SDS Final Report

Institute for Advanced Technology in the Humanities

Table of Contents

1. Overview

In 2000, the University of Virginia's Institute for Advanced Technology in the Humanities (IATH) and The University of Virginia Library's Digital Library Research and Development group (DLR&D) began a multi-year project called "Supporting Digital Scholarship" (SDS) funded by the Andrew W. Mellon Foundation. The project was co-directed by John Unsworth (Director, IATH) and Thornton Staples (Director, DLR&D). Worthy Martin (Interim Co-Director, IATH) replaced John Unsworth upon his departure from UVa, in the middle of 2003.

The SDS project's goals were to propose guidelines and document methods for libraries and related technical centers to support the creation and long-term maintenance of digital scholarly projects. The specific problems under examination were:

1. Structuring digital resources so that scholars can use them as primary sources;
2. The technical and policy issues associated with library adoption of "born-digital" scholarly research; and
3. Co-creation of digital resources by scholars, publishers, and libraries.

Over the course of the project, we have identified a host of other, closely related problems that confront the digital library and scholarly communities. Various legal, administrative, technical, and philosophic issues must be disentangled and resolved in order to make informed decisions about developing, collecting, and preserving this new type of scholarship. Both the creators and the collectors of this work need to be engaged with these issues and should develop a cooperative relationship.

We did not solve all or even most of the problems we identified. Some of the solutions must wait for technical developments, such as better tools, some require policy responses from the library and publishing communities, and some simply need more time for the scholarly community to internalize the opportunities afforded by digital resources.

While the SDS project clearly did not aim to set library or publisher policies, we undertook an extensive investigation of policy issues from the library's point of view. The results of those investigations are included here in Section 6. We hope that this report can be a step towards developing a useful and responsible digital library policy template.

On the technical side, we created prototype tools, such as the GDMS authoring tool (described in earlier SDS reports). We selected several born-digital projects, created by IATH Fellows with support from the UVa Library electronic centers and the computing support center (ITC). We then undertook prototype collections of these projects. Each "collection" provided valuable first-hand experience with the technical aspects of

transforming born-digital work into useful library resources (i.e., resources that can be maintained and accessed as primary scholarly works). The most recent work in this area is described in Section 5.

As part of this process, we identified an important method that can prove useful at all stages of collection. This method, currently called "significant properties," identifies the unique properties of digital works. It is discussed in Section 3. We also identified a key concept for assessing and collecting born-digital resources, which involves developing a limited set of collection levels that the library can offer. This is relevant to scholars and publishers as well as libraries, since it can lead to projects that are designed to be collectible. This concept is discussed in Section 4.

To summarize, the report contains six sections and four appendices. The next section examines the problem that the project set for itself. Section 3 discusses the significant properties method. Section 4 looks at levels of collection. Section 5 discusses our work in collecting born-digital projects. Section 6 contains the results of an extensive investigation of policy issues. The four appendices contain a list of committee members, an Author Questionnaire related to significant properties, a brief discussion of tools developed by SDS, and a summary of work mentioned in previous reports.

2. Statement of problem

The SDS project was a first attempt at exposing the issues associated with the long-term library support of born-digital scholarly materials. We have been concerned with two aspects of the problem: the born-digital scholarly materials and the library functions related to maintaining and providing access to those materials. UVa has therefore become an ideal laboratory for the project: several IATH projects have resulted in high quality born-digital scholarly materials and the Digital Library Research and Development group at the UVa Library has been building a robust framework for storing and accessing digital materials. We believe that the co-evolution of IATH and the UVA Library's digital library program has created an excellent context for investigating issues that are still well in the future for most academic institutions.

With regard to the UVa Library's digital library program, it is important to distinguish its goals from those of the "institutional repository" programs that are underway at many libraries. These programs strive to provide archiving services on demand, whereas a digital library must have selection policies associated with the value and veracity of content and with the library's ability to technically maintain the content and access modes for the materials in the long term.

Most of the projects that have incubated at IATH have created relatively complex electronic collections of heterogeneous digital materials, both digital surrogates of traditional materials and born-digital creations, in a variety of media and content types. As we come to the end of the SDS project, it is clear that, given the resources and appropriate support, scholars are increasingly creating projects like those at IATH. These projects are being built with digital primary materials that are either already in digital library collections or can be appropriately added to them. It is also clear that these scholarly projects themselves are a next generation of primary materials, upon which future digital scholarship will be built.

A basic goal of this project is to add scholarly digital materials into the stream of collection building activities of the library. While many of the issues in dealing with original digital scholarship are the same, they bring specific problems to collections development, data management, metadata, and digital library systems. We did not expect the SDS project to solve all of these problems, but we tried to foreground the scholarly user, examine key issues, and develop suggestions for policy and technical guidelines for the future digital library activities.

The problem that SDS is addressing is multifaceted. The overarching goal is creating, collecting, preserving digital scholarship, but it leads to a set of intricately related problems.

2.1. Complexity of digital scholarship

The term "digital scholarship" refers to a type of scholarship that is still young. Examples

include scientific reports published on the web in PDF files, specialized dictionaries distributed on CD-ROMs, intricately crafted web sites, journal articles distributed through on-line library catalogs, and so on. The unifying factor in these examples is the use of digital technologies to help create and distribute scholarly research and analysis. It is misleading to combine all types of digitally aided scholarship under one heading, since digital scholarship draws from a gamut of technical and scholarly tools, open-source and commercial, custom and standards-based.

Humanities digital scholarship uses the same range of tools but is showing more and more inclination for large-scale projects. A complex, constantly evolving web-based project that relies on custom-designed databases, scripts and dynamic search engines is light-years away from a journal article published in PDF format. Digital libraries have already begun to develop and test plans for collecting and preserving PDF files, CDs, and XML files formats, but many of the digital Humanities projects currently in development are actually thematic collections of digital resources, which are intended to support overarching intellectual arguments. These projects often have carefully designed infrastructures and elaborate interfaces that represent the information in a structured manner, so as to explain or support specific points in their arguments. The scholarship in these projects is not just in the content but also in the data mark-up, use of resources, and delivery. This is a notable departure from PDF files and illustrated narratives.

2.2. Characteristics of projects

To state the obvious, the most prominent characteristic of humanities digital scholarship is that it is produced by humanities scholars. Their projects are primarily concerned with literary and cultural analysis and documentation, but are produced in many fields. Perhaps one of the more notable elements of the more complex type of digital scholarship is its use of digital surrogates for resources. Traditional media, and electronic emulations of traditional media, can reproduce a painting or a photo of an original manuscript and include some kind of textual or audio description of the artifact. A digital scholarly work can create an XML structure to represent the resource, and attach all sorts of visual, textual, and auditory files to the structure. Metadata can provide additional historical and descriptive information.

This kind of work involves much more intra- and interdisciplinary collaboration than traditional format, and leads to what might be described as thematic research collections. Humanities digital works aim to present a complete set of evidence for their arguments: they use evidence that is directly related to the main argument as well as collections of resources that gives a virtual full-scale model of the subject. They give the user a complete set of information, a sort of virtual worldview that allows the scholar to make the argument and gives the user the tools to judge it.

2.3. Issues of structure: variable factors

All projects have some type of overarching structure, including an entry point and collections and subcollections of interrelated data, and most use the web to deliver information. While the particular details of the structure can vary wildly, especially in the early stages of development – each author has unique ideas about collecting and organizing digital research -- there are common characteristics that can be found.

2.3.1. Granularity

One of the primary characteristics of a digital Humanities project is the granularity of the data. In the case of articles that are encoded as PDF files, the project is a single object. Most digital scholarly projects tend to be collections of heterogeneous types of data, each type often having more than one function. Text, for example, can be found in surrogate text files; images of text; transcriptions; XML files that describe a text's structure; or text files that explain collections of texts. Texts that are created for the project often function as structural and descriptive metadata about the project as a whole and about the other parts, which also carry narratives that are primary content of the project. Projects have many categories of multimedia objects associated with them, all of which have a variety of possible formats. The analysis of a project for collection must account for the variety of formats included in the project, as well as for the content type and function of each of the granules.

2.3.2. Explicit/implicit relationships

Relationships among the various pieces of a project are a crucial, albeit often subtle, factor in creating a successful and useful scholarly work. A great deal of time and energy is exerted in deciding how to express relationships among resources in the project and how to explain those relationships to the reader. Much of this work is carried out in the project's back-end, in the information architecture, database tables, and scripts or applications that control the flow of information. When the project is preserved, these relationships need to be documented or recreated. This is a highly complicated and difficult problem on several levels, but a key is the documentation of explicit and implicit relationships.

The difference between these two is the level of communication between the parent and child objects. If the parent knows its children's IDs and location, the relationship is explicit. The Salisbury project's image archive, for example, documents a rigid explicit hierarchical relationship between the various components of the cathedral and its environs so that the archive can correctly associate images of the cathedral with the site plan. If, on the other hand, the children "know" who their parent is and the parent knows how to find them, the relationship is implicit. The Rossetti project uses work codes to signal relationships between certain paintings and poems.

2.3.3. Formalizing expressions of structure

Most scholarly projects of the type investigated in this project can be seen as highly organized networks of related units of content of different types. All of these units of content have at least one relationship to another. When collecting these projects into the digital library it is necessary to ensure that these networks of relationships are properly documented while fitting the whole into the digital library. For the projects involved in this investigation this is not too cumbersome because all of the content is newly created for the SDS project. The projects in their entirety can be imported into the digital library as long as the relationships are properly documented.

In the future this will become more of a problem, as projects are created around content that is already in the library, adding new content that represents the scholarly work of the project. The digital library must be ready to support multiple relationships among units of content without prejudice to any one context. The Fedora-based digital library architecture at Virginia is well suited to handle this challenge. The overarching collection is built as one large network of related units of content. A scholarly project that is a network itself can be integrated completely, becoming primary material for a next generation of scholars, creating new relationships for existing content and adding new units of content to the network.

Thanks to the World Wide Web, these projects are usually developed around a backbone of web pages that provide the linking texts that tie the disparate units of content together. Unfortunately, HTML does not provide a good way of formally organizing content, such that the content can be preserved separately from the presentation. Luckily, most of the scholarly content in these projects was contained in more formal formats, such as XML. Ideally, a project would be composed completely of formally defined XML-encoded content nodes, with the presentation of the data handled by equally formal but separate means. Again, a Fedora-based system proves to be well suited to the task.

2.4. Changing relation of library and author: involvement of library in creation and development, developing collectible projects

One unexpected effect of the growth of humanities digital scholarship is the increasing importance of the academic research library in the design and production of these projects. Regardless of what technologies are used to produce and distribute them, research libraries may be obligated to collect them and to preserve them for future users. This is a relatively straightforward task when the works produced are based on well-known standards and in stable formats. However, the technology being used to produce humanities digital works is constantly changing and the scholars themselves often do not have an adequate understanding of the tools that they are using. Librarians and archivists are gaining increasing technical expertise as they try to collect and preserve these works, and are becoming useful resources for scholars.

It is in the libraries' interest to be involved in the designing and building process,

whether by offering toolkits for authors, encouraging use of open source standards-based software, providing digitization and encoding guidelines, or offering technical support. Projects that are well designed and carefully built will be much easier to collect and preserve. The authors can benefit from such collaboration, too, since it can make the project easier to use and increase the project's likely lifespan.

This developing partnership takes the library into new territory. While some of these projects will be commercially published and distributed, some may only be available from the library's digital collection. If the library has also been involved in the creation of the work, it has acted as *de facto* editor and publisher and taken on a new relationship with the work.

2.5. Changing nature of library collection: selection, collection, archiving, and preservation

Digital materials have long been a part of library collections, and many libraries have or are developing policies and practices for handling and preserving newer formats as they come to them. However, these new guidelines in many cases do not encompass the more complex thematic collections that we are discussing here, but stick to simpler and smaller works (such as journal articles and CD-ROMs).

As authors become more ambitious and learn to use new technologies, their work will place as-yet unknown demands on the research library's ability to preserve them. While it is still possible to treat these works as special collections and to spend a great deal of time preparing them for collection and archiving, they are becoming more common and mainstream collecting practices will eventually be developed. In preparation for that point, libraries need to think about how to handle such projects at all stages: selection, collection, preservation, distribution, and deaccessioning.

3. Significant properties

The idea of trying to single out and measure certain intrinsic qualities of a project came from discussions between library and SDS staff. As part of those discussions, we tried to identify issues that need to be addressed from the beginning of the collection process, such as:

- What are the author's expectations from the library? What components does the author want or need to be collected?
- What kind of preservation can the library offer?
- What are the project's storage requirements, file formats, standards, etc.?
- Is the project completely documented (historically and technically)? If not, will the library have trouble providing the promised level of documentation?
- If the project degrades over time, will the library be willing and able to reconstruct its features?
- There should be a contract between the library and the author or depositor that clearly states the terms of collection and preservation

Further pursuit of those questions led to development of a questionnaire that collects basic information about a project and that can be used a starting point for discussions between library staff and authors or depositors. The SDS Author Questionnaire (included in Appendix 2) was developed as a tool for library selectors who are familiar with the subject matter that a given work covers but not with the work itself or the technologies behind it. It was designed to gather basic information about a project that can help selectors better understand the project and judge whether or not the library can collect and preserve the work.

The Questionnaire was then submitted it to four project authors (Jerry McGann for **The Rossetti Archive**, John Dobbins for the **Pompeii Forum Project**, Marion Roberts for **The Salisbury Project**, and David Germano for **The Samantabhadra Collection**). Their answers are available on-line at http://jefferson.village.virginia.edu/sds/questionnaires/. We then held a series of meetings with the authors and members of the UVa library selection and collection staff to discuss and critique the results.

Among the more interesting ideas to come out of these discussions was the notion of significant properties. These are those elements that are intrinsic to the project's identity and purpose. They include those parts of the project that contain the project's scholarship. These properties are in some ways non-fungible elements. They cannot be transferred, exchanged, or replaced without altering the project. Ideally, they should be identified early in the project's history, to be sure that they are well designed and adequately documented. It is difficult to form a definitive, closed list of significant properties, since they will vary from one project to another. Design elements,

stylesheets, databases, or content features that in one way or another encapsulate or explicate the project's reason for being are likely to qualify, however.

Identifying these properties serves several purposes. When combined with the library's existing selection mechanisms, they can be a useful tool that helps the library better understand the work and the implications for preserving it. They also can help the author more clearly express the project's goals and accomplishments and to better negotiate with the library. When identified and gathered together, they can be attached to a significant properties framework. If such a framework could be built and standardized, it could prove a useful tool for both libraries and authors.

3.1. Significant Properties and Preservation

When a work is collected, both the collecting library and the depositing author/creator need to agree what the library is preserving, what it can preserve, and whom it is preserving for. If the library intends to preserve the work's behaviors and features, the library must have information that will support emulation or migration. If it decides to preserve the project as a historical artifact – to preserve it in its collected state – the library will require a different level of information.

When the work is collected, the library staff may not possess a deep understanding of the subject matter and may therefore be unable to judge which elements will be most useful to future scholars. The author, on the other hand, may initially believe that every detail of the work must be preserved, but after some discussion may concede that parts of the work probably have only tangential value. Parts will have unexpected value, even if the work as a whole no longer functions as intended. They may become valuable as artifacts in their own right (a unique database or a custom application, for example). In that case, future preservation efforts would focus on preserving those useful elements, even at the expense of other elements. Perhaps the content is preserved elsewhere in a different format and therefore the library does not need to take heroic measures.

A thoughtfully designed significant properties framework could help authors and library staff identify what parts of a work are worth long-term preservation efforts and can aid future library staff trying to work with outdated technologies. If, for example, DTDs become a historical curiosity, the library may want to preserve them but migrate the content into a more current content delivery software or hardware. In this case, the challenge is not preserving the project's hardware and software, but preserving an understanding of the original intention behind the distribution of the information.

3.2. Identifying Significant Properties

There are several loose (and perhaps overlapping) categories for these properties:

- *Presentation.* This can cover visual and design elements related to the appearance, aesthetics, and look and feel of the project. It can include colors, fonts, decorative graphics, layout, and general design themes.

- *Function.* Those elements related to the organization and control of the data. Style sheets, java applets, databases, and scripts could all be considered functional properties. The site's interface(s) and layout might also fall into this category, especially if they control how data is presented to the viewer. For example, a map of an archeological site showing a picture of a building near the building's physical location could be considered a functional element. Links and indexes might also in this category, insofar as they control what data is displayed to the viewer.
- *Usage.* Properties related to the intended use of the project, or projections of future useful characteristics.
- *Content.* Elements that hold intellectual content or represent content, such as maps, essays, bibliographies, and databases.
- *Relationships.* Intellectual and encoded relationships, such as links and ID references. Intellectual relationships may or may not be specified in the software or documentation but can be implied from the way the project's databases, stylesheets, or DTDs are designed. Encoded relations can be intrinsic and extrinsic. Intrinsic relationships are supported by relational databases and mark-up technologies such as XML. Extrinsic relations are often between documents or objects, such as connections between databases and calls from an XML document to a non-standard application or object.
- *Navigation.* This can include structured or arranged paths, such as a table of contents or navigational buttons that move the viewer through the various components and information structures.
- *Development plans.* The long-term plans for the project. These plans may change as the project is built or may never be completed but they reveal which parts of the project the author considered definitive and which parts were more fluid. This information is useful for collection and preservation issues.
- *Historical value.* The author may have a good sense for which elements will have historical value. This can aid in future selection/deselection decisions.

One significant item that can be overlooked in this analysis is the importance of the project as a holistic entity. As a whole, the project has certain characteristics and features that may be lost if the project is decomposed into smaller parts that are more easily collected or preserved. Both the library and the author or depositor may want to keep this in mind when negotiating collection of a work.

4. Levels of collection

A library should potentially develop a set of collection options, ranging from a simplest level of collecting only project-level metadata to more extensive efforts. This can be associated with to a set of technical criteria, significant properties, and formal commitments that the library makes to the project's author. Note that these levels should be seen as a bottom line, either as the maximum commitment that the Library is willing to make or as the minimum commitment that the author is willing to accept. It may well be that at the time of collection the library may be able to collect the project and deliver it exactly as created, but because of some specific feature or technology it may not be willing to commit to always sustaining the project at that level.

The list below was developed both from technical experience developed over the course of SDS project and from discussions among the project team, IATH Fellows whose projects were used as test cases, and the library selectors for the areas that related to the projects. Note that each level builds on the levels below it.

Level 1: Collecting metadata only – At this level the project would be represented as a single object in the digital library which records that the project exists or existed in the past, and includes some descriptive metadata about the content of the project, people who were associated with it, etc.

Level 2: Saving the project as a set of binary files and metadata only – Only the most basic preservation would be attained at this level. Content files and possibly all the files associated with any custom software would be collected as standard binary files only. The same descriptive metadata would be collected as for level 1, along with technical metadata about the original formats of the files and any software that was necessary to use them. At this level, the assumption is that anyone interested in using the project would be on his or her own in trying to reconstruct it.

Level 3: The content can still be delivered as in the original – At this level, relationships among the content are preserved but no attempt is made to capture the exact action of the project or its look and feel. The user's experience may be different but the ability to navigate the connections that the author provided is preserved.

Level 4: Look and feel intact – The project operates and appears exactly as it was originally intended. The software may not be identical but every effort is made to recreate the user's experience as completely as possible.

Level 5: The project is completely documented – The project is preserved as a complete artifact, documenting its development and history. This could include ephemera such as e-mail archives from a project development team, reviews or citations of the project from other sources, documentation associated with grant proposals, etc.

5. Collection demonstration experiments

In the past year, work on the SDS project included two experiments that demonstrate different modes of approaching the collection task. Specifically, we were concerned with the stage of collection in which the structure of the scholarly project is formally expressed for the purpose of ingestion by a repository. In OAIS terms one might call this the SIP production stage. In the SDS context, the target repository was the FEDORA system that the Library is implementing (with further Mellon support), so the experiments could be described as demonstrating the effective creation of FEDORA SIPs under two modes of production.

The first mode is motivated by the injunction, "If you want it to be collected, build it to be collectable." It sounds superficial and tautological, but there is an important lesson at its core. If long-term preservation is to be effective and efficient (or maybe even viable), the repository's collecting, archiving and disseminating needs must be taken into account during the design and implementation of the digital scholarship project. We believe that one way this will come into general practice is via implementation of tools that aid scholars in authoring their materials. The tools should be designed to produce materials that accommodate the archiving needs of the repository. The GDMS Editor (see http://www.lib.virginia.edu/digital/reports/metadata/gdms.html), which we implemented as part of the SDS project, is intended to be such a tool and was used in the first experiment. This experiment, described in Section 5.2, demonstrates a wholesale recreation, rather than a completely new creation, of the project. The "recreation" aspect serves to provide opportunities for evaluation (e.g., was there an important part of the original project that could not be created in the new context or required inordinate effort to do so?).

Of course, many scholars set about creating their digital resources before any such repositories were designed, much less being implemented. These projects require a different mode of production in which the formal expression of the structure is built upon the existing, large and diverse digital project. Our second experiment demonstrates this second mode of production and is described in Section 5.3.

5.1. The Salisbury project

The Salisbury Project was created by Marion Roberts, a Professor at the University of Virginia, during her fellowship at IATH. The primary content for the project was the architecture of the Salisbury Cathedral. Professor Roberts created a formal description of that architecture and the description was used to organize a large collection of digital images of the cathedral. The original Salisbury project consisted of a well-organized HTML web site that gave access to the image collection; provided contextual information on four nearby locations and associated artwork; and included a teachers' guide, which contained a bibliography, a set of annotated links to other web sites, and guidance for classroom applications.

As described in earlier SDS reports, we handcrafted a process that transformed the EAD files and Dynaweb stylesheets into GDMS files and XSL stylesheets. Then the GDMS and XSL files and all of the image files were ingested, again by a handcrafted process, into one of the Library's early FEDORA-based test repositories.

In our first experiment over the last year, we worked with Professor Roberts and her graduate student to recreate the materials as if starting from scratch (please see http://dl.lib.virginia.edu/data/project/salisbury-html/salisbury). The formal description of the Salisbury Cathedral architecture with its numerous connections to the image set was correctly created using the GDMS Editing tool. In addition, the image collection was greatly augmented, as was contextual information on nearby locations (Old Sarum, the cathedral close, the old town of Salisbury and three parish churches). The newly created GDMS files fully captured the original and now augmented Salisbury project content.

Unfortunately, time and resource limitations did not allow us to recreate the necessary stylesheets to fully disseminate the Salisbury Project. However, we have demonstrated that the resulting files can easily be brought into the digital library as Fedora objects, representing a collection of the project as a whole and also integrating the five image archives of the cathedral and related locations into the Library's existing art and architecture collections. The Library has formally selected the project to be collected and, when time and resources allow, will create the rest of the XSL stylesheets that are necessary and create the Fedora objects required.

5.2. Expressing the Structure of the Rossetti Archive using METS

The Complete Writings and Pictures of Dante Gabriel Rossetti: A Hypermedia Research Archive (http://jefferson.village.virginia.edu/rossetti/) was begun by Jerome McGann, a Professor at the University of Virginia, during his fellowship at IATH. It was one of the first two IATH Fellow projects and has been under continual development during the succeeding years. That extensive effort has resulted in a very large number of content files. A major motivation for the Rossetti Archive has been to understand more fully and then to document the rich connections among Rossetti's works. Thus, relationships (of all the varieties discussed in Section 2.3) among the content files are crucial to the project. Indeed, the project's main premise was that the resulting digital resources would provide a model for building critical archives that would constitute the next step beyond paper critical editions of authors' collected works.

In an earlier phase of SDS, we completed a handcrafted process of ingesting the content files (text and image) into a test FEDORA-based repository. A major result of that effort was to demonstrate that the primary parent-child relationships among the content files could be effectively and efficiently represented in the test repository. As mentioned above, much of the new scholarly content in the Rossetti Archive is expressed in thousands of pointers among the content files. These relationships were still viable when the project was integrated into the test digital repository.

That process was extremely labor intensive, so in the last year we investigated a process by which the formal expression of the high-level structure of the project would be expressed in a METS file (or set of METS files) built on top of the original project files. A subset of Rossetti Archive, "The Blessed Damozel," was selected because it is one of the Rossetti's more complex works, manifested in both written forms (as poetry) and graphic forms (as a series of paintings). The results, we believe, will serve as a pattern for a METS expression of the structure of the entire project.

For our experiment with this mode of production the METS files were created by hand, but we wanted to investigate if the extensive detailed work needed to actually form the specific entries for the FEDORA-based repository could be automated. Again in OAIS terms, we wanted to investigate if the set of METS files could be an effective SIP for the FEDORA repository.

Our efforts did result in an effective METS expression of the structure of the selected portion of the Rossetti Archive, with regards to the information units of the content files (see Section 5.3 for more information). Due to time and resource limitations, we were not able to implement the stylesheets and programs needed to create the disseminators (in the FEDORA sense) for the digital repository objects from the METS files. The experiment was successful, however, in that FEDORA XML files were created via a single XSL stylesheet that defined all of the other characteristics of the objects and preserved the relationships among them.

We believe the process used for the selected portion of the Rossetti Archive will be an effective model for our later efforts with the whole project. The process is already serving as a model for a project (proposed to the Institute of Library and Museum Services by Ken Price, Co-Director of **The Walt Whitman Archive**, with IATH involvement) to formally express the structure of the Whitman Archive with METS files. Our experiment this year with the Rossetti Archive investigated the mode of building the structure expression upon an existing large project. The Whitman Archive will add an interesting twist to the question of collection because many of the texts in the Whitman Archive not only already exist, but are also part of a library production system (the texts are already disseminated as individual texts by the Electronic Text Center in the UVa Library).

5.3. About the METS files

An important element in the Rossetti Archive's design is the distinction between a "work" and an "instance." A "work" in this context is a concept that is expressed in various "instances," which could be a painting, sonnet, essay, drawing, and so on. "The Blessed Damozel" is a work that Rossetti created and recreated in poems, forms, and images. The concept of the work and instance here is similar to parts of the Functional Requirements for Bibliographic Records (FRBR). (For more information, see Barbara Tillett's SDS presentation "The FRBR Model" at http://jefferson.village.virginia.edu/sds/FRBR.htm.) Our METS expression of the Rossetti Archive structure has components for the works and instances as well as for

the project as a whole.

The METS file can be viewed on-line at http://jefferson.village.virginia.edu/sds/mets/. The files and their contents are described in the table below.

METS expression of Rossetti files	
project.xml	global copyright information, list of project staff, list of files, metadata about the project
work_damozel.xml	list of works, metadata about the works, list of file locations
instances_damozel.xml	list of files in the instance, specific metadata about these files, file locations

The project.xml file holds resources that are needed for project dissemination but do not have the fundamental content of the Rossetti materials (e.g., web html files and thumbnail images). The work_damozel.xml file contains metadata about "The Blessed Damozel" that is common to all (or most) instances and commentary about the work. The instances_damozel.xml file contains metadata and commentary about each instance.

This experiment used only information associated with "The Blessed Damozel" and the Rossetti Archive in general. However, we believe that the resulting METS encoding constitutes an effective example of how the remaining Rossetti Archive structure can be formally expressed in a suitable form for ingestion into a FEDORA-based repository.

6. Policy Issues

A digital repository serves an institution that collects, preserves, and disseminates digital scholarship. It must be guided by documented policies and procedures that 1) ensure that information is preserved against all reasonable contingencies, and 2) enable the information to be disseminated as authenticated copies of the original or as traceable to the original. To investigate these issues, we established a Policy Committee and charged it with recommending policy guidelines for building and operating repositories that support digital scholarship. (See Appendix 1 for a list of Committee members.) The Committee's recommendations are strongly drawn from the librarian's perspective, but authors and publishers have a stake in this work as well, and their opinions and interests should be considered when creating formal policies.

In developing these recommendations, the Committee drew on the June 2002 working group report "Trusted Digital Repositories: Attributes and Responsibilities," published by the Research Library Group (RLG) and Online Computer Library Center (OCLC). It discusses the concept of trusted, reliable, and sustainable digital repository. The Committee followed RLG's and OCLC's recommendation of using the Open Archival Information System (OAIS) framework of repository operation.

The remainder of this section is in three parts. Section 6.1 discusses the assumption the Committee decided to adopt. Section 6.2 discusses the general issues that were identified. Section 6.3 gives more detailed recommendations.

6.1. Assumptions

Current digital library literature reflects two implicit (and sometimes explicit) assumptions: first, that digital scholarly publications are and will be relatively simple, consisting of at most a few files, and that they will be created in or migrated to a handful of formats; second, that large, complex publications, with many interrelated objects and many significant functional properties will be too expensive for archives and libraries to collect. The complexity of collecting these projects is instead best addressed by emulation.

The Committee does not share these assumptions and feels that extensive and complex scholarly digital publications will become more common and that many will be deemed important enough to be collected and supported, regardless of cost.

On the other hand, there are some fundamental assumptions that the Committee believes are essential for successful operation of a digital library. These points must be discussed before any policy is drafted, since they are highly relevant to how the policy is interpreted and implemented. There are undoubtedly others that could have been included, but the Committee felt that these are the most important points. The brief discussions under each heading contain the Committee's recommendations.

Responsibility

While libraries are accustomed to building and preserving collections according to their own individual guidelines, digital collections require a community-wide approach. Responsibility for selection, collection, preservation, and access to digital work must be shared with other digital repositories (libraries, archives, museums, and related non-profit and for-profit organizations). This is not only more efficient and economical but ensures that the preservation community develops and maintains useful hardware, software, and technical and procedural standards.

Institutional memory

The long-term preservation and access of digital resources will be expensive (in both time and money), so the burden of preserving those digital cultural artifacts worth remembering must be shared. No single repository will be able to collect and preserve "everything" that is worth saving. A shared institutional memory will require hardware, software, and communication and procedural standards, which may be developed in cooperation with digital library user communities. Digital repositories must take the lead in developing these standards and in working with other communities that are studying and developing standards.

Control

Each digital library must control its own collection and maintain stewardship over all resources that it manages. That is, the library repository must have control over the files it collects. Digital content that is licensed to another organization or uses licensed access software is not fully under the library's control and will therefore be difficult or even impossible to collect. As a long-term strategy, the repository must cooperate with other repositories and with licensed content providers to develop strategies for developing preservation-friendly content and for transferring control of such content to a trusted repository (an example is the Mellon-funded electronic journal archiving project, at http://www.diglib.org/preserve/ejp.htm).

Future developments

It is very difficult to anticipate future technological advances, economic and political developments, and social changes. Current mainstream and cutting-edge technologies will inevitably be supplanted by new developments and breakthroughs, and while it may be possible to foresee advances in the next ten years, it is nearly impossible to know what will be mainstream in fifty or one hundred years. In all likelihood, a growing interdependence in scholarly communication between creators, producers/publishers, repositories, and users will place new demands on digital collections. Digital library policies and systems should therefore reflect current technical and social conditions but

must be allowed to evolve when appropriate.

OAIS

There is no standard or even widely accepted method for digital preservation and access and there is as yet no existing infrastructure that encourages cooperation and communication among digital repositories. However, RLG and OCLC have recommended the OAIS as the basis for a conceptual framework of an archival system that can preserve and maintain long-term access to digital information. The OAIS model has gained wide international acceptance as a framework for digital preservation and access and is being considered by the ISO. Any trusted digital repository should work within the broad framework of OAIS and participate in the ongoing international application of OAIS to the cultural heritage repository community.

Metadata

There are several emerging metadata initiatives that should be considered when developing a policy. Some deal with semantics and some with both semantics and syntax. In the latter category, the previously mentioned OCLC/RLG June 2002 working group report is of special interest. For example, the Metadata Object Description Schema (MODS), an initiative led by the Library of Congress (http://www.loc.gov/standards/mods/), is a descriptive metadata initiative for bibliographic information. The Metadata Encoding and Transmission Standard (METS) is a standard for encoding descriptive, administrative, and structure metadata of digital objects in XML and is maintained by the Library of Congress (http://www.loc.gov/standards/mets/). The Digital Library Federation endorses METS. SDS has studied both MODS and METS for recording metadata.

Long-term use vs. short-term costs

The current digital library literature assumes that for financial and technical reasons only relatively simply digital projects are collectable and that complex projects should be collected via emulation. SDS does not agree with this, and feels that short-term costs should not limit collection of complex scholarly projects.

6.2. Overarching issues

There are a number of overarching policy concerns that we have identified. This section will address issues that are directly related to specific aspects of digital repositories and indirectly related to more general matters.

As with the assumptions outlined above, the discussions under each topic contain the

Committee's recommendations.

Define the digital library's functions and responsibilities

A digital library's primary functions are similar to that of a traditional library: building large, organized collections of information resources that can be easily discovered and utilized by its designated user communities. However, functions traditionally performed by publishers and literary agents (editing, formatting, and distributing an author's work) have spilled over into the territory of the digital library. The digital library may need to clarify what responsibilities and functions it can fulfill, in part to educate authors and publishers who wish to deposit work and in part to identify what its policy must cover. If, for example, the library's resources allow it to offer an authorial workspace for building digital scholarly works and re-using resources already in its collection, policies must discuss issues such as security, copyright protection, and standards.

The library is free to create its own local selection policies

Unless legally required to collect and preserve certain materials, each library or archive is free to decide what it will collect and preserve. Local selection policies and procedures are answerable to the repository's user community, but creators of scholarly digital material should not assume that all scholarly material will automatically be preserved.

Levels of preservation

All parts of all works are not necessarily worth preserving. Nor are all parts of all works amenable to preservation. The care and maintenance of digital works is also expensive and some works may need intensive intervention and care as technology changes. A repository might choose to offer a menu of preservation options, ranging from the "bucket of bits" (no promises regarding long-term delivery) to ongoing high-level emulation, so that creators and depositors know what choices they have.

Policies for collections development must link to technical procedures about how and at what level materials are preserved, and at what level and how access is provided in the short- and long-term.

Control/authentication/access

There must be policies for access control, including authentication of users and disseminated materials, to ensure that all parties are protected. These policies must address the protection of the rights of producers, creators, and other rights holders after the point of collection, focusing especially upon fair use. The repository must have management mechanisms that serve these policies.

The library's policy also must clearly state what all parties involved can expect regarding these issues. For example, if the repository intends to maintain a "frozen" or fixed copy of the original work and disseminate a copy generated from the original copy, the depositor must be informed before the work is deposited.

Storage

There must be policies for storing resources, including service-level agreements with content creators. Repositories must have detailed plans for reliable long-term storage, whether in library-operated facilities or with outside providers. If resources will be stored in any 3rd-party infrastructures, the policies must cover service agreements, security, system maintenance procedures, disaster recovery, etc. Storage policies must support the repository's standards of preservation and access.

This issue will almost always require negotiation with the author or depositor before a work is submitted. There are technical and economic restrictions upon what the repository can reliably guarantee, so the author may need to settle for good-faith efforts.

Designated communities/knowledge base

There must be policies to identify the repository's communities -- user as well as author -- and those communities' knowledge base and particular needs. Libraries and archives have traditionally been obliged to protect the rights of creators and producers whose work they collected and to serve the interests of a designated user community. As in traditional media, the creators and users of digital scholarship often overlap, and there is likely to be a wide range of technical interest and competency. The repository should be sensitive to its communities' skill level and knowledge at all stages of policy design and implementation.

Updating policies as communities and technologies change

Policies must reflect both the current state of technology and the current state of the repository's designated communities. However, a system that neglects issues of long-term preservation and access inevitably risks obsolescence over time. On the other hand, digital information is prone to technological and intellectual vagaries, and even the most carefully thought-out techniques may fail. The repository must be prepared to regularly reconsider and update its policies as necessary.

Regular review of standards

There must be policies to guide the ongoing review and implementation of the

library community's digital standards. Assuming that the digital repository community agrees on some kind of standards and best practices for software, hardware, metadata, etc. and each repository regularly reviews them, the community should be committed to implementing these standards. Requiring creators to use standards-based tools and using standards-based tools for collection, preservation, and dissemination can accomplish this. The library community can also encourage the use of standards via university and professional organization policies covering the creation of digital scholarship.

Linking policies and procedures

Policies and procedures must be explicitly linked. Policy statements must document the reasoning behind the policy; relationships between the policy and any associated procedures; and the units at the repository that are responsible for oversight or implementation.

Creators vs. depositors

Traditional libraries rarely deal directly with the creators of the works they collect but instead work with a depositor, a legal entity that gives or licenses a work to a library. Digital repositories will continue to work with depositors but will probably also be negotiating with creators who are depositing their works. Repository policy must clarify this distinction, since there may be confusion over who is responsible for generating metadata, providing files, verifying project integrity, etc.

6.3. Policy

The committee began its deliberations with an explicit distinction between traditional and digital collecting, based on the widely held assumption that differences between analog and digital publications will require changes in library methods and policy. The committee first looked at activities and objectives associated with traditional library collecting and asked if any or all of them are relevant in the collection of digital scholarly publications, or if there are new activities that need to be added. The activities and objectives we focused on are:

- Selection
- Acquisition
- Preservation
- Description
- Discovery
- Access
- Control
- Deselection

The preliminary conclusion is that all of the traditional activities and objectives remain relevant. As new tools become available, the actual processes for carrying out the tasks may change and policies may need to be adjusted to determine how the tools should be used.

As noted above, the committee decided to use the OAIS as a basis for a conceptual framework of an archival system that can preserve and maintain access to digital information over the long term. SDS feels that the OAIS model identifies and handles key issues in a useful fashion and should be considered as a community-wide model. OAIS builds on the idea of information packages, which are conceptual structures for supporting long-term preservation. An information package contains a digital object and associated technical, administrative, and descriptive metadata. It is encapsulated by packaging information that binds, identifies, and relates content information and preservation description information and is discoverable via descriptive information about the package's content.

OAIS identifies three stages in digital archiving: collecting, preserving, and disseminating. Information packages or sets are associated with each stage: the submission information package (SIP), archival information package (AIP), and dissemination information package (DIP). These are used at the different stages and contain different types of information relevant to the content at that particular stage.

It seems logical that an analysis of repository policy should parallel this tripartite framework. We have therefore divided the digital repository's activities and objectives into three sections: 1) selection, submission, and collection; 2) archiving, control, maintenance, and preservation; and 3) discovery, delivery, and dissemination. Each section includes a recommended list of management tasks, a discussion of policies associated with that stage, and some specific issues that the Committee felt should be addressed in conjunction with particular policies. Note that, since some issues affect several aspects of the repository's functions, there is inevitably some repetition and some issues are addressed multiple times.

6.3.1. Policy guidelines for selection, submission, and collection

This section covers tasks associated with selecting and submitting works to the repository. In the OAIS model, the submission information package (SIP) is part of the ingestion process, which prepares works for storage and management in the archive. When the work is deposited, it is packaged in one or more SIPs and given to the repository. SIPs are joint productions of the repository and the work's creator or depositor, so it is important that each side understand what is expected of them and what they can expect from the other. Policies must consider the requirements of preservation, storage, and dissemination.

If a submitted work does not match the library's required formats and metadata or does not have adequate rights clearances, the library may decide that it is uncollectible. In that case, the library's policy must address how the repository staff

will intervene and work with the creator to make necessary changes if the quality of the content overrides refusal on a technical basis.

Repository management tasks

The digital repository manager should be given detailed social and technical gatekeeper functions related to the collection of resources into the digital library infrastructure. The recommended activities are:

- Work closely with the repository's designated community to advocate the use of standard practices when creating digital resources. This may include an outreach program for potential depositors.
- Negotiate for and accept appropriate information from resource producers and rights holders. Appropriate information includes:
    - Well-documented and agreed-upon decisions about what is selected for deposit, including required formats.
    - Effective procedures and workflows for obtaining copyright clearance for both short-term and immediate access, as necessary, and preservation.
    - A comprehensive metadata specification and standards for its implementation. This is critical for federated or networked repositories. It must include provision for rights metadata from content providers and technical metadata.
    - Procedures and systems for ensuring the authenticity of submitted materials.
    - Initial assessment of the completeness of the submission.
    - Effective record keeping for all transactions, including ongoing relationships with content providers.

Recommended policies

Policies need to cover the areas outlined below.

a. How to identify a "work"

There are existing criteria for identifying traditional works, but while those criteria can be used as a starting point for identifying digital works, the task is more complicated. Digital works often lack clear boundaries, normalized structure, or regular formatting, making it difficult to establish a stable

practical definition. While time and experience may lead to widely accepted conventions and guidelines for identifying a digital work, authors and publishers currently have virtually unlimited flexibility with digital media. A work may not have an appointed or intuitive starting point, a discernable logical file structure, or even coherent boundaries that separate it from other works. It may contain parts of other digital works or draw from common databases shared with other works.

Libraries are familiar with the philosophical issues involved in identifying what constitutes a work and have made practical accommodations for identifying and working with traditional media. However, it may be necessary to set policies regarding what is considered a work. For example, such a policy might require that, the work have a declared starting point, be reproducible as a discrete unit, and include specific administrative and technical metadata.

- **issue: responsibility for identification**

  If the library decides to develop a set of guidelines as to what constitutes a complete work, it should consider what role the author plays in identifying digital works. The library will likely not want to take on sole responsibility. The correlative question is whether or not the library can redefine what is and is not a work.

- **issue: sub-works**

  Many digital works contain databases or are complex web sites that undergo constant revision. Large text and image collections that take years to build are likely to be published in installments while still in development. A single work may contain sub-works that are works in themselves or are released as editions. A work's level of complexity does not predict the presence or absence of sub-works.

  If the sub-works are large or significant enough, it may be simpler to treat them as individual related works, although this can greatly complicate matters of preservation and persistence.

b. Selection

Selection policy must include documented programmatic guidelines for selecting materials that support teaching and research. Each library will have its own local selection policies and procedures and will not be obligated to accept or preserve digital material outsides its scope of collection unless the repository is fee-based or legally required to collect and preserve certain digital material.

c. Collection and acquisition

There must be policies for collection development (e.g., selection and retention) that link to technical procedures about how and at what level materials are preserved and how both short- and long-term access is provided. This crosses the boundaries of collection, management, and dissemination of digital scholarly publications. It is important that the expectations of the depositor match those of the repository. The repository may decide to offer different levels of collecting depending on how closely the work follows the repository's technical preservation requirements.

The policy must also document what makes a work technically collectible. This may cover minimal technical requirements, preferred standards, metadata, and update and revision specifications, including periodicity, as well as documentation of authenticity and rights requirements. If the library intends to automatically generate metadata for collected works, the policy must specify what type of information is required from the creator and how that information should be supplied (e.g., a list of all images and their URLs, formats, file size, etc.).

The policy must also state the library's requirements regarding copyrights and what kind of rights the library, the depositor, the creator, and the creator's heirs will hold after collection. The library may, for example, require the depositor to identify and acquire appropriate licenses and rights to all parts of the work.

• issue: levels of collection

A library need not collect all digital works equally. The library needs to have policy documenting its possible levels of collection and the criteria for making such selection decisions. Potential levels of collection may include, but are not limited to, collecting only project-level metadata and none of its individual resources; collection of selected resource files and metadata from a project without collection of the complete work and its project-level metadata; collection of metadata only for selected resources from a project without collecting either the resource files or the complete work; or collection of a complete work including project-level metadata and all individual resources and their associated metadata.

• issue: communicating with the depositor

The library needs to have well-documented agreements covering what is selected for deposit, including inventories of files and specific file formats. It should have documented procedures and workflows for

obtaining appropriate rights clearance for immediate and short-term access as necessary and for preservation, and for notifying depositors when the status of any or all parts of the resource changes. The library must also maintain records of all transactions with depositors.

• issue: technical requirements

The library must have comprehensive metadata and format specifications and agreed-on standards for implementation of those standards for collected works. Policy guidelines should include procedures and systems for ensuring the authenticity of submitted materials, checking technical requirements, and an initial assessment of the completeness of the submission.

• issue: collecting multiple editions

A stable or static work is relatively easy to administer, but many digital scholarly publications have large text and image collections, databases, and complex structures that are altered, revised, expanded, and developed over time. As a practical matter, though, the library may not want to collect active projects but instead will only collect complete works.

It may be more plausible to collect "states" of such projects. At various points in their growth, the project may have material that is sufficiently useful to be published and collected. The library should be able to judge if a proposed version is collectible, since it will be responsible for maintaining it. In some cases users will want real-time access to a work in progress, perhaps to see current entries in a database. Users may also need access to previous versions of a collection or database. This means that the library may find itself maintaining temporary versions of some parts of a work, multiple editions of others, and keeping track of a complex set of user permissions.

• issue: collecting all parts

A digital work's scholarly identity may be inextricably tied to the software that renders and indexes it, such as an XSL stylesheet that programmatically extracts and formats didactic text or a database of participants in an historical event. The identity of digital materials always involves an interplay of content and technology. The issue revolves around the extent to which specific renderings are considered essential, with respect to the identity and authenticity of

the intellectual object. Changing the work's look and feel in this case can seriously dilute the work's identity and value.

On the other hand, it may be impossible to collect and preserve the work in its deposited form. The library might try to break the work down into sub-works or, if the work as a whole is judged to have intrinsic value, separate data and data behavior when collecting the work. There may be instances where the library decides to collect only selected parts of a collection or work, or to collect content without the original rendering, providing instead a minimal or basic rendering. The repository might also decide to generate its own data behaviors to supplement or replace the work's original behaviors.

Alternative renderings may have a significant or even profound impact on the intended uses and purposes of the content. and can raise serious objections from depositors. The library should have mechanisms for judging when this is appropriate and for negotiating a solution with the depositor. The agreed solution should be documented in the collection agreement.

• issue: lifespan, revisions, versioning

Scholarship is an ongoing process. A large project may reflect years of preliminary research and take years more to finish. Indeed, there may be no concrete plans to ever complete the project. In most cases, though, there is some kind of intended or projected lifespan. Pre-collection negotiations must include the project's lifespan and how it might change in the future. The library's policies must discuss how long the library is prepared to support a project that is yet unfinished and how deselection policies might be applied.

It is the decision of a given library whether it will collect projects that are actively being created and maintained, as this implies assuming responsibility for the creation and maintenance environment. Policy will need to center around criteria for determining whether a work is complete and, if not, if the work is collectible and useful while production is ongoing.

The library must determine whether the work is complete and stable or will be undergoing revisions. If the work will be undergoing revisions and the library chooses to collect works where production is ongoing, the author and the library will need to determine whether or not some or all users need real-time access to the work in progress. The library will want to develop a set of possible options for this situation, including guidelines for authors who wish to offer limited real-time access to selected users. One question that quickly arises is

how much control the library has in selecting those users.

The library's policy must discuss how the library intends to handle revisions and versions of collected works and how conflicts with authors and project staff will be resolved. E.g., will the author expect to update or revise a work as he or she feels is appropriate? Will the library and creator be required to agree when a new version is ready to be collected? If the author wants to release revisions via controlled publication of stable editions or versions, the library must decide if it wishes to archive separate editions and versions. If the project is being developed with another unit or institution, that group should be involved at all stages of these decisions.

Another possibility is the collection of states (editions or versions) of the work, which may be useful in documenting creation and maintenance, functioning as and archival record of sorts. The library may collect states of the work in order to function as the primary means of access to the work. If successive states are collected, retained, and interrelated, the scheduling of collection of states also needs to be negotiated and documented in collection agreements. The economics of collecting states (the resources needed to collect, de-accession or interrelate successive states) will necessarily be an important consideration for the library in determining policy.

Clear lines of communication can avoid most problems, but clearly documented policies will help authors and depositors to understand what kind of control the library intends to exert and what issues the library is most concerned about. It is also important that all parties know who will be responsible for making decisions, accepting changes, editing, and making alterations.

• issue: deposition requirements

Deposition arrangements for a given work must be negotiated in advance with the depositor, including any special support or processing needs. The library may want to develop a standard set of deposition guidelines.

The library should ask for precise documentation when a project is deposited. This must cover not only syntactic information, such as data types and stylesheets, but also semantic information that explains what a particular tag in the DTD means or why the database uses a given set of tables. This helps the library make informed decisions about managing the project as well as helping future repository managers to preserve the project.

• issue: minimum descriptive metadata

The library must have a published set of minimum required metadata that it requires for deposited works. If a work does not have the essential metadata, the library may want to generate the missing data or help the author generate it. It is easier, of course, to generate this material while still building the work, so libraries may want to create cataloging tools or offer documentation and training to authors well before collecting their projects. Library metadata policies should address community outreach efforts that educate authors in proper tools and techniques for creating the minimum set of metadata for their works.

The library must have a comprehensive descriptive data specification and agreed-on standards, such as Dublin Core, TEI, MIX, or MODS. Metadata specifications must include both encoding practices (which elements to use) and best practices and any controlled vocabulary for use of the elements. Metadata policies must include standards for rights metadata and for representing technical metadata. This will prove especially vital for federated and networked repositories.

• issue: administrative and structural data

Policies must address who will be responsible for creating, identifying, and maintaining administrative data such as file inventories and creation and standards information. It should also consider who is responsible for creating structural data.

• issue: securing rights

Creators or depositors have responsibility for securing all rights for the work of others included in deposited digital works, and for making sure that these rights are transferable to the Library. All rights limits for each work used in a digital work must be specified in standard, machine-readable form. The library may need to provide copyright education, training in securing rights, and instructions for creating the required machine-readable rights record formats to creators and depositors. The library should have effective procedures and workflows for tracking rights status for both short-term and immediate access as well as for preservation; such procedures and workflows are necessary to support the Library's efforts to track and enforce rights for collected works.

d. Access and control

Once the work is selected, it will need descriptive, administrative, and technical data that describe its intellectual content, access methods, and rights in a normalized way that meets library standards. This descriptive information is required for the library's control over the deposited materials: the library must have control over the work and the files that make up the work, as well as relationships between files that reproduce the work on demand. Such additional normalized metadata are necessary to provide access to the works in the context of the rest of the Library's collections and metadata.

• issue: cataloging information

While issues of cataloging standards and vocabularies may seem of primary importance, the most important issue is actually policy revolving around inclusion in the Library's catalog or other access portals or gateways. The Library must document its policies on the inclusion of works in its indexes and catalogs, and the level or description (e.g., collection- or item-level) that will be found there.

If the work will be added to the repository's catalog, which means that cataloging information must be generated. This varies from descriptive metadata, in that library-specific formatting and vocabulary are used to provide specific points of access to resources. The library's policy must explain who is responsible for providing descriptive cataloging of works upon collection and if the author or depositor is expected to provide descriptive data to be used for cataloging. If so, they must be trained in how to generate and supply this data. The policy must also discuss the descriptive standards and practices that will be required.

• issue: restricting access

Some projects may want to restrict access to all or parts of a work, perhaps because of copyright, licensing restrictions, or privacy laws. When formulating policy, the library must decide if it can and should collect works requiring restricted access. If copyrights or licenses for the work require restrictions and the library is unable to comply with them, the work may be uncollectible. However, once the work is collected and is in the library's control, the library may, outside of legal and contractual restrictions, potentially make its own decisions about access. The issue of complex works containing a mix of available and limited access content must be addressed.

e. Deselecting

As with its traditional collections, the library must maintain a documented deselection policy for its digital collections. Such a document will describe guidelines governing deselection, including but not limited to relevance to current curricular needs, maintenance costs, and format and preservation issues. Such policies must be communicated to depositors at the time of collecting, and, if the collection agreement dictates, at the time of deselection.

f. Helping create new digital works

While traditional libraries support the creation of new scholarship by preserving and disseminating existing scholarly work and providing reference assistance and training, digital libraries will need to actively anticipate the needs of future digital scholars by collecting, preserving, and distributing digital resources in a way that encourages new scholarship. This new responsibility may include digitizing analog and print materials and allowing re-use of repository resources. Library policy must provide guidelines regarding recommended file formats for support and preservation; what level of authenticity the library can provide and the author can expect; and what administrative, structural, and descriptive metadata standards are required for creating digital resources. The library may also want to provide training in metadata creation to the author (or content creator), and technical documentation that explains what level of persistence and continuing dissemination of collected resources authors and depositors can expect.

An important question is what role the digital library will play in the continuing growth of digital scholarship. It may evolve into a quasi-publisher, providing authors with tools, training, resources, and publication opportunities. The policy should detail what kind of interface and support the library will provide to users and authors and what kind of role it expects to play in creating new materials. It should also discuss the author's responsibilities in producing a collectible work (e.g., the author is responsible for delivering a work that works within the repository's current infrastructure); when a work is considered collectable and deliverable; and how the authorial tools will handle revisions and updates.

There are also practical questions, such as how much space is needed to store resources, delivery specifications, interfaces to author tools, and so forth. If collected works can be made available to new authors for re-use, it is important to maintain the existing works' identity and authenticity. Re-use policy should protect the repository's resources and specify how those resources can be used, including metadata and credit line requirements. It must also state what kind of persistence and continued dissemination the library will provide for work created with in-house tools

and resources.

• issue: author workspace

If the library provides a workspace for authors or has in-house electronic centers that provide production support, there should be guidelines for communication and lines of responsibility between the various units. Those who are helping the author create a new work must be aware of policies related to collection, dissemination, and preservation. Policy should outline where the project's working files are kept, how much space an author has, who has access, procedures for deciding when all or part of a work is ready to be released, etc.

If the library supports the creation of resources by digital scholars or allows authors to re-use resources from its digital archive, the resulting works could be considered in-house productions. The library's policy should address how these works will be treated, especially if they are re-using already collected works, and if they will require special access controls (e.g., if they re-use resources with restricted access, will access to the re-used resources also be restricted? What if some of the re-used resources have different access and rights restrictions?).

• issue: standards and guidelines

The library must provide documentation for any required formats or metadata standards and vocabulary to the author and depositor. It must also explain the minimum level of markup and content needed for delivery. The library should also explain what technologies will be used for delivering the published work.

If authors want to re-use resources from the library's digital repository in a new work, the library must specify what technical and metadata standards the authors will use. If the authors are unwilling or unable to follow the specified standards, the library must negotiate an acceptable alternative. Otherwise, there may be conflict over who has the right to decide what technologies can be used in creating new works. When developing policy related to this topic, the library may want to confer with its user and author communities and with electronic publishers, so as to encourage cooperation and use of the agreed-upon standards.

• issue: support

If the library is providing staff, hardware, software, training, storage support, and/or guidance and quality control for authors it should describe these services in its policy, along with any conditions that it will impose on use of these resources. It should also discuss rights and ownership issues that may arise. These services can be quite expensive and may require a large investment of time and energy from library staff.

• issue: working with other institutions, repositories, and production centers

If other institutions, repositories, or production centers are involved in the creation of the work, the library must set up and maintain clear lines of communication and responsibility between itself, the author, any project staff, and whatever other units are involved. Agreements between the library and other institutions, repositories, or production centers should be documented and should provide for oversight of fulfillment of responsibilities and a means for resolving disputes.

Library policy should aim to streamline communication between the project's staff, the library staff, and non-library staff. It might outline rules such as including key library contacts in project discussion lists and management decisions and identifying key issues such as where resources will be stored, agreements for citing object usage and conditions of use, and required production or project participation credit for generating or hosting digital resources.

• issue: sharing works with other institutions and repositories

A work may be deposited in one place but shared with several other institutions. Authors at these other institutions may want to re-use the resources in this work for their own works, which may complicate rights issues, especially if there are limits placed on resource access. Library policies should address how individual elements in its collected works can be reused and under what circumstances (note that this may need to be coordinated with projects' particular access restrictions). It must also address issues such as whether the library can share digital resources and with whom; if it will allow another repository to collect a digital work that is based on its resources; and who is responsible for maintaining these resources.

• issue: working with commercial publishers and agents

If a work is commercially published or is distributed by an agent with a commercial interest in the work, the library will need to negotiate with the publisher or agent when soliciting a collectible version of the work. It is important to have clear lines of communication and responsibility between the repository staff and the publisher and to document any agreements or contracts.

It may simplify negotiations if the library has well-developed policies outlining the conditions under which a work is considered collectible, how the repository handles versions or editions, and so forth.

Other questions may arise. If the publisher has released multiple editions of the work, is the library obligated to archive them all? Can the work be re-used and distributed by another outside distributor? Will the work be restricted to certain users? If so, who will decide who should have access and how access will be restricted? What kind of expectation should the library offer regarding persistence and continued dissemination of the resource?

6.3.2. Policy guidelines for control, maintenance, and preservation

This section discusses issues related to creating and maintaining the archival information package (AIP, in the OAIS model). The AIP is produced by the library from the SIP and library-generated metadata, and is intended for internal use only. Once the SIP is deposited, the library can analyze its relevant properties and prepare it for ingestion into the repository. The library's control mechanisms can track the SIP from ingestion forward. These mechanisms should also be able to manage the AIP after it is moved into archival storage. Preservation mechanisms should then be activated to preserve the AIP.

Repository management tasks

The digital library repository manager's tasks include functions centered around maintenance of an AIP. Tasks include collecting and generating descriptive, administrative, and technical metadata and documentation, and monitoring and updating the metadata to reflect changes in technology and access arrangements.

The digital repository manager should be assigned detailed social and technical functions related to archiving of resources. The recommended activities are:

- Detailed analysis of an object or class of objects to assess its relevant properties. This analysis should be automated as much as

possible and should be informed by collection management policy, rights clearances, the designated community's knowledge base, and policy restrictions on specific file formats.

- Creation and verification of bibliographic and technical metadata and documentation to support long-term preservation of the digital object, according to its relevant properties and underlying technology or abstract form. The metadata should be monitored and updated as necessary to reflect changes in technology or access arrangements (which will influence the creation of preservation metadata).
- A robust system of unique identification.
- A reliable method for encapsulating the digital object with its metadata in the archive.
- A reliable archival storage facility, including
    - A program of monitoring media storage conditions
    - An ongoing program of media refreshment
    - Geographically distributed backup systems with regular frequent backups
    - Routine authenticity and integrity checking of the stored object, including references within the "object space"
    - Disaster preparedness
    - Response and recovery policies and procedures
    - Security policies and procedures

Recommended policies

a. Control

The library must generate its own descriptive, administrative, and technical metadata in order to control the works it collects. This is metadata above and beyond the work itself and is intended for internal repository-use only.

In this context, control involves knowing what the work contains (both intellectually and technically), where its files are located, how its content can be accessed, what kind of rights are attached to the files, and so forth. Control policies impact the physical collection of resources from the producer and ingestion into the library's infrastructure, including

determining what files and formats are selected for deposit, copyright clearance, and an initial assessment of the SIP's completeness. There must be consistent record-keeping policies for all transactions with content providers.

The library should also know how the files are structurally interrelated to form the work object; whether such interrelations are implicit (such as the relations in RDMS or id/idref relations in an XML document) in the work object, or explicit, such as between databases, or between databases and XML documents, or between XML documents. If such structural relations are not maintained, then the integrity and authenticity of the work object is compromised, perhaps disastrously.

Some libraries may give control over files in the collection to trusted outside agents. In that case, library policy must cover what kind of guarantees the agent must be required to demonstrate and what kind of security, redundancy, disaster recovery, and storage facilities are required. Responsibilities must be clearly assigned. Otherwise, the library must have its own back-up and data recovery policies and systems.

One potential area of conflict is the difference between a collected work (the SIP) and a distributed work (the DIP). If the library needs to alter a work in order to preserve and distribute it (i.e., the distributed version differs from the collection version), the creator may feel that the work has changed notably. The library's policy should justify and explain these alterations so that the library is not obligated to defend its decisions. It should make clear that, once the work is collected, the library's primary obligation is to that collected work. The policy should also outline a process for negotiating any disputes.

- **issue: physical control**

  The library must have physical control over material that it collects. However, parts of a work may reside on outside servers and for whatever reason cannot be physically collected. In this case, the library should have policies for ongoing assessment of the trustworthiness of the outside server, and for verifying that the agents running the server have adequate policies for protecting the works.

- **issue: persistent identifiers**

  The library must have a robust system of unique identification that tracks objects from deposit forward. If the work is stored in a third-party storage facility, the library's identification system must be maintained, even if the third-party has its own scheme for persistently identifying works

• issue: minimal metadata

The library must have a known set of required metadata, which is documented for content providers before collection. It should also have a reliable method for encapsulating the work's components along with its metadata in the archive so that data is not corrupted or lost over time.

The library must also have a known and documented set of internally generated metadata that will be assigned to all collected works at collection.

• issue: storage facility

The library must have a reliable archival storage facility, with documented abilities to perform the tasks listed above. If the library cannot provide these services itself it may choose to contract to a third-party facility.

• issue: checking state of digital collection

The library should develop automatic procedures for checking the integrity of a newly collected work before archiving it. The library should also plan to run programs to verify that the project works in the library's environment and to verify its links. If the work fails these checks, there should be procedures for identifying the problems and solving them. If large-scale changes are required, the library may need to negotiate with the author or depositor.

• issue: access

A depositor or the library may want or need to limit access to certain parts of a collected work. Parts of the work may have different usage rights and may require varying limits. The library will need mechanisms that allow it to control and record access to all or parts of a work. The policy should discuss how access would be tracked and recorded.

• issue: re-use

If the library is going to allow re-use of all or part of the work, the control policy should discuss how this would be enabled. There are several technical problems that may arise, especially if the current version of the work will be updated with new versions or editions in the future. Providing persistent links to multiple versions and portions

of a work is key functionality in supporting re-use.

b. Preservation

Preservation is a fundamental responsibility of a library. Digital technologies have a short lifespan, so the library needs to devote a substantial amount of time and thought to designing short- and long-term preservation strategies. Even a work that uses up-to-date open-source, standards-based tools might require intensive and expensive intervention in six month, five year, or twenty years. It is likely, depending on the variety of content encodings and renderings supported, that intensive and expensive intervention will be ongoing. Carefully planned and clearly explained policies can help the creator and the preservation staff anticipate and plan for this.

It is not enough to preserve the intellectual content, since the scholarly and historical value in a digital work is also in the mark-up, style sheets, databases, and the user interface. The library's digital repository may also be the refuge of last resort for unconventional but valuable research (analogous to items now collected by special collections). That said, all works are not worthy of preservation and some parts cannot be preserved.

A partial solution may be to offer levels of preservation, possibly ranging from the "bucket of bits" to emulation. It is best if these options and indeed the entire preservation policy are in-line with other libraries in similar communities. For example, if a library is attempting to preserve a work of unusual format but exceptional content, it could draw on a community of digital libraries that have a shared understanding of accepted preservation techniques and their application for such formats. In theory, if content and associated behaviors are maintained in standard forms (whether industry standards or internal library standards), it will be that much easier to preserve a work.

The library must provide documentation that clearly states what kind of preservation, persistence, and dissemination the library can and will provide. If the library is offering levels of preservation, the policy must explain the technical criteria used when deciding what level to apply to a given work. Preservation policies should include persistence of intellectual content and behaviors.

The policy should note that a work's long-term relationship to resources outside of the library's control is impossible to guarantee, since the library cannot be responsible for resources it does not control.

- **issue: standards-based projects**

If a work follows recommended standards for design, content, and delivery tools, it will (theoretically) be easier to preserve and control. However, if it uses commercial, proprietary, or otherwise non-standard tools it may require substantial effort to preserve.

Once the work is collected and is under the library's control, the library may choose to modify or alter the project to use standards-based tools. It may require the creator or depositor to cooperate with the library staff to make necessary changes. Given the current resource intensive nature of transforming proprietary and nonstandard or substandard works into standard works, this must be negotiated prior to collection.

Even so, the library may be required to make further changes as technology changes. As a matter of course the library should ask the creator to provide detailed documentation explaining how the individual elements of the work were built and how they work.

• issue: versions and editions

If the library decides to collect works-in-progress, the preservation policy must address versions and editions. If the library decides to collect states of a work, as discussed earlier, it may decide to provide continuing access to prior states. This may be necessary, especially if other projects re-use resources only available in older versions, but it is likely to prove a strain on the repository's storage facilities and access mechanisms.

If the library has done a great deal of post-collection work preparing the work for archiving and dissemination, it may consider allowing the creator to use a version of the library's copy of the work when preparing new releases. The library might also consider offering technical guidance, or even an authoring and editing workspace, to be sure that the next state is sound. This brings a new set of responsibilities and can have a significant economic impact.

Preservation policy must support updating and corrections to archived work, but should set reasonable limits (i.e., if a large database plans on publishing new versions every week for months on end, it may not be practical to preserve each version). It may also offer time limits on how long it can maintain access to all previous states.

• issue: persistence

If a work's content and associated behaviors are maintained in a standard form, the library theoretically should be able to preserve

these elements over an extended time. If a work does not follow these standards, it may have serious problems with data persistence and reliability.

The library must have a policy addressing the unique intellectual and system (file) identification of its works and describing its commitment to maintaining the persistence and integrity of such identification. It should also describe the form of citation that authors can use in citing the works that will provide the greatest amount of assurance that the referenced work can be located and retrieved.

References to works outside of the collection (i.e., outside the library's control) or references from works outside of the collection are a particularly difficult challenge. The library cannot control both ends of the links in these situations and the current lack of community-based standards for persistent identifiers and addresses makes it very difficult to regulate connections between resources in the repository and resources outside. Library policy should also consider more detailed questions about links, such as whether or not a work's links are embedded in the data representation or will be maintained separately.

If there are parts of the work that are treated as sub-works, the issue becomes more complicated. The author may want to release editions of one or more sub-works in addition to editions of the work as a whole, with library-assigned permanent identifiers for all of these overlapping parts. Parts of the work risk being preserved in different stages in different editions.

- **issue: significant properties**

When the library collects the digital objects that make up a work, it must create or obtain a detailed analysis of works and individual objects to determine their significant properties. Those properties help define the objects and should not be compromised. A digital object's significant properties dictate its underlying technical form, which must be documented and supported, and the amount of metadata that must be stored alongside the bytestream to ensure that the object is accessible at the agreed-upon level. The more significant properties deemed necessary, the more associated metadata that will be required. The significant properties might be an object's textual content, DTD, stylesheets, software for running video files, etc. Analysis should be automated as much as possible and informed by the collections management policy and restrictions on specific file formats. Identification of significant properties will assist libraries in developing object classes that can be used as benchmarks for

assessing the potential collectibility of digital objects, as well as aid in the management and preservation of its collections.

c. Authenticity

The library should have procedures and systems for ensuring the authenticity of materials in the collection. Authenticity in this context refers to the security of the objects that comprise the work as well as the work's characteristics (e.g., technical, intellectual). The AIP should contain metadata that makes this possible.

• issue: library-generated behaviors

The creator may feel that the only way to maintain a proper level of authenticity is to maintain the work's look and feel by preserving and maintaining the original stylesheets and formats, but if this is not possible the library must be able to identify and preserve the work's intrinsic value. Even if there is scholarly value in both the intellectual content and behaviors, the library may decide to replace or supplement collected behaviors with library-generated behaviors. In that case, it may want to authenticate both sets of behaviors.

• issue: versions and editions

If a work in progress is updated, the new state must be authenticated. If the previous state is still being disseminated, it must also be authenticated.

6.3.3. Policy guidelines for discovery, delivery, and dissemination

The deliverable version of the work is not necessarily the same as the collected version or the archived version. The OAIS delivery information package (DIP) contains the result set of the user's query to the repository user interface and finding tools and is assembled on the fly. It contains the requested data along with appropriate metadata. The DIP's content is not the same as the AIP or SIP: the disseminated version of the content may have a different look and feel than the submitted version. The library should consider informing users of this: if the DIP is markedly different, it can add a notice informing users that they are using the library's version of the work. This should be clearly explained in the policy.

Dissemination policies must be related into submission and archiving policies. The digital library repository manager's functions at this stage should relate to development of the DIP, such as analysis and documentation of the use requirements of the repository's designated communities.

The library is obligated to provide access to its digital collection, and standard use

should be supported, including reading, printing, and downloading and reusing digital resources (if allowable). The library must be able to control access according to license agreements with content owners and providers. The library controls the point and means of dissemination, but dissemination of collected works should not compromise the long-term preservation of those works.

The library should treat works in the collection as scholarly reference materials that can be reused in new scholarly digital works, but it must balance copyright laws, use restrictions, and collection agreements against its research obligations.

Repository management tasks

The digital repository manager must be given detailed social and technical functions related to discovery and dissemination of resources. The recommended roles are:

- Analysis and documentation of the user needs of the repository's designated communities.
- Ensure that the information to be preserved is "independently understandable" to the repository's designated communities.
- roduce well-maintained and documented technical metadata about the resources that matches the knowledge base of the repository's designated communities and with changing technologies.
- Make the preserved information available to the designated community. This should involve a system for discovering resources, appropriate mechanisms for authentication of digital materials, access control mechanisms in accordance with licenses and laws, mechanisms for managing electronic commerce, and user support programs. The information should be rendered in a useful format.

Recommended policies

a. Discovery/Delivery

A well-designed discovery tool and delivery mechanism is crucial to a digital collection. The library's minimal metadata and format standards must support the dissemination stage. The discovery interface and its specifications must be taken into account when the library is developing its collection and archiving standards.

The library's policy must describe a testing and review process for discovery and delivery and documenting a user interface. It should also describe who is responsible for implementing overall system development

and the testing and review process. The library should perform ongoing or periodic user studies to determine and document user needs, and, to the extent that it is economically feasible or practical, design discovery and dissemination apparatus that meet these needs

- **issue: assimilating with other on-line catalogs**

  If the library already has an on-line public access catalog (OPAC), it should decide whether or not its digital archive will have separate discovery and delivery tools.

- **issue: responsibility**

  The library is responsible for setting discovery, interface, and delivery specifications, but it must consider user needs.

b. Controlling access

If the library's collection agreement with an author or depositor limits access of the work, the library must have or develop mechanisms to fulfill the agreement. Since the library must have the technical means to impose such limits, it might be easier if the library has a pre-determined menu of options that can be offered to authors and depositors during the collection negotiations. It can offer, in accordance with emerging law and convention, a variety of copyright and access restriction profiles.

For example, the library might have three possible levels of access: a resource could be globally available, restricted by domain, or restricted to specific communities or individuals. Where restrictions are necessary, various techniques could be used control access, including (but not limited to) login identifications and passwords.

The library may need to keep access logs of who viewed what and when (although that raises thorny privacy issues) and policies covering who should be able to view that log and for what purposes. The library must have clearly documented policy in place about the purpose and confidentiality of its usage logs and the sharing of usage statistics with the depositor. For example, it might state that user information will be used to control access, and only for this purpose, and will not be used for any other purpose, such as tracking the reading habits of individuals.

- **issue: persistent linking**

  The library must be able to handle persistent links and bookmarking. Users will expect to be able to mark parts of works for later reference or citation. However, authors may want to encourage access to

individual resources within a work through the work's particular navigation apparatus. The library may want to declare whether or not it is willing to negotiate this point. It seems likely that it will not, since it would be a difficult aspect to control.

Technically, bookmarking and persistent links are a demanding problem, since while the various elements of the collected work will undoubtedly be assigned persistent identifiers (PIDs), the work's DIP may not be intended for long-term usage. If the DIP utilizes session-specific identifiers in its URL, the pages marked by a user will not persist once the user's session is ended. The library will need to develop a reliable method for users to find and repeatedly retrieve resources.

• issue: enforcing use restrictions

It is absolutely essential that the library be able to enforce restrictions related to copyright and ownership when disseminating and delivering resources. Policies must support enforcement efforts as well as user education about copyright and fair use restrictions. Policy must define who is responsible for protecting restrictions and how these restrictions will be communicated to users.

c. Authenticity

The library must verify the work's authenticity during dissemination. In order to establish trust, the library must describe in detail the methods and techniques used in ensuring and verifying authenticity. Policy must describe what kind of measures will be provided, what types of records will be kept, and who is responsible for overseeing and auditing this system.

• issue: granularity

The library must maintain an appropriate level of granular control on authentication and rights management. The repository must devise standards and techniques for ensuring and verifying the authenticity and integrity of works, as well as sub-works that comprise it.

The policy should note that if an item is outside the library's control (e.g., located on an outside server), the library cannot guarantee its authenticity. The library may be able to assess whether or not an outside source is reliable and can authenticate its own resources.

• issue: automation and record keeping

The authentication mechanisms should be as automated as possible,

but policy should describe security for the authentication system. The library must have documented policy in place about the confidentiality of its records and authentication statistics.

d. User and creator communities

The library must be familiar with the users that make up its community. This user community may in fact be made of several distinct groups, especially in larger research libraries, but as a whole it comprises the library's target audience. The library should analyze and document the use requirements of the repository's designated community or communities, so that it understands the community's technical skills, information needs, and support resources. The user community should be able to access and understand the information in the library's discovery and delivery tools without expert assistance. That is, the preserved information should be "independently understandable." Preserved information must be available to the designated communities, so there must be systems for discovering resources, mechanisms for authenticating digital materials, access control mechanisms in accordance with licenses and laws, an access rights watch, mechanisms for managing electronic commerce (if necessary), and user support programs that are appropriate to the designed community.

The creator community is in part a subsection of the user community. The library's policy should take into account the varying expectations and needs from the creator community, since these projects will put different demands on the digital archives.

• issue: technical metadata

As technology and the user community's knowledge base changes, the library's technical metadata should reflect user skills and expectations. I.e., the AIPs and SIPs must lead to DIPs that are useful for current and future users. Collected projects must contain well-maintained and documented technical metadata that is aligned with the designated community's knowledge base and with changing technologies.

• issue: changing forms of access

As the user community expands and gains new skills, it will place new demands on the library. Delivery mechanisms should provide continuous access via current methods of access.

• issue: user tools

There is a real need for authoring tools for building scholarly digital resources. Some tools may need to be developed with an eye to a particular community's needs, if it relies on a specialized set of resources and research activities. Tool developers will need profiles for subject areas in order to understand what resources are related to a given community.

Policy guidelines should consider a methodology for creating profiles of subject communities, analyzing their research activities, and analyzing scholarly digital resources to relate them to those communities. The library should designate staff to manage the development process for user tools. Responsibilities should include compiling a list of recommended or necessary tools; managing the purchase, development, and implementation of tools; and creating programmatic relationships between tools, the library's collections, and the digital library infrastructure. If the library is providing tools of any kind to scholarly digital resource developers, it must provide adequate documentation and technical support.

It can be expensive and technically complex to develop and implement such tools, so the library should carefully consider whether or not it is a worthwhile investment. Library policy might note that the development of such tools is desirable and that where economically feasible and justified by widespread need, the library will explore the feasibility of developing or assisting in the development of tools.

e. Electronic commerce

If the library wants to sell access to parts of the archive or market internally-generated projects, there must be clear policy statements on how this will be handled; who will be responsible for it; how fees will be determined, collected, and used; how to differentiate paid access from free access; etc. This is a very complex issue, and is further complicated by questions of re-use, ownership and authorship, potential remuneration for authors, copyright, and so forth.

7. Concluding remarks

As part of the process of developing this draft, we solicited comments regarding the selection and collection process from the authors and library staff. Among the notes that came back were the following observations:

- There should be a contract between the author/depositor and the library. This might be a boilerplate contract explaining the terms of collection and preservation, covering such points as copyright permissions, user access, editions, and versions, and working copies of the project.
- Define what makes a collectible project. This is a tricky question, since it may actually be easier to recognize what makes a project uncollectible. At the risk of sounding glib, a collectible project is perhaps a project that is designed to be collectible.
- Clarify copyright questions. If a project only has temporary copyright permissions for its resources, it must be able to identify which permissions will expire and when. The library must then be able to assign appropriate levels of access control.
- Dealing with multiple distributions of a project: commercial, library, and in-development. An author may want to maintain separate copies of a project for different audiences. For example, a project may be distributed by a commercial publisher as well as maintain a separate site for on-going development work.
- Development of local best practices and the importance of communication between the library and the author, even before the project begins.

The technical and administrative problems we encountered are formidable. Given the ephemeral nature of some of the technologies used in digital scholarship and the varied levels of technical skill in the authors, much of digital scholarship currently being developed may prove impossible to collect and preserve without substantial administrative and technical support and a great deal of patience from all parties.

In order to further both our own and other institution's collection experiments, there is a real need for tools that are designed to work in the context of digital scholarship and with the goals we have identified in this report. Without proper tools, accompanied by education and cooperation between communities, the many opportunities and possibilities of digital scholarship may prove difficult to fulfill.

In general, our work on the prototype collections and our investigation of policy issues have resulted in the extremely valuable experience we documented in the SDS annual reports. We remain enthusiastic about the challenges and opportunities encountered throughout the Supporting Digital Scholarship project.

Appendix 1: Committee Members

Policy Committee

Melinda Baumann, Library: Digital Library Production Services (2002-2003)
George Crafts, Library: Humanities Services
Bradley Daigle, Library: Special Collections (2003)
John Dobbins, Professor, Department of Art
Edward Gaynor, Library: Special Collections (2000-2002)
Leslie Johnston, Library: Digital Access Services (2002-2003)
Sandy Kerbel, Library: Science and Engineering Library (2000-2001)
Phil McEldowney, Library: Social Sciences Services
Daniel Pitti, Institute for Advanced Technology in the Humanities
Joan Ruelle, Library: Science and Engineering Library (2000-2002)
Thornton Staples, Library: Digital Library Research Group

Technical Committee

Rob Cordaro, Library: Digital Library Research Group
Kirk Hastings, IATH
Chris Jessee, IATH
Worthy Martin, IATH/Computer Science
Daniel Pitti, IATH Steve Ramsay, IATH
Perry Roland, Library: Digital Library Research Group
Thornton Staples, Library: Digital Library Research Group
John Unsworth, IATH/English Dept.
Ross Wayland, Library: Digital Library Research Group

Appendix 2: SDS Author Questionnaire

This draft was written in cooperation with the test case authors and UVA library selectors.

1. Basic project information

    a. Project full title

    b. Author name(s)

    c. Please give a brief summary of the project.

    d. What are the major parts of project (e.g., image catalog, database, bibliography)?

2. Explanation of project

    a. What files types and standards does the project use?

    b. Does it have a search engine?

    c. Are there any special requirements that the library would need to know about in order to disseminate this project, such as proprietary software that must be purchased or downloaded, specific browsers requirements, etc?

    d. Are there any parts of this project that are not available elsewhere (such as essays that are not available in other media)?

    e. Are there any parts that are published or distributed elsewhere?

    f. Can you provide or generate the minimum technical and descriptive information for all of your resources? (NOTE: if this question stays, it'd probably be helpful to attach a list of what the minimum metadata is)

    g. Do you want the library to collect all or only part of the project?

    h. Is the project finished or will there be future editions? If the work is still in progress, is there a projected completion date?

3. Checking for missing information

    a. Do you have copyright permissions for all resources used by the project?

    b. Can you give a provenance for all objects in the project (e.g., date of creation, who created it, and any appropriate description)?

    c. Will you be able to obtain any necessary permissions from resource

owners?

4. Specifics about dissemination

   a. Is it acceptable to allow access to individual resources or parts of the project via outside links (such as the library's image catalog)?

   b. Does the library need to restrict access to any elements in the work? This may be required by your copyright agreements.

Appendix 3: Tools

As part of SDS's work, we have worked towards developing tools for aiding authors, users, and repositories. This work is well worth pursuing, but in most cases we were unable to do more than take the first steps towards building useful and reliable applications. The WebCollector and GMDS Tool were discussed in earlier reports.

WebCollector

The Java source and class files for the WebCollector application are available on-line at < http://icarus.lib.virginia.edu/software/webcollector/>.

GDMSTool: GDMS XML Graphic Editor

GDMSTool is a graphic XML editor tailored specifically for the General Descriptive Modeling Scheme (GDMS) DTD. It is currently available in for Windows 2000 (available at http://icarus.lib.virginia.edu/software/gdmstool/gdmstool.zip) and Mac OS X (available at http://icarus.lib.virginia.edu/software/gdmstool/GDMSTool.sit).

DOCA

The digital object collection application (DOCA) is a tool for allowing users to create links to resources in a collected work (i.e., create a bookmark to a web page). The application we worked on was intended to allow people from the UVa community to create their own collection of references to digital objects from the UVa library's Central Repository. The objective was to develop effective techniques for referencing digital objects from the Central Repository. The test version was not finished, due to lack of time, but the diagram below shows the high-level view of the tools' functionality.

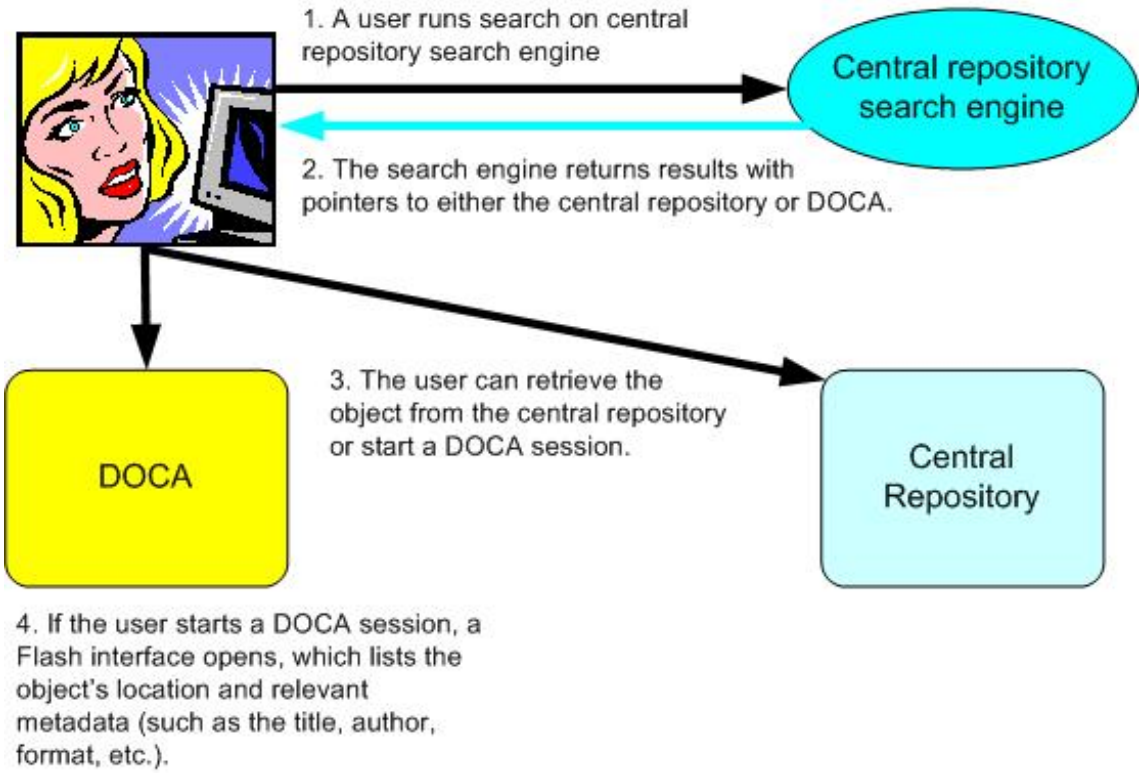

Figure . DOCA

The Java source and classes for the test version are available on-line (http://icarus.lib.virginia.edu/software/doca/).

Appendix 4: Summary of previous work

All SDS annual reports are posted on the IATH web site, at <http://jefferson.village.virginia.edu/sds/>, and are available to the public.

The list below shows the tools and projects that were under development during each year of the project. The artifacts listed are all included in the annual reports.

2000

Tools: FEDORA, WebCollector, Granby
Projects: Salisbury

2001

Tools: FEDORA, GDMS, WebCollector, Granby
Projects: Salisbury, Rossetti, Pompeii
Artifacts: Content models for Salisbury and Rossetti, GDMS DTD, XSLT stylesheets for Salisbury and Rossetti

2002

Tools: METS, GDMS Editor
Projects: Salisbury, Pompeii, Tibet, Whitman, Rossetti, Blake
Artifacts: GDMS code for Salisbury and Pompeii, Tamino report