Creating Virtual Collections in Digital Libraries: Benefits and Implementation Issues

Gary Geisler
Interaction Design Laboratory
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360, USA
+1 919 489 2759
geisg@ils.unc.edu

Sarah Giersch, David McArthur, Marty McClelland
CollegisEduprise, Inc.
2000 Perimeter Park Drive
Morrisville, NC 27560, USA
+1 919 376 3424
{sgiersch, dmcarthur, mmcclelland}@eduprise.com

ABSTRACT
Digital libraries have the potential to not only duplicate many of the services provided by traditional libraries but to extend them. Basic finding aids such as search and browse are common in most of today's digital libraries. But just as a traditional library provides more than a card catalog and browseable shelves of books, an effective digital library should offer a wider range of services. Using the traditional library concept of special collections as a model, in this paper we propose that explicitly defining sub-collections in the digital library—virtual collections—can benefit both the library's users and contributors and increase its viability. We first introduce the concept of a virtual collection, outline the costs and benefits for defining such collections, and describe an implementation of collection-level metadata to create virtual collections for two different digital libraries. We conclude by discussing the implications of virtual collections for enhancing interoperability and sharing across digital libraries, such as those that are part of the National SMETE Digital Library.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, standards, system issues, user issues*

General Terms
Design, Standardization

Keywords
Digital library, collection, metadata, user services

1. INTRODUCTION
Most of the digital library research and development to date has centered on issues related to the technology and content of digital libraries [10]. This work has focused on issues such as developing effective ways to digitize and store resources, how to efficiently deliver resources over the network, providing ways to search for resources, and how to enable digital libraries to interoperate.

These are fundamental issues to be sure, but to be viable in the long run a digital library must be more than a collection of digital objects that can be efficiently stored and transported. Just as the traditional library evolved to provide services to make its contents more accessible to its users, the effective digital library must develop a range of services to assist its users in finding, understanding, and using its contents. Moreover, in its digital form the library has the potential to not just emulate traditional libraries in the services it provides but to improve and extend them by capitalizing on advantages inherent in the medium.

One important area where the digital library can extend the services it provides beyond that of the traditional library is in integrating and highlighting user contributions. With the exception of especially unique or noteworthy contributions, the traditional library is rarely eager to receive resource contributions outside of its usual channels, as the effort needed to catalog and integrate outside contributions into a physical library is substantial. Digital libraries, on the other hand, are more often willing to receive contributions. It has been demonstrated that a combination of minimal submission data and basic verification procedures can result in high-quality digital library contributions with low rejection rates [7]. Such contributions enhance the value of the digital library by increasing its size and diversity and the process of cataloging and integrating contributed resources into a digital library often requires less effort.

However, the aspects that make digital libraries built from user contributions valuable—diversity of content, potential for large growth—also create potential drawbacks. For example, search and browse facilities enable users to find resources based on features such as author, subject, and keywords, but as a digital library grows, finding specific resources of interest among the entire collection can become more difficult. At the same time, the prominence of a given contributor's contributions becomes diminished as the library grows.

One way to help users find resources of interest in a digital library while ensuring that contributors receive recognition is to borrow a concept that has long been part of traditional libraries: the special collection. By defining and making available *virtual collections* we believe the digital library can extend the special collection model and—at a modest cost—provide benefits to both its users and contributors that will encourage its own growth and viability.

JCDL'02, July 13-17, 2002, Portland, Oregon, USA.

2. WHAT IS A VIRTUAL COLLECTION?

Traditional libraries often contain, in addition to their main holdings, special collections. In these settings a special collection is generally defined as a group of related materials that is given some form of special treatment. The special treatment might be due to the rare or delicate nature of the materials (rare books or antique maps, for example), or because the library wants to highlight the materials in some way (the collected papers of a noteworthy poet).

In contrast to traditional libraries, the special or sub-collections of digital libraries can be much more fluid. Where the holdings of a traditional library are physically constrained to a single space and a single ordering, resources in a digital library can be distributed across many servers, can be owned by different organizations, and can be displayed in many different orderings and arrangements. This fluid nature makes defining collections and sub-collections in the digital environment less straightforward. Adhering closely to the traditional library definition, one could consider a digital library collection to be all those resources that reside on a single server, or alternatively, all those that can be accessed though a given library's interface, even if they are physically distributed.

As suggested in [4], however, even a broad definition of a collection in the context of digital libraries can be ambiguous. It can, for example, be influenced by the point of view of those making the definition. The people responsible for managing resources stored in separate databases might think of each as a distinct collection, while an end user with access to them all is more likely to consider them a single collection.

Defining sub-collections can be even more flexible as there are many possible factors that can suggest how sub-collections can be formed. A sub-collection can be defined by including all those resources that share a topic or other significant attribute (the collection of all butterfly images), those contributed by a specific organization (the collection of Insect Museum resources), or those used for a specific purpose (all resources used for the online course in butterflies).

These sub-collection examples are instances of collections that cannot be easily replicated in traditional libraries. They are made possible by exploiting advantages the digital environment inherently provides: objects can exist in multiple collections, collections with the same objects grouped in different ways can co-exist, collections can be created dynamically and exist for varying amounts of time. They become virtual collections and as such—in contrast to the traditional library—enable a digital library to provide a limitless number of sub-collections based on a wide range of features.

3. BENEFITS OF CREATING VIRTUAL COLLECTIONS

Although it is common for traditional libraries to create and maintain special collections, many digital libraries do not attempt to provide a similar service. Most digital libraries do create the most basic of virtual collections—the result set dynamically created from a search request or category browsing—but rarely do they explicitly create and promote the sort of virtual collections described above. Doing so, however, can benefit both the users of the digital library and those who contribute resources to it.

3.1 Benefits to Users

A digital library that contains virtual collections helps its users in several ways. A new user who may be intimidated by a digital library's search interface or the number of results returned by a query might be better introduced to the digital library through the more easily exploreable partitioned set of resources in a virtual collection. A directory of the virtual collections contained by a digital library, as shown in Figure 1, can provide a good introduction and overview of the library's contents to new or casual users.

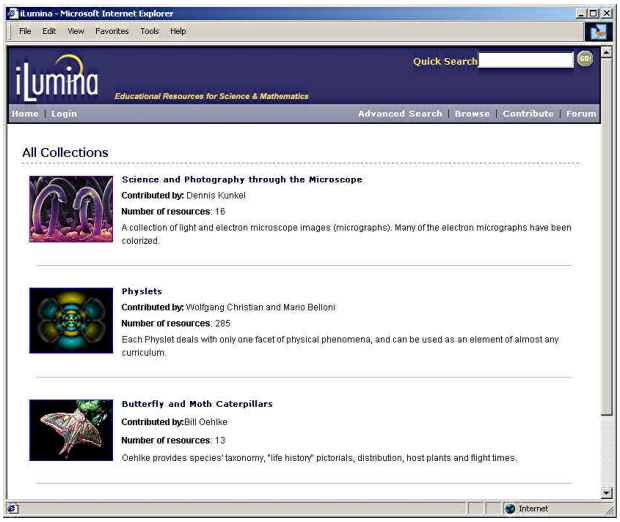

Figure 1. Virtual collections available in the iLumina digital library

Associating resources with virtual collections enables those resources to be found more easily, either by browsing the contents of a highlighted virtual collection (launched from a page such as that in Figure 1) or through standard search and browse interfaces. Figure 2 shows how virtual collections are available from the browse page of the Open Video Project, a digital library of video resources.

Figure 2. Virtual collections as browse choices

Adding virtual collections to search facilities, such as that of the iLumina digital library of educational resources shown in Figure 3, enables a user to perform a standard search but restrict it to a specific virtual collection, which could provide a more manageable and higher-quality result set than by searching the entire digital library.

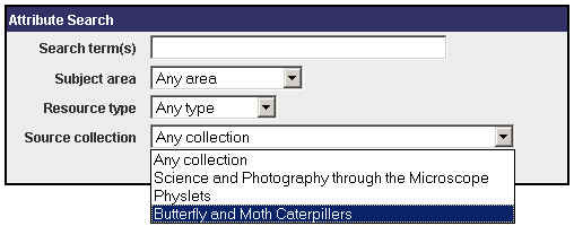

Figure 3. Virtual collections as search criteria

Looking at the use of the digital library from a "work-oriented perspective" [8], other benefits to the user stem from a more productive use of time. In [8] it is suggested that sub-collections can facilitate work by isolating a group of related content and enabling a user to focus on those resources. Defining virtual collections makes it easier for users to find and work with such groupings of related content, either through a listing of available collections as in Figure 1, or by a "related resources" link based on virtual collection associations and tied to specific resources. Additionally, the virtual collection description might include links to related information outside the digital library, thus guiding users to more materials for their work.

3.2 Benefits to Contributors

In most cases those who contribute resources to digital libraries (at least not-for-profit libraries, such as the new National SMETE Digital Library (NSDL)) are not directly compensated, yet digital libraries often depend largely on contributions for the content they provide. It is, therefore, in the best interests of the digital library to find ways to encourage new and repeat contributions. Virtual collections can benefit contributors in several ways. First, they provide an alternative distribution outlet. Contributors often have collections in which they have invested effort in creating and would like to see used more widely. Because a digital library will generally have a much larger base of regular users than the contributor, contributing the collection gives the contributor's resources more exposure.

By grouping a contributor's resources through a virtual collection the digital library helps maintain the resources' association with the contributor and in effect provides publicity and recognition to contributors. By explicitly highlighting virtual collections and the people and organizations that contributed them (and providing contact information and links to those contributors, as in the collection details page shown in Figure 4, for example), a digital library can increase the visibility of contributors. As a result, contributors are likely to benefit from more traffic to their own web sites, and can point people to their contributions at hosting digital library.

Virtual collections can not only help improve the "brand" of a set of resources and support their distribution, but can also offer basic infrastructure services. In some cases, such as with the Open Video Project where the resources (video files) are quite large, contributing resources enables the contributor to share resources without the overhead of storing and managing them, while retaining an association with them. If a contributor owns a large number of resources, this is a significant benefit itself, and one that has been taken advantage of by several Open Video Project contributors.

Finally, if the digital library shares information about resource usage, either directly to its contributors or as is increasingly common, through most recommended or top-10 lists, the contributor can gauge the relative demand of his contributions. This is helpful not only to contributors and the users of the digital library, but also "helps new contributors understand what is considered a good item" [7].

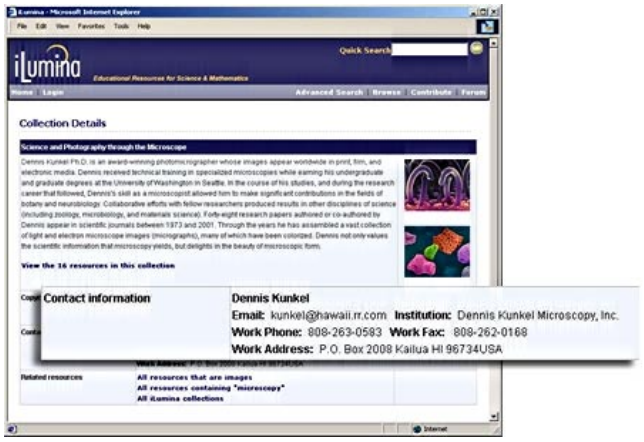

Figure 4. Details of a virtual collection including contributor contact information

4. IMPLEMENTING VIRTUAL COLLECTIONS

The benefits of virtual collections do not come without a price, of course. For a digital library to be able to easily create and remove virtual collections, to associate resources with different virtual collections in a flexible way, and to help users find and use the virtual collections, the library must have a structured approach to representing these collections. Moreover, to make creating such collections practical, this approach should also strive to minimize the costs associated with creating virtual collections.

In the remainder of this paper we describe an approach to implementing virtual collections based on our work in creating collections for two different digital libraries. We first review the current research related to representing collections in digital libraries and describe the costs and benefits of different types of metadata used to represent collections. We then describe how we used this information to define a collection-level schema for the iLumina and Open Video digital libraries and discuss practical issues related to implementing the schema.

4.1 Collection-Level Metadata

Metadata is a key element of any library, traditional or digital. Metadata is used by libraries to describe and organize item-level resources and by users to search and browse the library. Collection-level metadata performs a similar function for collections and is used in traditional libraries for discovery across collections.

Work on collection-level metadata from several fields including archives, museums, libraries, and the Internet is relevant to the design and implementation of virtual collections. As outlined in [13], each field defines collections differently and has different standards governing collection description. The past few years have seen a movement to create a standard for collection description that is informed by, yet transcends, the fields from

which it is derived. Work in the UK and the US has resulted in the formulation of goals for collection-level metadata and the definition and development of schemas to describe collections.

Based on work with the eLib working group on Collection Level Descriptions, the RIDING Clump Project created a searchable database of collection descriptions to provide information about what was available in its libraries [1]. The purpose of its scheme was to describe any type of collection—physical or virtual (electronic), networked or otherwise. RIDING collection metadata should allow users to discover, locate and access collections, search across multiple collections and allow software to provide services based on user preferences.

The Research Support Libraries Programme (RSLP) Collection Description Project developed a model allowing all the projects in its program to describe collections in a consistent, machine readable way [14]. The RSLP builds upon the RIDING goals above by requiring that collection metadata allow the refinement of distributed searching approaches based on the characteristics of collections.

The Alexandria Digital Library (ADL) is a research digital library project focused on geo-referenced geo-spatial information. The goal of the ADL is to create a single model that supports the four roles ADL identified for collection metadata: collection registration, network discovery, user documentation, and management [4].

Several themes emerge from this survey of requirements. First, it highlights the importance of establishing standardized collection-level metadata schemas that can effectively describe and manage a diverse set of collections and their metadata. Second, it argues that the schemas must support a number of functional library services that enable users to access collections and items, to search for materials, and to comprehend and use them effectively.

4.2 Types of Metadata and Costs of Creation

One challenge to creating collection-level metadata noted in the literature is the potentially high cost of production. Metadata can be automatically-generated or human-created [11] with the latter clearly imposing more significant costs in terms of human effort and time. In the context of collections, [4] describes two types of roughly corresponding metadata: *inherent* metadata, or information that can be extracted from the resource objects themselves, such as total number of objects or total file-size of the collection; and *contextual* metadata, or metadata which involves human judgment to create, such as a textual description of a collection of resources.

There are significant advantages to utilizing inherent metadata as much as possible. Because it can be generated automatically, inherent metadata has minimal costs associated with its creation and maintenance and can be updated on a regular, automated schedule. In contrast, human-created metadata is time-consuming, error-prone, costly to create, and more likely to be inconsistent. A person assigned to create metadata may only perform this task on an occasional, as-needed basis, and it may be a lower priority task than others for which that person is also responsible. Inconsistencies in metadata assigned to resources can arise due to variations in a given cataloger's judgment over time and because different catalogers may make varied judgements in cataloging resources.

There are drawbacks to relying solely on inherent metadata to define virtual collections, however. A risk in complete automation is the loss of many of the benefits of creating a virtual collection. Contextual metadata is important because it enables us to give some character and cohesiveness to the virtual collection. Indeed, a collection

> … is likely to be more than an accretion of all it contains; it has been gathered for a purpose. Human-created metadata is thus vital for articulating the scope, intent, and function of a particular collection—attributes that are likely to make the collection easier to locate, and easier to use. [11]

The use of contextual metadata and human judgment in selecting resources to be included in a virtual collection has other benefits. Virtual collections can be described "in terms of expected use in addition to being characterized by the terms they actually contain" [11]. Resources can be more carefully chosen for inclusion in a virtual collection, with consideration of expected use, resulting in a more concise collection of high-quality resources that is easier to for the user to search or browse.

It is important to recognize, however, that a collection-level schema that relies heavily on contextual metadata is relatively costly to implement and thus less likely to be maintained in the long term. A more viable approach is to define a schema for virtual collections that balances the costs and benefits of each type of metadata. In short, a cost-effective schema should include useful inherent metadata, supplemented by contextual metadata that captures human judgments of a collection's nature and the selection of criteria for inclusion in the collection.

4.3 Virtual Collections in iLumina and Open Video

Based on our review of the current research on representing collections, our first goal in implementing virtual collections was to develop an appropriate collection-level metadata scheme. However, to ensure the schema was sufficiently general, we applied it to two very different digital libraries: iLumina, a library of sharable undergraduate teaching materials for science, mathematics, technology, and engineering; and Open Video, a shared digital video repository and test collection.

Each digital library contains more than 1000 items and accepts contributions from anyone, subject to review before being made publicly available. Substantial collections of resources have been contributed to each digital library by a single person or organization. In iLumina's case, these collections include a group of light and electron microscope images (micrographs), several hundred "physlets," or small physics applets, and a collection of images and related information of butterfly and moth caterpillars. In the Open Video digital library, substantial contributions of video footage have come from a handful of organizations, including the Internet Archive, Carnegie Mellon's Informedia Project, and the University of Maryland's Human Computer Interaction Laboratory. A unique aspect of the Open Video Project is the large size of its video files, which has encouraged contributions from some organizations that lack the resources to store their video themselves.

In both iLumina and Open Video the resources of their sub-collections can be found through various searching and browsing mechanisms. However, for reasons discussed earlier, we felt that creating virtual collections to represent the contributed sub-

collections would benefit both the contributors and the users of these digital libraries. Specifically, our primary motivations for developing virtual collections were similar in each case: to highlight the work of authors/creators who contributed a critical mass of materials on a topic, to streamline the creation of item-level metadata, and to provide users with another way of accessing and understanding the items available.

4.4 Defining a Collection-Level Schema

RSLP's collection description schema was chosen by iLumina and Open Video because the set of elements was universal (it wasn't created to meet the needs of a specific digital library), yet provided the flexibility for customization, if needed. The RSLP schema was also selected because it is based on the Dublin Core schema [2]. Dublin Core is a common item-level schema used by many digital libraries, which would facilitate mapping elements and exchanging data. Previous work from the RSLP and the Collection Description Focus [6] resulted in thorough documentation, which facilitated understanding and implementing the schema in a relatively short amount of time.

Other projects currently use the RSLP schema to describe large, unrelated, relatively static physical collections in a digital environment. Our unique contribution is to use the RSLP schema to describe "born digital" objects of varying granularities, with varying relationships and at varying stages of collection growth in a digital library. However, the fact that RSLP is typically applied to physical collections meant that some elements and cataloging notes were not relevant to describing collections in a digital environment. This did not minimize the universal nature of the element set or render the schema unusable, but it did require us to review all the RSLP elements and choose the ones most appropriate to digital collections.

iLumina and Open Video formulated requirements used to select RSLP elements, and, more generally, to measure the success of implementing the RSLP schema. These included:

- Identifying a subset of useful elements that would yield informative and easily understood collection descriptions
- Minimizing the cost, in time and resources, of creating collection-level metadata

Useful collection descriptions can be created by identifying a subset of elements relevant to users, by ensuring that metadata is complete within a collection description and consistent across collections and by presenting descriptions in an easily understood interface.

Low-cost metadata creation can be accomplished by harvesting metadata automatically, by requiring the collection creator, rather than a cataloger, to describe their collections and by providing an efficient cataloging tool, such as the one shown in Figure 5.

Using the complete RSLP schema, collection-level descriptions were created for iLumina, Open Video and their sub-collections. It was important to identify the metadata source (item-level record, collection creator, subject-area reviewer) to track the cost of creating metadata. By starting with the complete schema we identified elements that aided understanding the collection. This process also identified extraneous elements, which were not included in the collection record interface and the collection cataloging tool being developed. The resulting subset of elements met our requirements: collection descriptions could be created with minimal cost while providing sufficient information to aid discovery.

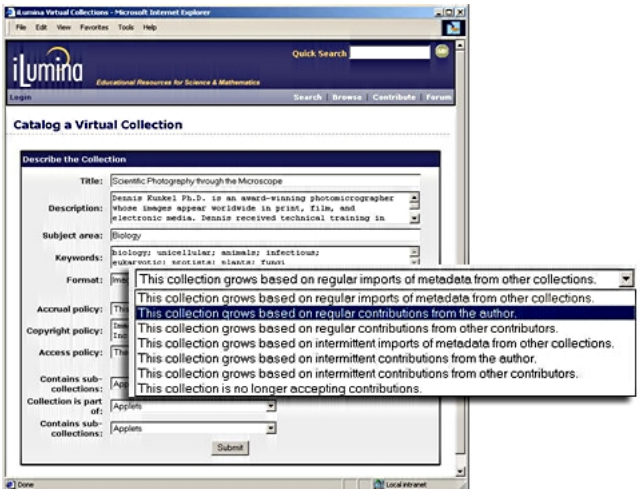

Figure 5. Cataloging tool for describing virtual collections in iLumina

4.5 Implementation of the Collection-Level Schema

Table 1 shows the subset of elements used to catalog collections in iLumina and Open Video; the RSLP suggested use for each field; and, the iLumina and Open Video decisions about the type of data to include in each field and who provides the information.

The RSLP schema contains 46 elements. Originally, iLumina and Open Video implemented a subset of eleven. After another iteration of testing and design, iLumina has implemented sixteen and Open Video seventeen RSLP elements. The element subset records four types of information about collections: a description, access policies, relationships to other collections, and collection owner contact information. The subset was chosen because the initial cataloging process consistently yielded data for these elements. The subset also matched the types of data reflected in item-level records. This provided users with consistent information between item and collection.

During the initial implementation, four listed elements were not included in the subset: Type and the three elements related to time. Because the RSLP schema has been used to describe physical collections, the developers created a controlled vocabulary to distinguish between collection types. We used the type element during the initial cataloging process, but found that the collections in iLumina and Open Video were often of the same type, so the same vocabulary terms were used repeatedly with no distinction. Also, the terms would have to be explained to collection contributors and users, which could be a barrier to cataloging, using, and comprehending the collections. Recently, however, Open Video incorporated the type element into its schema as a means to distinguish between virtual collections created for different purposes, such as collections organized around a specific contributor and collections containing resources from different contributors intended for a special purpose, such as a test collection. Rather than using the RSLP controlled

Table 1. iLumina and Open Video implementations of the RSLP schema

Field Name	RSLP Definition	iLumina Implementation	Open Video Implementation
Collection Title	The name of the collection	Manually entered by contributor	Manually entered by cataloger
Short Description	Not an RSLP metadata element	Manually entered by contributor	Manually entered by cataloger
Collection ID	A formal identifier for the collection	Automatically generated unique collection identification number	Automatically generated unique collection identification number
Keywords	Keywords or subject descriptors associated with items in the collection; recommend using LCSH level 1 terms	Keywords from item-level records are automatically displayed for manual selection.	Manually-entered with representative terms from Keywords field of resources
Subject Area	Free-text or structured text statement of the strength(s) of the collection	Subject terms from item-level records are automatically displayed for manual selection.	Automatically populated with value of Genre field (i.e., Documentary, Lecture, Advertising)
Role	The type of role, or function, carried out by a person	Manually selected from a drop down menu. Current roles (as defined by RSLP) include: Administrator, Collector, Creator, Owner	Currently not implemented in Open Video
Contact	vcard (contact information)	Contributors manually provide contact information	Contributors manually provide contact information
Format	Free-text field of the collection's physical or digital characteristics including number of items, total duration of items and physical or digital space requirements	Media Types (i.e., Image, Video, Audio) from item-level records are automatically displayed for manual selection.	Automatically populated with unique values from Format field (i.e., MPEG-1, MPEG-2, MPEG-4)
Type	The type of collection	Currently not implemented in iLumina	Manually entered by cataloger
Description	Free-text summary of the collection	Manually entered by contributor	Manually entered by cataloger
Date Collection Began	Range of dates over which collection was accumulated	Manually entered by contributor	Manually entered by cataloger
Date Collection Updated	Range of dates of individual items within collection	Automatically gathered from item-level metadata	Automatically gathered from item-level metadata
Accrual Policy	Free-text statement covering accrual policy, accrual method and accrual periodicity of the collection	Provide standardized statements which can be manually selected from a drop down menu	Provide standardized statements which can be manually selected from a drop down menu
Copyright Policy	Free-text statement of the legal status of the collection	Automatically populated with copyright statement from item-level metadata and manually modified.	Automatically populated with copyright statement from item-level metadata and manually modified.
Access Policy	Free-text statement; cover issues such as allowed users, charges, etc. and may give reasons for any restrictions in place	Provide standardized statement which can be manually modified	Provide standardized statement which can be manually modified
Contains Sub-collections	A collection contained within the current collection	Collection titles and IDs are automatically displayed for the contributor to manually select.	Manually-entered references to other defined collections
Collection is Part Of	A collection that contains the current collection	Collection titles and IDs are automatically displayed for the contributor to manually select.	Manually-entered references to other defined collections
Related Collections	Identifier or name of a second collection that is associated by provenance with the current collection	Collection titles and IDs are automatically displayed for the contributor to manually select.	Manually-entered references to other defined collections
Resource Count	Not an RSLP metadata element	Automatically generated through a query	Automatically generated through a query
Thumbnail	Not an RSLP metadata element	Manually created by contributor	Automatically generated through a query

vocabulary for Type, which classifies collections by curatorial environment, content or policy, Open Video created a new vocabulary more appropriate to its online resources. iLumina has not yet incorporated the type element but the new vocabulary Open Video has implemented for Type may be integrated into iLumina in the future.

Initially, iLumina and Open Video did not implement the three time elements because they seemed more applicable to physical collections. However, informal feedback indicated that users preferred to see metadata about time because it aided understanding the collection. As a result, iLumina and Open Video included two elements (Date Collection Began and Date Collection Updated). After identifying a useful subset of elements that would inspire our schema, we next considered how to minimize the cost of determining values for those elements and

how to extend the possibilities for expressing relationships between collections.

The cost of creating collection-level metadata can be reduced by automatically populating fields in the collection description. In iLumina and Open Video, "manually-entered" metadata is provided by the collection creator via a cataloging tool or by a subject-area reviewer when the collection-level metadata is examined. "Automatically populated" metadata is derived from querying specific fields of item-level resources within the collection. The Alexandria Digital Library uses "automatic" to describe the process of harvesting inherent metadata from the resources themselves and not from item-level descriptions [4]. In the case of iLumina and Open Video, collection descriptions are completely comprised of contextual metadata that is manually entered either at the item or collection level. Currently three fields in the subset can be automatically populated with metadata from the item-level description. For collection description to be cost-effective, the cost of item-level metadata creation must be minimized and more fields in the collection description must be automatically populated.

Implementing a cataloging tool with a usable interface for collection contributors is another way to reduce the cost of creating collection-level metadata. In the prototyped collection cataloging tool shown in Figure 5, metadata for other fields will be supplied in drop-down menus with standardized vocabulary or text boxes that can be modified by collection contributors or reviewers. This will ensure consistency in collection description. Also, a well-designed interface with clear instructions should minimize the cost of metadata creation in terms of a contributor's time. For example, when a collection record is rendered in XML, the elements retain their RSLP attributes; however, field names were changed on the interface (RSLP attribute "Concept" becomes "Keyword"; "Super Collection" becomes "Collection is Part Of"). iLumina and Open Video hope to pass the majority of the cataloging costs on to its collection contributors as a trade-off for having the collection publicized. iLumina will incur some cost through the involvement of the subject-area reviewer as they error-check metadata and recommend changes.

One aspect of metadata creation that iLumina contributors and subject-area reviewers share is identifying the relationships between collections and expressing them through the relational fields (Contains Sub-collections, Collection is Part Of, Related Collections). These relationships can be applied to collections of varying sizes and granularity, as in Figure 6, which shows the relational fields for the iLumina Digital Library, and in Figure 7, which shows the Physlets virtual collection within iLumina.

Though relationships are currently noted manually, in the future, relationships between collections could be inferred automatically. For example, when the Physlet record notes in field "Collection is Part Of" that it is contained in iLumina, then the iLumina record would automatically reflect the Physlet collection in the field Contains Sub-collections.

As collection-level metadata becomes widely used, we believe the relational attributes will be essential not only for discovering resources within single repositories, but also across digital libraries, such as the National SMETE Digital Library (NSDL). As it is currently envisioned, the NSDL will be a highly distributed set of collections and sub-collections tied together by a core integration system that coordinates services on the collections for users across the country [15]. However, the larger and more distributed the NSDL becomes, the more difficult it will be for users to find valuable resources and the (often small) collections they need. By explicitly representing not only a wealth of virtual collections, but also the relationships among them, regardless of their physical location, a collection-level metadata schema should greatly improve the navigability of the NSDL.

Contains Sub-Collections	Scientific Photography through the Microscope Physlets Butterfly and Moth Caterpillars
Collection is Part Of	NA
Related Collections	NA

Figure 6. Collection relationships for the iLumina collection

Contains Sub-Collections	NA
Collection is Part Of	iLumina
Related Collections	Applets Physics All iLumina collections

Figure 7. Collection relationships for the Physlets virtual collection

4.6 Concluding Thoughts on Implementation

Implementing a subset of the RSLP collection schema in iLumina and Open Video has been a success so far. The collection descriptions are one way to reward contributions to these digital libraries while providing another perspective of their holdings. Furthermore, our collection-level metadata schema can be used to describe informal dynamically-created collections, as well as more traditional ones that persist indefinitely. Although the current implementation therefore demonstrates several benefits of virtual collections, a number of extensions are possible that promise to add to the functionality of our collection-level schema and the virtual collections built on it.

Some ideas for further research include:

- Implementing a cataloging tool that further streamlines collection description by using drop-down menus and standardized vocabularies and by harvesting/mapping information from item-level records within the collection.
- Identifying a process to capture data relevant to collection- and item-level metadata regardless of which record is created first.
- Considering an extension of the RSLP schema to include education-related elements from the IMS metadata schema [5].
- Developing a separate metadata record specifically for intellectual property rights information that would be applicable to item and collection-level records.
- Providing contributors (and libraries) with services that will help them track the use of specific virtual collections, and potentially improve them. For example, contributors could be provided a URL that shows the web log activity related to their contributed resources.

- Devise methods for sharing virtual collection records across federated digital libraries that would give more visibility to authors and their materials.

Except for the last point, these ideas represent ways the current approach to creating and using virtual collections within a single digital repository could be enhanced. In the concluding section, we consider some implications of virtual collections across a large federated set of repositories.

5. CONCLUSION

The collection-level metadata schema we have developed and have started to test with iLumina and Open Video has already enabled us to define virtual collections that benefit both library users and collection-providers in several ways. But in a broad sense, the most important beneficiary may be the digital libraries themselves.

Virtual collections encourage us to see a digital repository not as a unitary structure, but as a modular construction comprising many sets of resources, some small and others large, some separate and others overlapping, some stable and others transient, some defined by the library managers and others established by library users. We think this is a compelling perspective. In fact, large-scale digital libraries are increasingly adopting just such a modular structure.

As we have noted, for example, the NSDL already consists of many dozen collections, housed in physically separate locations. One thing our work on virtual collections suggests is that each of the component repositories of NSDL would profit by describing its collection using a schema such as the one we have implemented. Furthermore, within these physically separate collections, there is no reason why sub-collections cannot also be described using the schema, just as we have described the virtual sets embedded in Open Video and iLumina. This leads to a perspective in which the NSDL is seen as a rich, possibly cross-linked, hierarchy of virtual collections.

There are a number of reasons this perspective could be attractive to NSDL. In the first place, as we have noted, it is often costly to create metadata. Item level metadata is the most costly of all, since it describes the "atomic" digital objects in a collection. However, it is often unnecessary to incur this cost: for example, when all members of a set of objects are similar, item descriptions are redundant. In such cases, collection-level descriptions will be more cost-effective than item-level metadata. However, a given repository may have some distinct items, as well as sets of similar components. This means that descriptions of resources in a collection should neither be fixed at a low level of granularity (item-level metadata) nor or at a high-level (complete collections), but must change as needed. In other words, a cost-effective way of describing a collection will require the flexibility of virtual collection metadata schemas such as the one we have presented here.

Virtual collections may solve other problems that also loom for the NSDL. One is that some collections that are contributing their materials to NSDL are often (rightly) reluctant to share all of their metadata. For example the Michigan Teacher's Network (MTN) [12] describes thousands of Web resources that can be successfully used with students in the classroom. These descriptions are, in effect, item-level metadata, where items are the reviewed learning resources. However, if MTN shared these records with the NSDL they would, in effect, be giving away all of their intellectual property, and might consequently lose their viewership (users might go directly to NSDL). One way for MTN to solve this problem would be for them to share (virtual) collection-level metadata—at whatever level of detail is appropriate—rather than item-level information. In this way NSDL could provide users with a broad description of resources, while the originating collection would still provide the most detailed information, and therefore, presumably keep most of their viewership. Indeed, their audience could expand, since NSDL, as *the* national digital library for science and mathematics resources is likely to draw many more users than its contributing collections alone can.

These examples suggest that collection-level metadata, if it is applied flexibly and cost-effectively to virtual and real collections, at various levels of granularity, can be often be more useful than item-level metadata. However, unlike item-level metadata, collection-level descriptions are only beginning to be investigated by the digital library community. Many hurdles remain.

In addition to the extensions and challenges noted in the previous section, for instance, it will also be important to test the generality of the schema we are developing by applying it to other repositories and collections, both within and outside of the NSDL, and which have no connections to either iLumina or Open Video. Planned usability studies will provide important data about the usefulness of our schema, both in terms of how easily and effectively contributors can create new virtual collections and how useful end-users find these collections to be.

Technical details concerning the transaction of collection-level descriptions among federated repositories will also need to be worked out, if metadata is going to be shared across a distributed library at low cost. Fortunately, many of the protocols that have been tested for item-level metadata should also apply straightforwardly to collection-level schemas as well. For example, the Metadata harvesting protocol [9] enables collection providers to easily expose their metadata to services providers. By agreeing on a standard collection-level metadata schema it should be as simple for repositories to exchange collection information as it now is for them to share item records.

6. ACKNOWLEDGMENTS

This work is partially funded by NSF DLI-Phase 2, grant #0002935.

7. REFERENCES

[1] Brack, E.V., Palmer, D. and Robinson, B. "Collection Level Description - the RIDING and Agora Experience", D-lib Magazine, September 2000.

[2] Dublin Core Metadata Initiative. Available at http://www.dublincore.org/

[3] Heaney, M. "An Analytical Model of Collections and their Catalogues." Available at: http://www.ukoln.ac.uk/metadata/rslp/model/

[4] Hill, L. L., Janee, G., Dolin, R., Frew, J. and Larsgaard, M. "Collection Metadata Solutions for Digital Library Applications", Journal of the American Society of Information Science, 50(13), p. 1169-1181.

[5] IMS Global Learning Consortium, In. "IMS Learning Resource Metadata Specification." Available at: http://www.imsproject.org/metadata/index.html

[6] Johnston, P. and Robinson, B. "Collection Convergence - The Work of the Collection Description Focus", Ariadne, 29, October 2001.

[7] Jones, P. “Open (source)ing the doors for contributor-run digital libraries”, Communications of the ACM, 44(5), May, 2001, p. 45-46.

[8] Levy, D. M. and Marshall, C. C. “Going digital”, Communications of the ACM, 38(4), p.77-84, April 1995

[9] Lynch, C. “Metadata Harvesting and the Open Archives Initiative”, ARL Bimonthly Report 217, August 2001.

[10] Marchionini, G. (1999). “Augmenting Library Services: Toward the Sharium”, Paper presented at the International Symposium on Digital Libraries.

[11] Marshall, C. C. “Making metadata”, Proceedings of the third ACM Conference on Digital libraries, p.162-171, June 23-26, 1998, Pittsburgh, Pennsylvania.

[12] MTN. “Michigan Teacher’s Network”, Available at: http://mtn.merit.edu/about/index.html, 2002.

[13] Powell, A. (ed). "Collection Level Description: A review of existing practice", An eLib Supporting Study, August 1999.

[14] Powell, A., Heaney, M. and Dempsey, L. "RSLP Collection Description", D-lib Magazine, September 2000.

[15] Zia, L. “The NSF National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) Program”, DLib Magazine, 6 (10). October, 2000.