

Digital Archives

© 1995 Kenneth J. Varnum
All Rights Reserved--No Reproduction with Authorization
varnumk@omri.cz

Introduction

The creation of a digital library involves the solution of numerous problems and issues, from designing computer programs to handle arbitrarily large numbers of users and databases to designing intellectual models on which to base the digital library design to filling the library with resources. The whole process is of course highly recursive, with answers to one set of issues raising new questions elsewhere. This essay will explore the latter two points, leaving issue of computer program design as one more easily resolved with a clearer conceptualization of what is to be designed and how it will be used. The first section of this paper will define the digital library so as to allow further discussion. The second section will introduce a model for conceptualizing a digital library and the final section will introduce the topic of resource development within the context created in the preceding two sections.

I. What is a Digital Library: Toward a Working Definition

The term "digital library" is widely used in such a variety of contexts and has been endowed with so many connotations that to discuss the underlying concept without first exploring it is to invite misunderstanding. Thus, before beginning to describe criteria for discussing digital library models or specific conceptualizations of one, the term should be described and refined to ensure a clear understanding of what is being discussed. This description will be general since the concept is evolving in a number of different directions simultaneously, and no one model has come to the forefront. Given this situation, this will be a working definition for this paper, not necessarily for application elsewhere.

A. What a digital library should look like

A fundamental premise of this essay is that there will be no single "digital library." No one library will be perfectly applicable to all situations or kinds of materials, let alone be the repository of all knowledge. Thinking that one digital library will fill all needs is as unreasonable as stating that there is just one kind of library--that there is one organizational scheme, one set of materials, one fixed range of subjects that all libraries should have. This statement is patently false: traditional libraries use a number of cataloging systems, the Dewey Decimal and Library of Congress systems being the most common in the United States, but by no means universal. Certain libraries specialize in specific types of resources (periodicals, microform publications, handwritten manuscripts, etc.) or certain topical areas (such as the Folger Shakespeare Library, for example). Still others generalize, the Library of Congress and the New York Public Library being but two examples.

Many heralds of the digital future speak with hushed tones about the wonders of decentralization and how it will change the face of information retrieval and access. These people have, perhaps, become so caught up in the future that the present has disappeared. One should remember that information is already highly distributed. In fact, the University of Michigan Digital Library initiative, for one, recognizes this certainty and is developing a set of standards and protocols rather than a specific model. As both Bill Birmingham and Mike Wellman described¹, and as is logical given the capabilities of decentralized computing, there will not be a digital library, but rather digital libraries, interconnected but independent. While their connections may not be "virtual" T1 lines, they rely on a similar arrangement, Interlibrary Loan (ILL). The true advantage of the digital library will not be decentralization or interconnectedness, but the speed at which a user can access information from whichever digital library within the overall infrastructure that possesses it.

So, digital libraries will be networked and interconnected, and almost certainly interdependent. If we can isolate a single digital library from the broader network, what will it look like? It should have whatever set of resources are of most use to its immediate clientele. And it must be

¹ Lectures to ILS 605, October 5, 1994 and November 30, 1994, respectively.

flexible in terms of how it handles data. Since the way information is displayed to the user (the way the "books" are arranged on the shelves) can be separated from the organizational schemes used by the library itself, the organizational scheme that most benefits a set of users should be employed for that set. This flexibility is one of the digital library's main selling points. It can be many things to many people because it can be customized. Thus, it should be able to attract any user who might currently not use the traditional library because the organizational scheme is perceived to be incomprehensible.

As evidenced by the foregoing, a digital library can be defined broadly or narrowly. As a working definition, then, a digital library is a collection of electronic resources which can be searched from a common access point sufficiently seamlessly that the user neither notices nor cares where one collection leaves off and the next begins. In this definition, the digital library is not even a new creation--we already have and use them, although they are not so labeled. Dialog, the well-known collection of on-line databases, is a digital library, as is Lexis-Nexis. User interfaces for these databases are not as developed as they might be, and require a great deal of training and experience for successful use. They are, nonetheless, digital libraries, which we have come to define here as a large number of disparate information resources available "under one roof," as it were, and although the vocabulary and syntax needed to search one file may be different from that needed for another, the process is essentially the same. The future digital library, though, will undoubtedly be much larger and more complex, should adapt to the needs and expectations of the user, and should make irrelevant such hindrances as language and location of a particular item.

B. Criteria for Judging Implementation of Model

Such a loose definition of the digital library needs amplification. In this section, a set of three basic criteria for measuring the utility of a digital library will be outlined and discussed. A digital library implementation must be: **scaleable**, **useful** and--most important--**useable**.

1. Scaleable

One of the largest assumptions of the digital library project is that digital materials will be created or found to fill the virtual shelves of the library. In the current, testbed, phase of digital

library development, the assumption that sufficient materials can be found to fill the single shelf being used is not far from the mark. Such testbeds are essential to ensure that the architecture created for the digital library is robust enough not to come down with a crash when subjected to "real" use. The word "real" is in quotation marks here because, just as the digital library itself is an amorphous and slippery concept, so are ideas of how a such library will be used, even initially. Regardless of whether a digital library testbed is endowed with more resources or subjected to more users, it will still be a testbed, which will not function the same way as a real-world working model.

Any digital library design must, therefore, be able to expand to be able to handle many more users and provide access to many more resources than even seems likely to be extant. One lesson (and possibly even a rule) of the information revolution of the last few years is that use grows far faster than expectations. To cite one brief example, America On Line, an Internet access provider, had to stop advertising for new customers in mid-1994 until it could expand its computing resources and rewrite its source code to handle the use loads it unexpectedly achieved, and announced plans at the end of 1994 to double its subscription base (to three million people) by the end of 1995.²

While such rapid rates of growth can not continue indefinitely, it is reasonable to expect the number of service providers and users to increase significantly over the foreseeable future. Any system must be expandable beyond what seems reasonable at the time it is created.

2. Useful

The idea of usefulness is so subjective that I do not believe there are concrete measures by which a system's usefulness can be judged effectively until it is up and running. At that point (or during testing, when real users work with the system), an estimate of user satisfaction can be made. How this should be measured is an open question. Much research in such relatively well-understood areas as on-line database retrieval focuses on relevance of materials obtained through a searching tool. But the problem of relevance is relativity: who is the judge of relevance? Even if a test database is constructed, with articles carefully selected to fit into one of a small number of categories, who is to say that if a

² The New York Times, National Edition, 6 December 1994, p. C6.

particular user's search obtains an item outside the assumed category that it is not relevant? It was selected, for better or worse; there may be some aspect of it that matches the searcher's needs. One solution to the relevance morass has been suggested in the University of Michigan's NSF Digital Library Testbed (UMDL) grant proposal, which recognizes the failure of traditional means of measuring usefulness of materials culled from a database, and by extension, of the database itself. The UMDL will evaluate the database according to another, very subjective, measurement--the value of the information obtained to the user who obtained it.³ The substitution of "value" for "relevance" does not do much to clear up the vagueness of the original concept, but it does peg an easily comparable, albeit subjective, measure on a piece of information. Perhaps the only way to judge the usefulness of a digital library implementation is to allow it to compete with others, all of which provide access to the same information resources, and to see which one is used more. A less wasteful alternative might be that when a digital library is perceived to be better to use than the traditional library for a given purpose, the usefulness criteria will have been met.

3. Usable

Usability is even harder to define than usefulness, and is perhaps even more subjective a measure. Several indications of usability are the following:

Can users find information on a topic? If the system allows users to find information on a topic of interest to them, then the system has passed the most important test of usability. Obviously, a digital library will not contain information on every topic immediately. There will be a prolonged development stage before even an approximation of "all knowledge" is available. In the short run, though, if a digital library can provide the bulk of its users with information resources they need, it will be useable.

Is the information provided by the digital library at an appropriate level to the user? A useable digital library will have information in a variety of formats and for a variety of purposes--from cursory and introductory to thorough and intensive. Not all users will want highly detailed,

³ UMDL Proposal, p. 44.

footnoted, and researched information; for many purposes, an overview will do. The age and education level of the individual user must also be taken into account. A grade school student will not be able to use a graduate-level explanation of how and aircraft flies, just as a Ph.D. candidate will have little use for a high-school text on political party theory.

Is the access interface sufficiently easy to use that it can be employed by people at different educational levels and/or ages? No matter how efficient or effective the program that matches users with resources, results will be useless if the user cannot effectively instruct the computer what he wants. The interface must be intuitive, and must be flexible enough that more advanced users, who better know how to use it, can access more advanced functions. A WWW-browsing tool like Netscape or Lynx might be an appropriate model for the interface for individuals with little knowledge of how a library is organized--much as inexperienced library users go right for the card catalog or on-line public-access catalog without first looking through a thesaurus of subject terms LCSH [Library of Congress Subject Headings], for example). For more advanced users, a less scripted interface would be appropriate.

A wonderful metaphor for this, coined by Yuri Rubinsky, is that, much like Disneyland, a digital library must keep the technology (the "magic", if you will) hidden. In Disneyland, the magic is the tunnels beneath the entire park--the same tunnels are beneath Space Mountain as Mr. Toad's Wild Ride. In the digital library, the tunnels become the programming--completely transparent to the user. There is, in Mr. Rubinsky's phrase, no "difference between 'asking a question' and 'doing research'⁴." Usability varies dramatically depending on the specific user; what works for one group of users will not necessarily work for others. The system must therefore be able to communicate at a variety of levels.

C. Conclusions

The three criteria outlined above should not be construed as the final word on digital library evaluation; other measures can and should also be used. For example, economic measures are likely to be a key factor for many libraries or archives developing a digital element. However, these are not

⁴ Yuri Rubinsky, *Electronic Texts the Day after Tomorrow*, p. 12.

discussed because, while they are important, they should not be the driving force by which a digital library should be judged. While the cost of the technology and development that go into digital libraries can be added quite readily, this is not a criterion that is likely to have a lasting effect; technology costs have been dropping steadily with no end in sight; the computing power needed to search vast databases will drop. In an article which advocates the quickest possible development and implementation of the digital library, Brian Hawkins of Brown University writes,

The electronic library is specifically both a solution to the economic problems facing libraries and a vehicle for a new functionality that promises to transform scholarship and bring the cultural, social, and economic benefits of information to many.⁵

It is this dream that underpins much of the enthusiasm for the digital library and is guiding some of the initial popular enthusiasm for it. At the same time, people interested in the digital library should think carefully about what a digital library should be, and be careful to create a system that meets the criteria outlined above, and others, and does not provide features that are wonders of programming but do not serve a particularly useful or useable function. A specific conceptual model is needed to avoid this danger. This is the subject of the next section.

II. Archives as a Model for the Digital Future

The name "digital library", in many ways, makes a great deal of sense in the direct interpretation of the phrase. At the same time, though, such a literal interpretation might mislead both builders and users of the digital library. In this term, the word "library" receives most of the emphasis. Perhaps the phrase is popular because it embodies the familiar, the neighborhood library, filled with books. While the library might be the most visible or user-friendly real-world model to pin to the developing virtual equivalent, there is another model, not quite as prominent in the public eye, that has much to offer the digital library. Like the library, this entity has centuries of tradition to draw on. Unlike the library, it is less confined by concepts of organization and access.

⁵ Brian Hawkins, "Creating the Library of the Future: Incrementalism Won't Get Us There", New Scholarship: New Serials, 1994.

The model alluded to is the archive. Archives have a great deal to offer the digital library for several reasons. First, archives already deal effectively with collective bodies of information, and not with individual items. Where the traditional library focuses on acquiring specific (albeit many) books, periodicals, etc., , archives collect information in a less distilled form--original papers, files, and documents by the thousand. The archival approach to collective description and arrangement seems very appropriate to the digital environment simply because it is impractical to catalog everything currently available in electronic format at the item level. Archives deal with collective description.

Second, archives approach resource sharing much differently than do libraries. While libraries tend to acquire as many different books within their field of expertise as possible to ensure that any needed resource is available on-site without the need to resort to interlibrary loan, in the archival understanding, collections are unique (though it is in most cases possible to copy collections onto photocopy or microfilm, duplication is seldom undertaken except to make copies of particularly used or valuable collections for in-house use) and are shared. "Although university libraries have practiced resource sharing for years, for many the preferable option nonetheless remains local ownership of as much of the universe of published scholarly material as resources permit. Interlibrary loan services are thought to be inefficient; the lending library, understandably, attends to its own readers' needs before addressing those of readers elsewhere. The difficult economic circumstances in which research libraries currently find themselves argue for new models, and electronic information technologies seem to hold particular promise."⁶ Much the same holds true for archives--but they already share detailed meta-information and stand to gain a great deal from the electronic sharing of the resources themselves. Either resources must be shared or the researcher must travel. Archives already share detailed holding information, often including complete finding aids, over RILIN, a nationwide computer database of archival holdings.

⁶ Anthony M. Cummings, Marcia L. Witte, et. al., "University Libraries and Scholarly Communication: A Study Prepared for The Andrew W. Mellon Foundation", The Association of Research Libraries, November 1992, (URL <http://www.lib.virginia.edu/mellon/ch10.html>).

A. What is an archives?

Unlike most libraries, archives already think in terms of collections of information, not of individual items. Archivists see the world in terms of **documents, files, series, record groups** and **archives**. I will provide a brief definition of each of these in both the traditional archival and digital library environments. A **document** is a single item, the building block of a collection. It could be a two-sentence letter, a thousand-page manuscript, a video cassette or any other tangible item. In the digital library, a document is similarly any one item-- as simple as an electronic mail message or as complex as a relational database.

The next more general level in this hierarchy is the **file**, which contains at least one document. If this is a file of papers, then there could be hundreds of pages; if videocassettes, then a handful), and are described by topic or genre of contents ("computer purchase" or "correspondence", for example) and the dates spanned by the material. Archivists rarely describe the contents of a file with more than a topical or genre description, and then only when the specific contents are of special significance historical or monetary value.

In the digital library, the archival file has a broader definition, as a collection of not necessarily similar documents. Two understandings of "file" with this definition can be grasped: literally, as the contents of a subdirectory in some storage device, or more figuratively as a compound document. The latter definition needs a bit of explanation. A compound document is one which is comprised of several independent parts that are brought together only upon display--the individual pieces of it are stored separately and are likely the responsibility of different individuals. This, then, is how an information resource in the digital environment should be understood: as an individual file.

Series are groups of files on a related topic or of a common genre. They are gathered and described collectively because archivists believe that an organization or entity can be understood only through examining its collected output. (Sometimes there are subseries within series, but this is more for ease of use than ease of description.) In the digital library, a series is the highest level of the hierarchy--a collection of resources stored together--or, in simpler terms, a self-contained digital library.

The upper two levels in the archival world are related: the **record group**, which is comprised of all the series from a single person or office, and the **archives** itself, which is the collection of record groups. A records group could be as small as a few inches of paper thick or as large as several thousand linear feet of materials. The archives, the collection of all record groups under one roof (administrative or literal) can likewise be large or small. The digital equivalents of these two terms can be seen as a specific digital library and of an overarching network of digital libraries, respectively.

B. The importance of context

I have mentioned the importance of context from the archival perspective several times in the preceding section. As Ann Gilliland-Swetland noted,⁷ context is the archivist's meat and potatoes: knowledge does not exist in the abstract. While this organizational principle might seem somewhat strange or inefficient at first blush, it must be remembered that archives contain millions of individual documents, most of which have meaning only when viewed in conjunction with other documents so that a coherent understanding of the collection can be developed. There are few individual documents that explain everything about their creator. To understand one fact or piece of data, one must understand the process by which and the reason for which that piece of data was generated. Archives therefore describe information collectively in **finding aids**, written by the archivist who processes the collection. A finding aid describes in words the origins and organizing principles, and strengths and weaknesses of the collection from the researcher's point of view without, generally speaking, explicitly listing the contents of the collection item-by-item (exceptionally interesting documents aside). Collections and their finding aids are in turn indexed, usually in a card catalog, by subject, name and other access points. Keywords are assigned to aid in cross-referencing collections. Researchers generally start with the card catalog to find the appropriate collection to examine. While these layers of cataloging are time-consuming, they ensure that collections may be accessed from a number of points without having to create a surrogate for each item within the collection. In the digital library, however, such cataloging becomes much more cost-effective and can be automated, especially if the materials in the collection

⁷ Talk at the University of Michigan's School of Information and Library Studies on 4 November 1994

itself are already in digital form. Thus, the "digital library" is already highly "archival" in its functioning.

Modern archives have only just recently begun to obtain (**accession**, in archival parlance) electronic information in significant quantities and have therefore only recently begun to grapple with electronic information. This is so for two primary reasons. First, archival accessions lag behind the present by years, often decades, since documents are only transferred to an archives after their immediate relevance has faded, and even then often not until the storage space was needed for some other purpose. Second, even in the age of the electronic office, paper is still viewed as the reliable storage medium and remains a far more common office output than any digital product.

Even so, the mix of paper and electronic information will gradually move toward the digital as offices come to rely increasingly on digital means of creating documents, and particularly on electronic storage media. As more material is in electronic format, archives will be faced with greater quantities of electronic information. The electronic office being a relatively new development, the electronic portion of information accessioned has been small, but is growing.

To meet the rising tide of electronic accessions, archives, just as digital libraries, need to resolve appraisal, preservation and access issue related to electronic records. Many of the functions that have been documented by archives in the past--communication, decision making, interactions with other entities, recording thoughts and plans, and so forth--are now electronically based. The trend is towards greater automation. When new accessions are made from offices, it is now normal, if not yet the norm, for that accession to have digital components. Archivists need to decide how to treat the digital component of accessions (which might consist of files stored on someone's hard drive, a department's materials for a multi-media report to the board, or a multi-level spreadsheet with complex formulae and annotations, to name just a few examples). Additionally, hypertext documents (such as those on the World Wide Web) present their own problems to this study--what to do when all that is accessioned is essentially a list of pointers to other sites, perhaps with commentary from the person who created the list, perhaps not. While this list could well explain a person's or organization's interests and

activities, without the items pointed to there is no context by which to understand the collection as a whole.⁸

C. Digital libraries

In this section, I will examine in more detail how archivists see electronic records and explore potential solutions to the problems they present within the context of the digital library. A satisfactory solution must take advantage not only of the strengths of electronic documents themselves, but also ensure that users can get access to them. Just as in traditional (i.e., paper-based) archives, storage and preservation solutions are not acceptable if they so limit access by researchers that the items might just as well not exist.

A digital library is one that preserves, organizes and makes available electronic records. Such records are those that are either created or stored in digital format on a computer. They may be hand-entered, scanned, or created by a computer from individual databases or documents according to some automatic process. Electronic records differ from paper records in that they do not need to have a physical form to be considered records. Rather, they can go through the entire record life cycle (creation, active use, temporary storage, and final disposition) in a purely electronic world and never be fixed on paper or any other physical output. Digital documents can be the compound product of two or more people or the result of a long period of revision and rewriting. As such, the familiar concept of a first, second, or final draft loses some of its meaning. This essay, for example, has been written and re-worked in several stages over six months and bears little resemblance to the "first draft"--which itself never saw paper. There is essentially no trace or evidence of the revision process between the first version and the last because there is no method in use for capturing intermediate stages of the process. While intermediate drafts may never be intended for publication, with traditional writing techniques (typewritten or handwritten drafts) intermediate stages are often extant, allowing later researchers to trace the author's thought process and work patterns.

⁸ For a discussion of this topic, please see Caryn Stein and Weston Thompson "Using Electronic Manuscripts to Document Student Life: An Introduction for Archivists", (URL http://www.sils.umich.edu/~wthackt/607project_homepage.html).

Furthermore, electronic documents have no "original," at least not one that can be readily determined. Each electronic copy of this essay, for example, the one stored on a floppy disk as a backup, is in every sense as authentic as the one on the hard drive that the author is actively working on. But if work were done on the backup copy, at another site, and brought back to the original computer, the backup version becomes the new original and replaces the old. Copies of copies could be made, and still no degradation would occur. As hardy as electronic documents are, they are also curiously fragile. Once a file is erased, or if its data somehow become corrupted, it is often irretrievably gone (as many a student can attest), as surely as if the paper version of a document had been burned.

Archivists are given far more materials of all kinds than they have room or need to keep; their decision is often (unlike the library) not what to acquire but what not to preserve. The process of deciding what to keep is called **appraisal**. While archives have more or less resolved the issue of what to keep and what to discard for paper documents with a set of appraisal guidelines, such guidelines are lacking for electronic materials. Nonetheless, a pair of opposing tendencies can be discerned.

At one pole is the idea that digital storage will allow absolutely everything to be preserved for all time. While disks and other media may provide compact storage of large quantities of information, they are more susceptible to office hazards than paper--beyond fire and water, the two main threats to traditional files, disks can be damaged by heat, dust, magnetic fields, mishandling, and so on. Additionally, the prospect of creating a comprehensive and usable index for so vast an amount of information is a daunting task that current theory and practice are not able to meet. Finally, it ignores the problems of constantly advancing and changing technology.

At the other pole is avoidance of digital documents and electronic storage at all costs. If we do not know how to handle digital documents, why not just convert them into a form we can handle? When the digital file (a complex, multi-part document--refer to the definitions, above) is printed, electronic enhancements are lost. The more complex the file, the more noticeable the loss because links among and between the files--even if they are recorded in text, the ability to move from one to the other, or to look up and perform calculations on data from different documents maintained in separate corporate offices

and to summarize them all on the fly--will be lost. Although each component of this file is an individual document, from the archival standpoint the collection of documents is the important thing. Each individual section, without the others, loses meaning because it has been isolated from its context. Although individual items may have some intrinsic value, in the absence of the whole, information is lost. There is simply no two-dimensional way to represent such information accurately.

A second objection to avoiding electronic documents is that, despite our advances in storage and retrieval technology, the volume of material to be maintained is huge and mostly unnecessary. Furthermore, this option is not even applicable to complex, multi-media documents which often include, or have links to other, non-text (image, sound, motion picture, or mixtures thereof), files. Archivists, much like digital librarians, are caught in a bind: it is not technologically possible to maintain everything in its electronic form, but it is also not possible to print it all out and store it with traditional means.

Archivists' desire to appraise and preserve electronic data in its original form is not the only force in action. Just as donors are creating information in electronic formats and asking archives to preserve it, so are users coming to expect electronic information resources and automated access to all records in archives. The advantages of information in electronic form are partially in capturing data that had been previously lost, and partially in the ease of manipulating digital information.

D. What gets preserved

It should be clear that not all document-creating entities warrant preservation of their entire documentary history, or even necessarily of "snapshots" of significant stages along they way. In the middle ground, many archivists are beginning to view the digital archives as a way to streamline the accessioning and preservation stages of archival activity into one smooth process. This concept has two facets: preservation of the information, and preservation of the context in which that information is considered worth preserving. The first facet is simpler to handle, and if that is all that is needed, can be accomplished by printing the documents (with the necessary loss of context). The second facet is, and perhaps the trickiest challenge facing both archives and digital libraries, is that neither has as yet

decided how to handle complex digital documents. As mentioned above, such documents have just slowly begun trickling into archives, so there has been no impetus to deal with them. Likewise, digital libraries are just beginning to grapple with the problems of rendering and describing such documents.

The crux of this problem is that office-generated electronic documents can be far more complex than either their paper (which have only a flat, two-dimensional existence) or electronic publishing (which are designed together in a conscious way) cousins. Electronic documents share the worst features of both types, as far as archivists are concerned. And the fact that a document's appearance and content are almost completely independent of each other has opened up another field for debate. Noted electronic records scholar Charles Dollar comments that, "It is the logical structure of documents that makes them intelligible to humans. The physical relations of a document largely involve style and rendition.... Unlike paper documents, where logical and physical relations are inseparably linked together, the logical and physical relations of electronic documents are in fact separated and stored independently of one another".⁹ WWW documents, for example, are created as flat, unformatted text and are tagged with formatting instructions. A browser can be set to display any kind of tagged text in any manner. The task of formatting information has passed from the information creator and to the information user. This may be viewed as a wondrous advance in the digital library community, but not by archivists.

E. How to describe materials

Archivists traditionally first determine whether a set of documents is worth preserving, and then describe them according to provenance and then according to content. Assuming that an archivist has both the informational (the data itself) and technological (the computer and software that processed the data originally) components of an electronic record series, how does he go about analyzing it? According to David Bearman, the selection criteria for records in general is "evidential historicity," which he defines as

⁹Charles Dollar, Archival Theory and Information Technologies: The Impact of Information Technologies on Archival Principles and Methods, Oddo Bucci, ed., (Macerata, Italy: University of Macerata, 1992),p. 36.

the sum of all information that can be determined about an accountable transaction, which is defined as the relationship between a record and an activity determined by archivists to require evidence. The information which determines evidential historicity is derived from analyzing the data, the structure, and the context of records, each of which testifies explicitly and implicitly.¹⁰

Thus, determining the provenance¹¹ of electronic records is a process similar to that of determining provenance of paper records--through understanding the documents themselves, the way they are structured, and the office or individual that created them. Much as the potential richness of a digital document far exceeds that of its paper counterpart, such documents' provenance likewise can appear much more complicated and detailed. However, this complexity is misleading; figuring out the creating agency for a hypertext annual report within a large corporation--a report which likely brings together spreadsheets with complex notes and formulae from the financial department, audio and video messages from the corporation's directors, and more images of products from the sales and production divisions--looks complicated, but no more so than for a simpler, paper equivalent. Some specific office is still responsible for collecting and organizing the information; it might be the computer department in our example here, not public affairs, but there is still a function that can be described and used as an organizational tool.

Content description is a relatively simple process that involves examining the materials and describing them generally at the file level (as noted earlier). Describing specific methods of building digital libraries is beyond the scope of this paper; suffice it to say that description of electronic documents poses at least as significant a problem for the digital library as for the traditional one. The task of matching specific single documents, which themselves frequently have little meaning taken individually, to a specific research question, is a task daunting for the most trained reference archivist. The best solution, it would seem, would be to describe appropriate collections of documents in terms of

¹⁰David Bearman, "Archival Principles and the Electronic Office", in Angelika Menne-Haritz, ed., Information Handling in Offices and Archives, New York: K.G. Saur, 1993, p. 178.

¹¹"Provenance" is an archival term referring to the origins of a document. It is a slippery concept, but in broad terms it refers to the purposes for which the document was create. In this way, the function of the person or office that creates a document is more important than the individual creator.

function, role, topic, actor and recipient of actions. There will still need to be a sifting and sorting process involving the end user; to think otherwise is unrealistic.

F. Conclusions

Archives and digital libraries have a great deal to offer each other. Archives have experience in organization and description which I think the digital library should draw upon. Collective arrangement makes a great deal of sense when information will be widely dispersed among documents; once the appropriate collection is located by the researcher, then a search engine might be used effectively to find the requisite subset of documents.

If the digital environment brings the ability to self-publish to everyone, or even to many, some kind of selection criteria will need to be invoked. While the caveat that "not everything needs to be preserved" may soon no longer be necessary for reasons of computer memory limitations, it will always be necessary for more subjective reasons--most of what will be composed in digital form will be unimportant or trivial and will not serve any purpose if preserved. Determining what documents find their way to the digital library should also be informed by principles of archival appraisal--that most collections contain a great deal of chaff for every kernel of grain. The next section continues with the theme of selecting from among the many resources that will inevitably be available.

Finally, the idea that archives, as broad collections of information, need to have a context to be intelligible also transfers to the digital library. A screen display with the answer to a question is next to meaningless if there is no sense of the process by which that answer was determined. The library might be the more common metaphor for digital information collections, but the archives is more appropriate. And in the end, it seems to me, the digital library, like the archives, is nothing without context. If there are no relationships among the many information resources available through the digital library, there might as well not be a digital library.

III. Virtual Resources for the Virtual Library: Collection Development

The final section of this essay will look at the importance of a collection development policy (CDP) to the digital library. A CDP bridges the gap between the theory of what a digital library should be and what they are in practice; it is the map that leads the way to the goal. Digital libraries, like archives, will exist in a world of information overload--there will be too much information available to include it all in every, or in some case in any, specific library. CDPs in the digital library will therefore not be as concerned with acquiring resources (their focus in traditional libraries), but with separating the wheat from the chaff. It seems certain that the digital library's shelves will be filled; quality, not quantity, is the issue.

A. Comparing Collection Development Policies: Archives and Digital Libraries

A CDP, briefly, is a statement of focus (what the library will and will not collect), clientele (who they are, their educational level and general interests), and an outline of the decision-making process (who decides the specifics of what resources should be acquired).

1. Focus

The start of a CDP is an explicit statement of the archive's focus. Since an archive generally endeavors to document a specific facet or range of human activities and must carefully direct limited resources to attain that goal, archival CDPs often include explicit statements of subjects that are not of local interest. While the digital future may bring vast monetary savings to libraries, it is unreasonable to expect the present largesse being bestowed on digital library projects to continue indefinitely. Another important feature of a collection development policy is an explicit statement that resources will not be selected or excluded because of the political, religious or other views expressed within them. Just as librarians and archivists have traditionally upheld the principle of equal access to information, so must the digital librarian.

2. Clientele

A second section characteristic of a well-developed CDP is an explicit acknowledgment of the intellectual interests and abilities of the people who will be using the library's resources. This is a larger concern in the present for libraries than archives because archives are generally consulted by experts who come to the archives specifically for resources available there; libraries tend to draw a larger cross-section of the local population who often do not know what the library has to offer. Digital libraries will likely be faced with clienteles more like that of the present-day library. Budgets aside, academic research and town public libraries have different sets of users for whom they obtain different kinds of resources. In a research library, grade school texts would generally not be part of the general collection, just as highly technical scientific works would generally not be found in a small public library. A library should select materials that will both be of use and be accessible to its patrons.

The digital library must make similar decisions, but at a different level. Digital documents can include meta-information on the level of information included, making searches self-limiting. While this process might be hidden from the user, it must be done carefully and accurately. High-school aged users will find the digital library useless without materials accessible to that educational level, while the layperson will find abstruse technical descriptions of chemical reactions unintelligible. It goes without saying, though, that the system must not be designed to purposefully prevent users from finding information at something other than their presumed level.

3. Decision making

A third important section of a CDP is an explicit description of the resource selection process. Are decisions made by committee, by an archivist expert in that field, or by the head archivist? Who resolves disputes within the committee? When there are several resources available on the same topic, what are the guidelines for selecting the best resource for that library?

In the digital library, a less centralized arena, these specific questions may not directly apply. However, a good CDP for a digital library will both encourage long-term planning for eventual

accessions of documents (in the traditional archive sense) as well as be flexible enough to take advantage (and recognize that an advantage exists) of resources which suddenly appear.

B. Developing a Collection Development Policy for the Digital Library

A CDP is similar to a mission statement: it details the reason for collecting a certain range of materials and explains how that process is undertaken. A policy should be relatively constant over time--if it can (or does) change at a whim, the collection will lose focus and holes will appear in it. A well-written policy can be a powerful document in the defense of freedom of information and freedom of access--but it should be written as to ensure flexibility.

In neither digital nor traditional libraries does a collection development policy state explicitly what resources should be purchased (except to the extent that periodically revised reference sources might be mentioned). In the digital library environment, however, such prescriptions are also to be avoided because sources will come and go. At present, WWW and other servers come on and off line with abandon, and often refuse to connect to users if too many are already accessing them. Furthermore, a truly useful resource could be entered into a digital library from one server, only to disappear when the creator moves on. Since the network will be decentralized--resources will be not be stored on one machine--there is very little to prevent the only copy of a resource from vanishing into the void--the equivalent of "out of print" except that, in the digital environment, there would not necessarily be a library which possessed a copy from which it could be "borrowed" or copied. Making a back-up of rare materials would side-step this issue, but that presumes the knowledge that no one else on the distributed network has a copy, either.

The need for a collection development policy in the digital library environment is probably even greater than it is for the traditional library. The digital library is acting in something of a vacuum--there are conventions and expectations about what should be found on the shelves of traditional libraries, but these rules of thumb do not transfer easily or directly, not is there enough experience on the digital library side to have learned new guidelines. Since the mass-use digital library is a rapidly evolving and somewhat novel concept, a collection development policy for it is all

the more urgent. A digital library which includes resources simply because they are available might be doing well in these early, testbed, phases of development, but a "working" digital library so created will be a very poor one indeed. It is assumed that resources will develop at a pace at least as rapid as that of the technologies behind the digital library itself; given the experience of computer technologies in general over the past several decades, this is probably not a poor assumption.

C. Conclusion: What Next?

While the digital library, as presently conceived, is a new idea, in its practical aspects it is less revolutionary than evolutionary. Entirely new modes of thinking are not required; rather, new ways of applying old ideas and interpreting them will be called for. The digital library has many analogies in the present world of libraries and archives. Although the differences between traditional archives and libraries are presently fairly broad, I think they will diminish in the electronic world of information storage and retrieval to which we seem to be heading. The archival world has a great deal of experience in organizing large amounts of seemingly disparate information into a cohesive whole and describing it so that the researcher can pick her way through mountains of information to find the specific groups of items that will answer her question. Collective description and organization is the key to constructing the digital library, at least in its beginning phases. While organizing at this level of granularity will certainly not be sufficient for all time, it must be remembered that these are extremely complex problems; the hierarchical descriptive techniques of archives make it possible to hang more detailed descriptions and catalogs on the existing tree without having to reinvent it from the top down. And there is a logic to building systems that can be focused on ever-decreasing levels of a hierarchy as experience is gained through working at the upper levels.

As strong as the inclination to start fresh may be, I think that would be a mistake. The current world of library and archival science has a great deal to offer the digital library environment. Professionals in the information field do not have expertise in programming, but in helping the end user (who, it must never be forgotten, is not a computer expert, is not a librarian or archivist, and, for that matter, is frequently not knowledgeable about the current information resource, the public library) find

the data he wants. Whatever systems are created and thrown into the marketplace for public use, the ones that are easiest to use and most successful at locating appropriate information resources will be the ones in highest. The successful ones will take advantage of what both the traditional and digital libraries have to offer.