet and Web-based technologies to improve it. Building on the results of this workshop, the NSF began funding NSDL projects in 2000, and to date it has awarded over 180 grants. These awards cover a number of areas including collection development, services, and basic research. The specific work described in this paper is funded under grants to Cornell University, Columbia University, and the University Corporation for Atmospheric Research (UCAR) for "Core Integration" (CI), which includes coordinating the architectural, organizational, and policy infrastructure for the NSDL.

The initial technical work by the CI team led to an architecture and data model motivated by three basic functions: selecting web-based STEM resources, searching across them, and facilitating access to them. The architectural paradigm to do this is essentially a *union catalog*; a metadata repository (MR) of Dublin Core records corresponding to resources developed and curated by NSDL collection projects and other contributing organizations. The MR is implemented as an Oracle™ relational database, in which individual metadata records are stored in a series of tables.

Dublin Core metadata records, which include URLs to the corresponding digital resources, are ingested into the MR via OAI-PMH [29]. As part of the ingest process, these records are processed to normalize dates and various controlled vocabulary elements. Other services, including CI-managed search and resource archiving, use an OAI-PMH server[12] to harvest these normalized records and thereby obtain information they need (e.g., to build search indexes from metadata).

The search service uses the Lucene[13] full-text indexing system to index both the harvested metadata records describing the resource and the text content of the first HTML page referenced by the metadata record. The archive service uses the Storage Resource Broker [10] developed the San Diego Supercomputing Center. It does monthly web crawls of all the digital resources identified in metadata records harvested from the MR. The archiving service identifies a collection of linked pages that it considers most representative of the actual resource, and it creates a snapshot archive of these pages.

From the user perspective, resources in the NSDL catalog and underlying services are available through a central portal at <http://www.nsdl.org>. This central portal will soon be supplemented by portals for specific learning communities, funded through the NSDL Pathways program [2].

The NSDL central portal and underlying metadata repository-based architecture was first deployed in December 2003. Over the past two years the size of the collection has grown to over 1.1 million resources, with metadata records harvested from over eighty OAI-PMH data providers.

2 Utility of a metadata repository as a digital library architecture

In general, the large-scale use of Dublin Core and OAI-PMH in the NSDL MR has proven their utility for providing basic digital library services, but has also revealed a number of implementation problems. The most outstanding of these relate to metadata quality [6] and OAI-PMH validity, especially XML-schema compliance. As a result, the administrative costs of maintaining the MR have been unexpectedly high. These technical issues will be described in a future paper.

However, our focus in this paper is to examine the existing NSDL architecture, and digital libraries in general, from a broader perspective. This section reviews scholarship from the education community that examines requirements for education-oriented digital libraries and the functionality needed to meet those requirements.

Digital libraries are valuable for education because they offer access to and opportunities for use of online primary resources. But, in order to be truly effective as educational tools, they need to do more than provide access to quality resources. Reeves wrote "The real power of media and technology to improve education may only be realized when students actively use them as cognitive tools rather than simply perceive and interact with them as tutors or repositories of information." [49]. Marshall also noted that digital libraries need to be more than repositories and must support the full life cycle of data, information, and knowledge, and knowledge construction in general [36].

This broader functionality mandates an information model for digital libraries that is richer than a collection of simple web pages or static documents. Wiley [60] and others use the notion of *learning objects* to indicate a collection of information, which includes not only one or more primary resources, but also the dynamic *educational context* for the information. Context includes social and cultural information; patterns of use; pedagogical goals, the nature of learners' educational systems; and the learners' abilities, preferences and prior knowledge [37]. Information context can be quite complex, reflecting the breadth of audiences served by a digital library and the differences in how audiences use and manipulate information.

A number of researchers have examined the many facets of this contextual information. These include:

- Capturing opinions, comments, and reviews about library resources [39] and their history of use [43].
- Describing the community of users involved in the creation of a learning object [48].

- Capturing learner interactions and connecting their profiles to learning objects [38].
- Associating teacher recommendations and correlations to state education standards [47].
- Tracking and storing the search keywords that led to eventual use of the resource [4].

The basic record-oriented data and metadata model employed by most digital (and traditional) libraries has a limited ability to fully model this multi-dimensional information context.

First, metadata records, and metadata repositories, primarily represent individual item properties. They often fail to completely model contextual relationships [43] that surround resources and do not distinguish among the multiple entities – resources, metadata, agents, ontologies – that are part of that relationship structure. Furthermore, because they are frequently based on fixed schema or models, they are difficult to adapt to evolving information needs. The NSDL metadata repository, for example, only distinguishes between collections and items and represents only the membership relationship between them. Because the MR is stored in a relational database, each new relationship requires schema redefinition. This lack of flexibility has proven problematic due to the changing requirements over the span of the NSDL activity.

Second, the static nature of metadata records, which are generally created once by resource creators or catalogers, is also problematic. Resource context is dynamic, expressing changing patterns of use, preference, and the shifting cultural environment. Recker and Wiley write "a learning object is part of a complex web of social relations and values regarding learning and practice. We thus question whether such contextual and fluid notions can be represented and bundled up within one, unchanging metadata record" [48].

Third, an information model that is metadata-centric inevitably runs against the problematic fuzziness of the "data or metadata" distinction[14] [19]. For example, we have noted above that one of the useful forms of contextual information is annotations. Are these metadata (about something) or data in their own right? There is no one answer, but an architecture that imprints the distinction between data and metadata makes it difficult to deal with such ambiguities.

Finally, we have also noted the importance of information reuse – the ability to take primary resources and combine them into aggregate learning objects or lesson plans [46] and recursively reuse and re-factor new objects. Because the physically-bound information units in the traditional library were not amenable to such reuse, a metadata – centric approach – managing descriptive records – was possible. However, a digital library needs to be *resource-centric*, providing the framework for managing, manipulating, and processing content and metadata and the seamless line between them.

3 Information modeling for complexity and context

What is the proper information model that overcomes the limitations of the metadata-centric approach? In searching for an answer to this question, we should be wary of throwing out collection of cataloging records, and ignoring the value that uniform metadata has for "order making" over heterogeneous information [30]. However, we need to incorporate these catalog records into a richer foundation that represents structured and unstructured descriptions, heterogeneity and homogeneity, metadata and content, static and dynamic information, complex relationships, and a host of other complexities.

This section describes the framework of a richer information model that accommodates complexity and context. We develop this model by describing the "item problem"[15], building from the basics of providing search and access to homogeneous items (resources) in a digital library and progressively adding complexity. We argue that, although the context of this description is the NSDL, the problem described here generalizes over a variety of digital libraries and information environments.

3.1 Representing stuff[16]

Figure 1.

As noted earlier, the initial goal of the NSDL was to provide selection, search, and access over URL-accessible STEM resources. This limited goal was met by the well-known union catalog model, where the library is represented as a set of uniform (Dublin Core) metadata records that reference resources via their URLs. Note that in this model the representation of resources is second class – they are not represented themselves but only exist indirectly via references (URLs) from metadata.

3.2 Describing stuff in multiple structured and unstructured ways

Figure 2.

While Dublin Core provides minimal descriptive interoperability, it should co-exist with other richer domain-specific and purpose-specific formats [24]. In addition, as noted in the previous section, unstructured descriptions such as comments and

annotations are often as useful as structured metadata records. Furthermore, these descriptions, both structured and unstructured, originate from multiple contributors. This additional complexity stresses the foundations of the union catalog model, that is based on one set of providers (the library catalogers) creating and managing a uniform set of descriptions. Two new modeling issues arise. First resources need to be modeled alongside descriptive records, since resources provide the anchoring point for linking together multiple descriptions[17]. Second, modeling of agents and providers becomes important in order to represent branding of resources (who selected or created the resource?) and its distinction from branding of metadata (who provided the metadata?). Branding is a useful tool that helps users discern the quality of digital resources.

3.3 Adding more types of stuff

Figure 3.

As noted above, the model already needs to represent multiple types of descriptions, the agents that provided them, and the resources they describe. But, resources themselves are not homogeneous. Digital libraries collect an expanding variety of information resources – images, audio, and text, and extending to many more complex resource types such as data, simulations, multi-media learning objects and the like. This raises additional modeling complexity – in particular how best to accommodate uniformity at the user interface level while simultaneously representing the special characteristics of each type of resource. In addition to description (metadata) issues, there are access and presentation issues, since different information types may require different access protocols and helper applications, all of which must be represented in the information model.

3.4 Its not always clear what stuff is

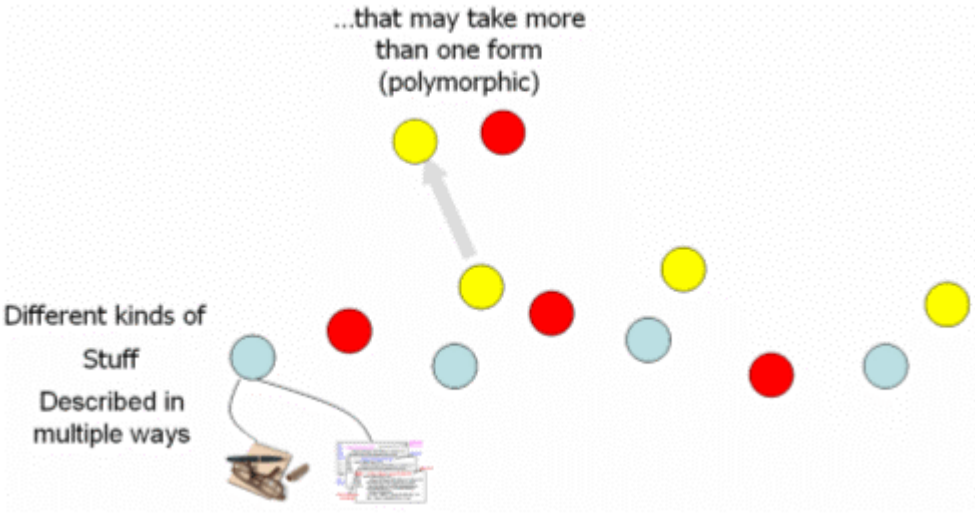

Figure 4.

Resources in a digital library are not always easy to characterize. For example, is an "e-book" a book or software [33]? Is information always either metadata or data? Can't an agent sometimes be an information resource? These are just a few examples of type complexities when modeling information. Rather than forcing the entities in the digital library into disjoint slots, the type structure of the information model must be polymorphic. Any entity should be able to adopt different characteristics and behaviors, depending on the context of access or use.

3.5 Allowing users to customize stuff

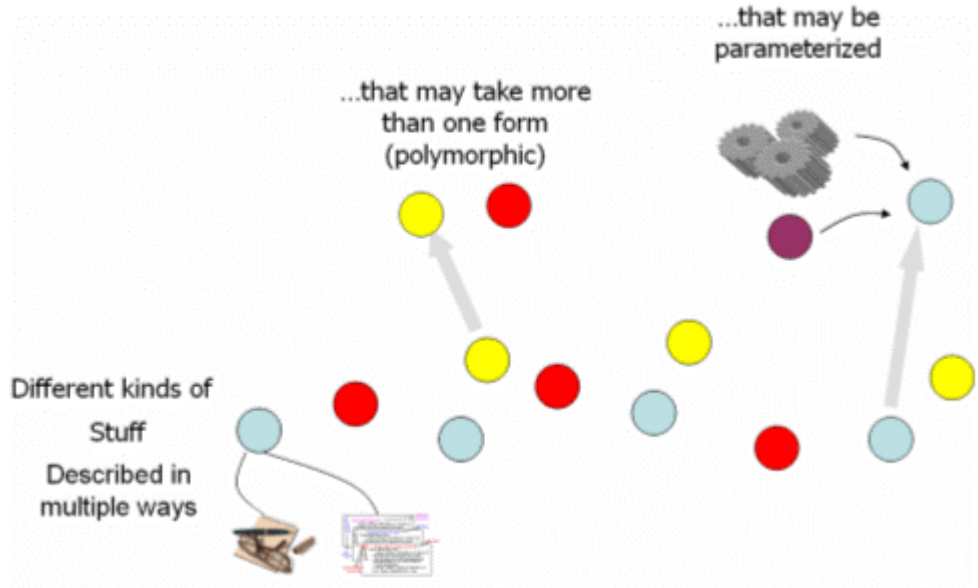

Figure 5.

Early digital library work formulated the notion of digital objects, which are packages of information with multiple disseminations available through service requests [22, 26]. Most modern digital library systems implement this functionality using standards such as complex object containers [12, 32] that encapsulate the metadata and data streams associated with a digital object. A service request may then include a parameter that specifies the nature of the dissemination request – for example, a request for a PDF or a LaTex dissemination of a scholarly document.

In a service-oriented architecture, these disseminations may be produced computationally as well as statically. For example, rather than storing an image in multiple formats and resolutions, it is possible to respond to a user request (e.g., 300dpi jpeg) using a single archival form (TIFF) that is processed by an image manipulation web service. This functionality is particularly attractive in an educationally focused digital library where customizing content according to diverse user needs (e.g., language) is desirable.

The information model must therefore model services alongside text, data, images, and other information and must characterize the interactions of those services with the other information units in the library.

3.6 Expressing the relationships among stuff

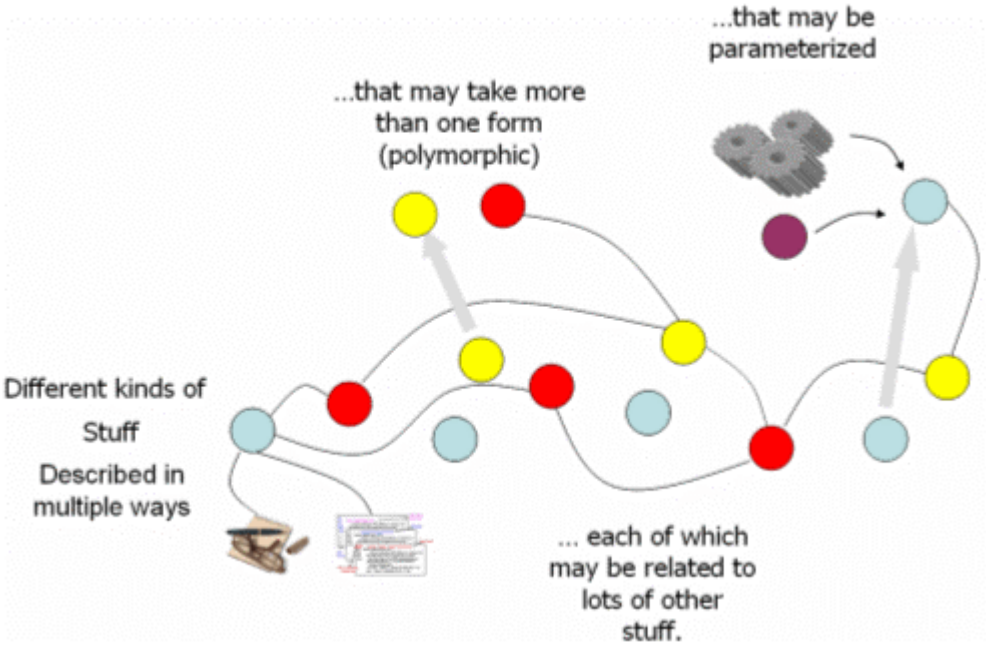

Figure 6.

With the goal of providing better support for the *collocation* objective [55], the library community has experimented with a number of information models for modeling bibliographic relationships. One example is the Functional Requirements for Bibliographic Records (FRBR) [1]. Our experience in the NSDL has shown that

bibliographic relationships are just one part of the problem. Other relationships include those between metadata and providers, resources and curators, resources to subject taxonomies, learning objects to state standards, and others. As the library expands, we anticipate needing to model other community-specific relationships. The information model must represent this graph of interconnected information nodes and the ontologies that provide meta-information for the relationships. Furthermore, these relationships need to evolve and not be rigidly constrained by static schema.

4 The Information network overlay

The previous section suggests an information model that is graph-based, with semantically linked edges and nodes that are flexibly typed and web service aware. We use the notion of an *information network overlay* (INO) to represent this model.

The INO borrows from two pre-existing bodies of work. Overlay networks have been employed in a number of applications to represent a set of edges or connections projected on top of a set of nodes that exist in some other network context – e.g., the Internet. Two particularly well-known application areas are network routing [5] and searching in P2P networks [18]. Semantic graphs, expressing the relationships among web resources, are fundamental to the semantic web [13] and have been employed in educationally-oriented applications such as Edutella [41]. In fact, our application of information network overlays employs semantic web technologies, integrated into Fedora.

The concepts underlying the INO are illustrated in the following figure, with the following layers:

- The *primary resources* or raw data selected for the library are shown at the bottom layer of the illustration. In the NSDL these are the web-accessible STEM resources. But, as noted earlier, these raw materials also consist of data sets, the agents and organizations that contribute to the library, and services.
- The *information network overlay*, shown at the next level up in the illustration, is locus for modeling library resources, their descriptions, and the web of information that builds around them. It is first populated with the primary resources or references via metadata to them, which are shown as red nodes. The association and derivation of these nodes with the primary resource layer is shown by the solid red arrows. The dashed red arrows in the INO indicate initial relations between these nodes, such as the collection/item relations in the NSDL. In the NSDL, the initial populating of the INO is done via metadata harvest from collection providers, essentially duplicating the functionality of the phase I metadata repository.
- The *Access-Controlled API*, shown as the next level up in the illustration, provides full programmatic access to the INO. This includes read and write access to the components of the data model – data, documents, metadata, agents, relationships, etc. – and searching over the relationships – e.g., "find all resources contributed by DLESE"[18].

- The API can then be used by external contributors – e.g., users, services, ontology classification services, and the like – to enhance the information in the INO. These API-channeled requests, indicated by solid green arrows, add both new nodes (such as learning objects that aggregate existing resources) to the INO, indicated in green, and new relationships among the nodes, indicated by green dashed lines.

This bi-directional flow, the representation of primary resources from the underlying raw data layer and contextual information from the upper layer, allows the INO to evolve over time into an increasingly rich information space. In the same way that amazon.com is an information source that extends far beyond a product catalog, we expect that digital libraries built on the INO model will reflect expanding communities of knowledge built over the resources in the library.

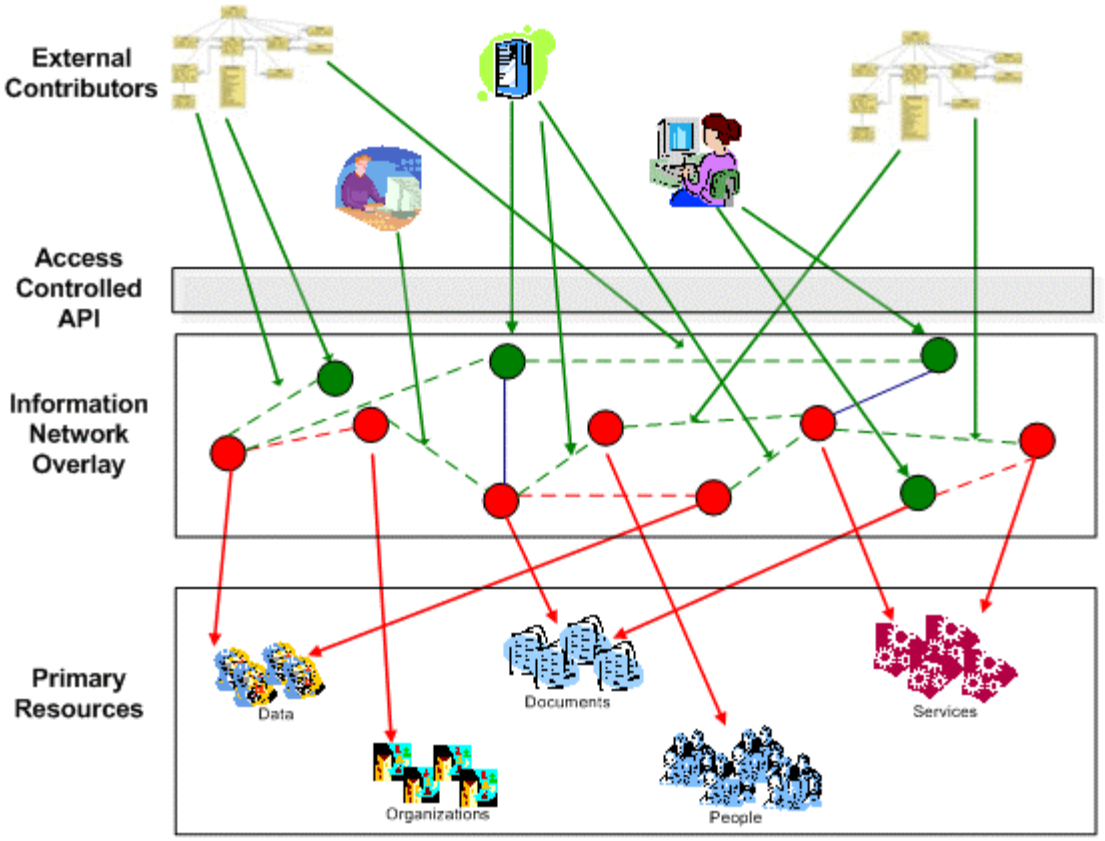

Figure 7.

Our platform for implementing the INO is the Fedora open source repository software[19]. Fedora has been deployed in a variety of applications including institutional repositories, archives, museums, commercial library projects. The rich object model underlying Fedora and exposure of the model through a web service interface makes Fedora an ideal framework for implementing the INO.

Each node in the INO corresponds to a Fedora digital object. The Fedora digital object model offers considerable functionality, combining traditional content management, service-oriented architectures, and semantic web technologies. The model allows the aggregation of local and remote data in multiple formats. Web-accessible services may then be associated with the data aggregated in a digital object. As a result a digital object is accessible in multiple representations, some of them direct transcriptions of aggregated data, and some of them produced dynamically by the associated web services. In the NSDL context, this provides the technical foundation for reuse and construction of complex learning objects [46], mixing primary resources with teacher commentary that can then be presented dynamically in multiple formats (e.g., as PowerPoint slide shows or Flash presentations).

Each edge in the INO corresponds to a semantic relationship expressed within the Fedora digital object model. Examples of relationships between digital objects in the INO include well-known management relationships such as the organization of items in a collection, structural relationships such as the part-whole links between individual chapters and a book, and semantic relationships useful in educational digital library organization such as subject, grade, and curricula appropriateness. Fedora defines a base relationship ontology using RDFS [17] and provides a slot in the digital object for RDF expression of relationships based on this ontology. Assertions from other ontologies may also be included along with the base Fedora relationships. All relationships expressed in digital objects are mapped into Kowari [57], a native RDF triple-store. The RDQL [52] and ITQL [56] query interface to this triple-store is exposed as a web service. Like any web service, this triple store query service may be associated with a digital object, allowing disseminations from digital objects that are parameterized by their semantic context.

5 The NSDL data repository – NSDL Phase II

To distinguish our work from the metadata repository (MR) in the first phase of NSDL, we call our INO implementation the *NSDL data repository* (NDR). The complete technical details of the data model implemented in the NDR are out of scope for this publication. The following three examples of functionality illustrate some of the features of the model. The complete NDR consists of multiple instances of these data model elements combined with other elements. For example, the multi-sourced metadata template, described in Section 5.1, is repeated for all 1.1 million resources in the NSDL collection.

The NDR is currently implemented as a single Fedora repository maintained by the CI team. In the future, we expect to implement the NDR as a set of federated Fedora repositories.

Each example that follows has an accompanying figure, where the circles represent nodes in the information network, which are implemented as Fedora digital objects. Each circle is color coded to correspond to the information type that it represents in the context of the modeling example. The lines represent semantically loaded relationships

between the information units. Like all relationships in Fedora, these relationships are stored within the digital object and then indexed in the Kowari triple store. These relationships are then searchable through graph queries.

5.1 Multi-sourced, multi-format metadata with branding

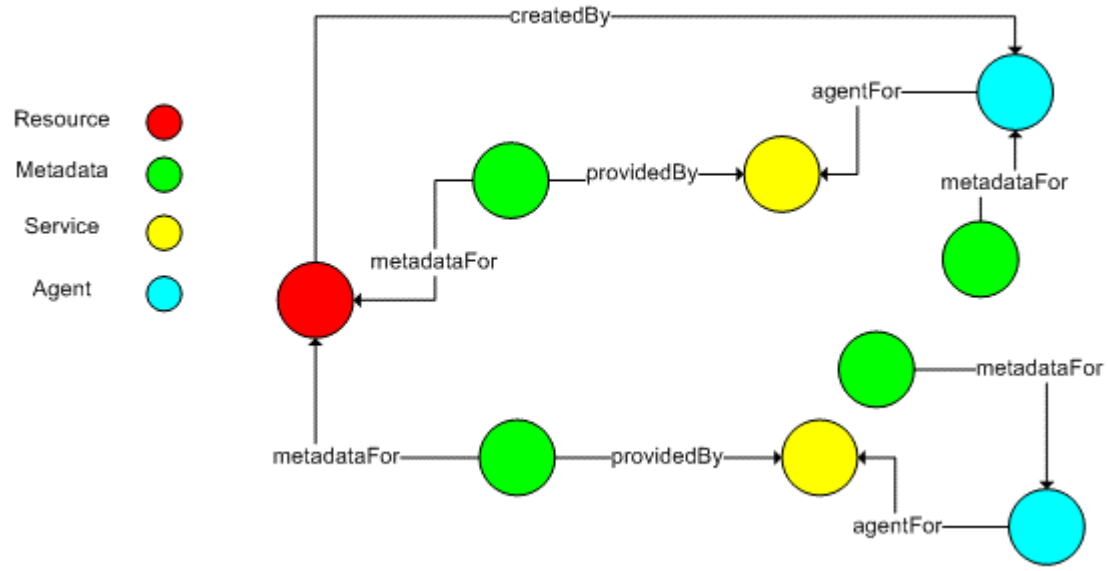

Figure 8.

The illustration above shows the NDR model that associates multi-sourced, multi-format metadata with a resource. Each metadata digital object in the illustration aggregates multiple formats originating from a single metadata provider. Building on Fedora's dynamic dissemination capability, some of these formats are computationally generated from a base format. The linkage of the metadata digital object to its provider, and the resource to its creator/selector provides branding information. Branding is important in any library where data and metadata originates from multiple sources. This branding, linked with a reputation system, can provide information for determining the quality of resources and their descriptions.

5.2 Unstructured Annotations and Reviews

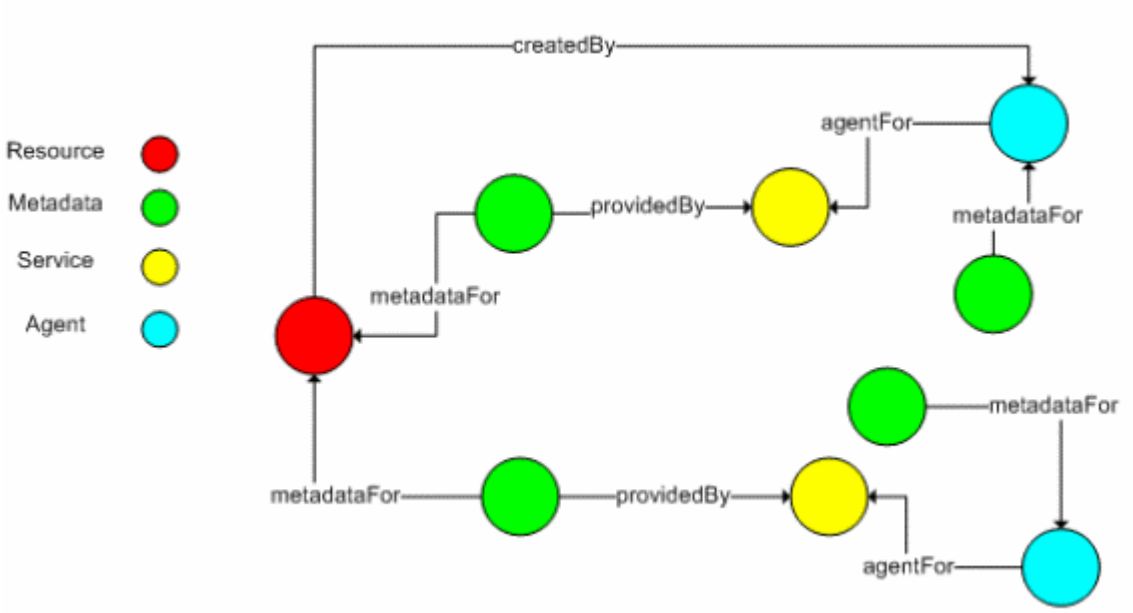

Figure 9.

Although structured metadata is useful for a variety of purposes, we have noted the utility of unstructured annotations and reviews. The model represents these annotations and reviews as resources in their own right – their status as an annotation is due to their association with an "annotationFor" relationship to a target resource. This is one example of polymorphism in the INO, whereby a node may assume multiple characteristics. Another example is a resource that is also an "agent". Another is a "collection" that is also an "item" that can be aggregated in other collections. The Fedora digital object allows this without constraints of single-inheritance object-oriented architectures. Essentially, a digital object may assume any combination of type identities.

5.3 Collections and Aggregations

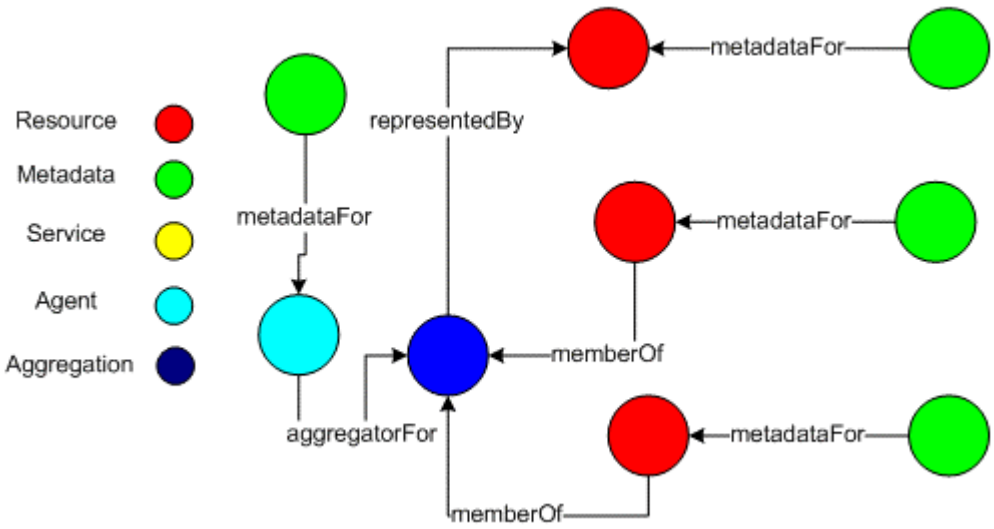

Figure 10.

The Phase I NSDL architecture only allowed one form of aggregation, which expressed the relationships between a metadata provider and the set of metadata harvested from the provider. The NDR implementation includes a general set-based aggregation model that allows any number of resources to be included in any number of aggregations. As indicated in the illustration, an aggregation is "representedBy" a resource. This resource supplies the semantics for the aggregation. For example, an aggregation may indicate a set of resources that correlate to some state educational standard. In this case the resource at the tail of the "representedBy" link expresses that standard. Aggregations are themselves "resources", which may be nested in additional aggregations. These semantically meaningful aggregations in the data model are the foundation for rich contextualization of resources in the library.

5.4 NDR Status

As of the writing of this paper, the initial load of the NDR using data from the pre-existing metadata repository is almost complete (over 1.1 million records). The resulting INO graph has about 1.5 million nodes, and the graph has about 10 million explicit edges (with a number of other implicit edges that are Fedora specific). In a future paper, we will report on our experiences with RDF triple stores in the process of scaling an INO up to this level. In particular, we have noted a number of performance characteristics of triple stores that apply to other semantic web-based applications.

With completion of the data load, we will release the specification of the NDR API to the NSDL community. This will begin the process of "fleshing out" the INO with additional contextual information added by the NSDL community. We anticipate some interesting results that characterize the nature of the INO as it develops over time.

6 Conclusion

In the age of Google, what is a digital library anymore, anyway? Just asking the question is bound to raise passions. Despite our zealous defense of the successful work of the digital library community over the past decade, the amazing success of commercial web search engines has changed the playing field. Search and access over a set of resources, while important to any digital library, are not sufficient. Digital libraries need to distinguish themselves from web search engines in the manner that they add value to web resources. This added value consists of establishing context around those resources, enriching them with new information and relationships that express the usage patterns and knowledge of the library community. The digital library then becomes a context for information collaboration and accumulation – much more than just a place to find information and access it.

Our work in the NSDL has demonstrated that the familiar metadata-based model is not sufficient for this type of functionality. We have designed and implemented an information network overlay within Fedora, which includes the full functionality of the existing metadata repository, but models relationships, services, and multiple information types within a web-service based application. This rich information store will provide the basis for the next stage of work, implementing an expanding suite of user-visible library services that fulfill the "laboratory for sharing and building knowledge" envisioned in the original NSDL report [59].

Acknowledgements

This paper incorporates work from many people in addition to the authors. The Fedora group, especially Chris Wilper and Eddie Shin, deserves credit for their hard work implementing these ideas within the Fedora open source software. Members of the NSDL group, especially Tim Cornwell, Elly Cramer, and Naomi Dushay, played major roles in the formulation of the NSDL data model and its implementation within the NDR. The NSDL project as a whole owes acknowledgement to Lee Zia, who has championed the project at NSF over many years. The work described here is based upon work supported by several grants. NSDL NDR work is supported by the National Science Foundation under Grants No. 0227648, 0227656, and 0227888. Work on information network overlays is supported by the National Science Foundation under Grant No. 0430906. Work on Fedora is supported by the Andrew W. Mellon Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Andrew W. Mellon Foundation. Special thanks to Lucy Lagoze who showed Carl Lagoze how hard it is for a student to use web search engines and provided some lessons on the importance of context and patterns of use.

Thanks, also, to Mike Keller and Vicky Reich for permission to adopt and adapt a title they used in an earlier, informative paper [23].

References

[1] "Functional Requirements for Bibliographic Records," International Federation of Library Associations and Institutions March 1998. <http://www.ifla.org/VII/s13/frbr/frbr.pdf>.

[2] *New Pathways to the National Science Digital Library*, 2004 <http://www.infosci.cornell.edu/news/NSDL_Pathways.pdf>.

[3] "Report of the Science, Mathematics, Engineering, and Technology Education Library Workshop," National Science Foundation, Washington, DC, Workshop Report July 21-23 1998. <http://www.dlib.org/smete/public/report.html>.

[4] J. Abbas, C. Norris, and E. Soloway, "Middle School Children's Use of the ARTEMIS Digital Library," presented at ACM/IEEE Joint Conference on Digital Libraries (JCDL '02), Portland, OR, 2002.

[5] D. G. Andersen, H. Balakrishnan, and M. F. Kaashoek, "Resilient Overlay Networks," presented at 18th ACM SOSP, Banff, Canada, 2001.

[6] W. Y. Arms, N. Dushay, D. W. Fulker, and C. Lagoze, "A Case Study in Metadata Harvesting: the NSDL," Library Hi Tech, 21 (2), 2003.

[7] W. Y. Arms, D. Hillmann, C. Lagoze, D. Krafft, R. Marisa, J. Saylor, C. Terrizzi, and H. Van de Sompel, "A Spectrum of Interoperability: The Site for Science Prototype for the NSDL," *D-Lib Magazine*, 8 (1), 2002. <doi:10.1045/january2002-arms>.

[8] D. E. Atkins, *Report of the Santa Fe Planning Workshop on Distributed Knowledge Work Environments: Digital Libraries*, 1997 <http://www.si.umich.edu/SantaFe/report.html>.

[9] K. Auletta, "Under the Wire," *New Yorker*, January 17, 1994.

[10] C. Baru, R. Moore, A. Rajasekar, and M. Wan, "The SDSC Storage Resource Broker," presented at CASCON'98, Toronto, 1998.

[11] D. Bearman, G. Rust, S. Weibel, E. Miller, and J. Trant, "A Common Model to Support Interoperable Metadata. Progress report on reconciling metadata requirements from the Dublin Core and INDECS/DOI Communities," *D-Lib Magazine*, 5 (January), 1999. <doi:10.1045/january99-bearman>.

[12] J. Bekaert, P. Hochstenbach, and H. Van de Sompel, "Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library," *D-Lib Magazine*, 9 (11), 2003. <doi:10.1045/november2003-bekaert>.

[13] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific*

American, (50), May, 2001.

[14] C. L. Borgman, "Digital libraries and the continuum of scholarly communication," *Journal of Documentation*, 56 (4), pp. 412-430, 2000.

[15] C. L. Borgman, "The invisible library: Paradox of the global information infrastructure," *Library Trends*, 51 (4), pp. 652, 2003.

[16] C. L. Borgman, "What are digital libraries? Competing visions," *Information Processing & Management*, 1999 (35), pp. 227-243, 1999.

[17] D. Brickley and R. V. Guha, "RDF Vocabulary Description Language 1.0: RDF Schema," W3C, Recommendation February 10 2004. <http://www.w3.org/TR/rdf-schema/>.

[18] A. Crespo and H. Garcia-Molina, "Semantic overlay networks for p2p systems," Stanford University, Palo Alto 2003.

[19] R. Daniel Jr. and C. Lagoze, "Extending the Warwick Framework: From Metadata Containers to Active Digital Objects," *D-Lib Magazine* (November), 1997. <doi:10.1045/november97-daniel>.

[20] E. Fox, R. M. Akscyn, R. K. Furuta, and J. J. Leggett, "Digital libraries," *Communications of the ACM*, 38 (4), pp. 22-28, 1995.

[21] F. Fukuyama, *The end of history and the last man*. New York, Toronto: Free Press, 1992.

[22] R. Kahn and R. Wilensky, "A Framework for Distributed Digital Object Services," Corporation for National Research Initiatives, Reston, Working Paper cnri.dlib/tn95-01, 1995. <http://www.cnri.reston.va.us/k-w.html>.

[23] M. A. Keller, V. Reich, and A. C. Herkovic, "What is a library anymore, anyway?," *First Monday*, 8, May 5, 2003.

[24] C. Lagoze, "The Warwick Framework: A Container Architecture for Diverse Sets of Metadata," *D-Lib Magazine*, 2 (7/8), 1996. <doi:10.1045/july96-weibel>.

[25] C. Lagoze, W. Arms, S. Gan, D. Hillmann, C. Ingram, D. Krafft, R. Marisa, J. Phipps, J. Saylor, C. Terrizzi, W. Hoehn, D. Millman, J. Allan, S. Guzman-Lara, and T. Kalt, "Core Services in the Architecture of the National Digital Library for Science Education (NSDL)," presented at Joint Conference on Digital Libraries, Portland, Oregon, 2002.

[26] C. Lagoze and J. R. Davis, "Dienst - An Architecture for Distributed Document

Libraries," *Communications of the ACM*, 38 (4), pp. 47, 1995.

[27] C. Lagoze, S. Payette, E. Shin, and C. Wilper, *Fedora: An Architecture for Complex Objects and their Relationships*, 2005 <http://arxiv.org/abs/cs.DL/0501012>.

[28] C. Lagoze and A. Singhal, "Information Discovery: Needles and Haystacks," *IEEE Internet Computing*, 2005 (May/June), 2005.

[29] C. Lagoze, H. Van de Sompel, M. Nelson, and S. Warner, *The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0*, 2002 <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

[30] D. Levy, "Cataloging in the Digital Order," presented at The Second Annual Conference on the Theory and Practice of Digital Libraries, 1995.

[31] D. Levy, "Digital Libraries and the Problem of Purpose," *Bulletin of the American Society for Information Science*, 26 (6), 2000.

[32] Library of Congress, *METS: An Overview & Tutorial*, 2004 <http://www.loc.gov/standards/mets/METSOverview.v2.html>.

[33] C. Lynch, "The Battle to Define the Future of the Book in the Digital World," *First Monday*, 6 (6), June 4, 2001.

[34] C. A. Lynch and H. Garcia-Molina, "Interoperability, Scaling, and the Digital Libraries Research Agenda," IITA Digital Libraries Workshop May 18-19 1995. <http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html>.

[35] C. A. Lynch and M. A. Keller, *googlization, digital repositories, distance education, and privacy*, 2005 <http://www.learningtimes.net/acrlarchive.html.

[36] B. Marshall, Y. Zhang, H. Chen, A. Lally, R. Shen, E. A. Fox, and L. Cassel, "Convergence of Knowledge Management and E-Learning: the GetSmart Experience," presented at ACM/IEEE Joint Conference on Digital Libraries (JCDL '03), Houston, TX, 2003.

[37] K. Martin, "Learning in Context," *Issues of Teaching and Learning*, 4 (8), September, 1998.

[38] G. McCalla, "The Ecological Approach to the Design of E-Learning Environments: Purpose-based Capture and Use of the Information about Learners," *Journal of Interactive Media in Education*, 7 (Special Issue on the Educational Semantic Web), 2004.

[39] F. McMartin and Y. Terada, "Digital Library Services for Authors of Learning Materials," presented at ACM/IEEE Joint Conference on Digital Libraries (JCDL '02),

Portland, OR, 2002.

[40] National information Standards Organization (U.S.), *The OpenURL Framework for Context-Sensitive Services*, 2003 <http://www.niso.org/standards/resources/Z39_88_2004.pdf>.

[41] W. Nejdl, B. Wolf, and C. Qu, "EDUTELLA: A P2P Networking Infrastructure Based on RDF," presented at WWW2002, Honolulu, 2002.

[42] A. M. Ouksel and A. Sheth, "Semantic Interoperability in Global Information Systems," *SIGMOD Record*, 28 (1), 1999.

[43] P. Parrish, "The Trouble with Learning Objects," *Educational Technology Research and Development*, 52 (1), pp. 49-67, 2004.

[44] S. Payette and C. Lagoze, "Flexible and Extensible Digital Object and Repository Architecture (FEDORA)," presented at Second European Conference on Research and Advanced Technology for Digital Libraries, Heraklion, Crete, 1998.

[45] President's Information Technology Advisory Committee: Panel on Digital Libraries, "Digital Libraries: Universal Access to Human Knowledge," PITAC February 2001. <http://www.itrd.gov/pubs/pitac/pitac-dl-9feb01.pdf>.

[46] M. Recker, *Instructional Architect*, 2004 <http://ia.usu.edu/>.

[47] M. Recker, J. Dorward, and L. M. Nelson, "Discovery and Use of Online Learning Resources: Case Study Findings," *Educational Technology and Society*, 7 (2), pp. 93-104, 2004.

[48] M. Recker and A. Walker, "Collaboratively filtering learning objects," in *Designing Instruction with Learning Objects*, D. A. Wiley, Ed., 2000.

[49] T. C. Reeves, *The Impact of Media and Technology in Schools: A Research Report prepared for The Bertelsmann Foundation*, 1998. <http://it.coe.uga.edu/~treeves/edit6900/BertelsmannReeves98.pdf>.

[50] V. Reich, "LOCKSS: A Permanent Web Publishing and Access System," *D-Lib Magazine*, 7 (6), 2001. <doi:10.1045/june2001-reich>.

[51] G. Salton, *Dynamic information and library processing*. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

[52] A. Seaborne, *RDQL – A Query Language for RDF*, 2004. <http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/>.

[53] M. Smith, M. Bass, G. McClellan, R. Tansley, M. Barton, M. Branschofsky, D.

Stuve, and J. H. Walker, "DSpace: An Open Source Dynamic Digital Repository," *D-Lib Magazine*, 9 (1), 2003. <doi:10.1045/january2003-smith>.

[54] J. Surowiecki, *The wisdom of crowds : why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*, 1st ed. New York: Doubleday :, 2004.

[55] E. Svenonius, *The intellectual foundation of information organization*. Cambridge, Mass.: MIT Press, 2000.

[56] Tucana Technologies, *iTQL Commands*, 2004 <http://kowari.org/oldsite/271.htm>.

[57] Tucana Technologies, *Kowari metastore*, 2004 http://www.kowari.org/>.

[58] J. Ward, "A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative," presented at Joint Conference on Digital Libraries, Houston, 2003.

[59] F. Wattenberg, "A National Digital Libraries for Science, Mathematics, Engineering, and Technology Education," *D-Lib Magazine*, 1998 (October), 1998. <doi:10.1045/october98-wattenberg>.

[60] D. A. Wiley, "Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy," in *The Instructional Use of Learning Objects: Online Version*, D. A. Wiley, Ed., 2000.

[61] L. L. Zia, "The NSF National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program," *D-Lib Magazine*, 8 (11), 2002. <doi:10.1045/november2002-zia>.

Notes

1. Although searchable catalogs of digital documents were introduced early in the history of computing [51]. general use of the term "digital library" began in the early 1990's [16] [20].

2. <http://www.google.com>.

3. <http://www.handle.net>.

4. <http://www.dublincore.org>.

5. <http://openarchives.org>.

6. <http://arxiv.org>.

7. The completeness of which will be substantially enhanced by massive scanning efforts such as Google Print <(http://print.google.com>.

8. <http://www.doi.org>.

9. <http://shibboleth.internet2.edu/>.

10. A search reveals that there are over 13,000 appearances of this term on the web, including a webcast by Clifford Lynch and Michael Keller on the subject [35]

11. This is a reference to Francis Fukuyama's 1992 book [21] reflecting a similar myopic euphoria about political economy.

12. <http://services.nsdl.org:8080/nsdloai/OAI>.

13. <http://jakarta.apache.org/lucene/docs/index.html>.

14. We note that this problematic distinction was one of the initial motivations [44] for the Fedora architecture, which is used to implement the model described later in this paper.

15. Thanks to a former colleague who is now at amazon (and who will remain anonymous) for this phrase. The correspondence between information modeling problems at amazon and digital libraries is not coincidental. Amazon is perhaps the premier example of an information environment that provides users with contextualized and rich information built over a basic data layer (its products).

16. The use of this term is borrowed from Godfrey Rust [11]. "People Make Stuff, People Use Stuff, and People Do Deals About Stuff"

17. Collecting metadata from multiple metadata providers raises interesting equivalence problems. The ability to determine that two descriptions are about the same resource is based on heuristics and subjectivity.

18. <http://dlese.org>.

19. <http://www.fedora.info>.

(The title of the article was corrected on 11/16/05.)