LIBER QUARTERLY, ISSN 1435-5205

K.G. Saur, Munich, printed in Germany

Paper and Digital Repositories in the United States

by DAVID F. KOHL

INTRODUCTION

What you've asked me to talk about today is basically what can libraries do with all the stuff they have and continue to get. Where do we put it all; what do we do with it? As we know, libraries have three core functions: collecting, organizing and preserving the key documents of the human enterprise. And, with apologies to Saint Paul, the greatest of these is preservation. For without preservation neither of the first two ultimately matter. My assignment this morning is to bring you up to date on one specific aspect of the preservation function, library repositories, and indeed, library repository developments in the US. The plan for this morning's presentation is the following: after a brief background review to give us a context for American developments we will examine first the various kinds of print repositories and then the various initiatives for electronic repositories. Because other presentations at this conference deal with electronic repositories, the main focus today will be on U.S. print repositories.

CONTEXT

In today's world, discussing repositories is not quite as straightforward as one might think. In fact, to adequately understand the thinking behind library repositories and hence their development, it is necessary to identify clearly a series of paradoxes, or perhaps dilemmas is a better word, with which the library manager must struggle. For example, librarians collect materials so that they can be used, but the use of materials causes their destruction which prevents their further use. There are other relevant dilemmas. One is requesting more budget for the purchase of new materials when there is not room to store the materials already owned. And a third dilemma: is it enough to save the information or is it also necessary in some sense to save the artifact as well. As fundamental dilemmas we need to keep in mind that it is not possible to resolve them; it is only possible to achieve a careful and insightful balancing of tradeoffs. As we will see, all attempts at library repositories balance these tradeoffs in different ways and with varying degrees of creativity and success.

Until World War II American libraries were their own repositories. But following the War there was a huge expansion of both higher education and of libraries in the U.S. Respectable academic library size went from collections in the thousands of volumes (around 300,000 average) to collections in the millions (around 2 million). Fremont Rider made a considerable reputation for pointing out that this explosive growth was

taking place not in just one or two libraries but across the board and that continuing to increase library capacity by building new buildings simply was not feasible (Rider, 1944). The first impulse for many library professionals was to continue to use the library itself as the repository but just more efficiently. Rider's solution, as most of you may recall, was microform collections. This would allow more efficient use to be made of existing library space through the application of technology. (Sound familiar?!) Microforms (because microfilm was soon joined by microprint and microfiche) enjoyed a considerable vogue for many years. Vendors began producing huge microform sets (Chadwyck-Healy, UN Docs, British Sessional Papers, etc...and even libraries joined in production as through such major initiatives as the National Register of Microform Masters (NRMM) project).

Nevertheless, although still in use, their day has clearly passed. Patrons never liked them, Rider's math turned out to be fundamentally flawed, archival problems surfaced with the technology, and finally, a new technology - digital information - developed which seemed more promising. Rider's proposed solution also did not provide a good balance for one of the core dilemmas mentioned earlier. Microforms preserved the information but did not preserve the artifact. While most of us are comfortable discarding a newspaper after microfilming it - with the exception of Nicholson Baker (2001), of course -, was it really possible or desirable to discard a Gutenberg Bible after microfilming it?

As a consequence, Americans in the post War period also began experimenting with a parallel strategy which sought to balance the tradeoffs of the growth dilemma differently. This is the strategy which we are discussing today - a library repository, by which is meant a facility, physically separate from the library dedicated exclusively to the storage of library materials (rather than the provision of the full range of library services such as cataloging, reference, ILL, etc.). The first such post war U.S. effort (foreshadowed by the New England Depository founded in 1942) was the Center for Research Libraries (CRL). Begun in 1951 as the Midwest Inter-Library Center. Its current name was taken in 1965 when its scope was expanded from a regional one to encompass a national agenda. It differed from the Rider approach in three important ways. It focused on an external, consortial, paper oriented approach. Its tradeoff to the paradox of whether to preserve the information or the artifact was to do both, but in a more cost effective manner by reducing duplication among libraries of ultra low use materials. It also addressed (in a minor way) the paradox of requesting new money when there was not enough room to house already owned materials by freeing up stack space at the member library through transferring ultra low use material to the Center. The consortial approach was also helpful in that a group of libraries could share the cost of housing this ultra low use material. It should be noted, however, that it was not a pure repository since it also operates in a minor way as a cooperative collection development agent for its members, i.e. purchasing some low use materials which none of the members own. The Center today has some 90 full members from across the U.S. and stores around 3.5 million volumes most of which have come from member libraries. Although very successful for

a while and still in use, a combination of management problems, scalability issues, and ownership issues across state lines has taken it largely out of the repository spotlight in the U.S. Further information is available at its web site.

The real repository development is more recent. Since the 1980's in the U.S. a combination of increasing fiscal austerity in higher education (reducing the ability of universities and colleges to build new libraries) and the continued growth of print collections even at slower rates have created widespread and severe space problems which have given a new urgency to the repository agenda. These new repositories go far beyond the Center for Research Libraries. Characterized by a highly specialized offsite building, major, rather than token transfers of library material, and ownership by an individual library or a state defined, such repositories are serious library warehouses. Physical plant design for these new repositories has split into basically two approaches - what could be called traditional/opportunistic and the Harvard model. Although sharing much in common, there are also pronounced differences. Common elements include special attention to humidity/temperature/lighting, processing and retrieval workspace, shelving by materials size rather than subject class, identification of stored materials in member library OPACs and massive storage capacity. To illustrate this last point, an important one, a comparison between CRL and a post 80s repository might be useful.

CRL has a present collection of 3.5 million volumes. This may seem a lot. But consider, this collection was built over a 51 year period from an average of roughly 60 members nation wide. This represents an average of 68,627 volumes being transferred annually from roughly 60 members or just over a thousand volumes a year. As noted earlier, since the CRL has also functioned as a common buying club for some materials these means even fewer materials being transferred. In contrast, just one of the five Ohio repositories – Ohio Southwest Regional Depository - serving just four academic libraries has developed a collection of 1.5 million volumes in under 8 years. This shows a transfer rate of 187,500 volumes per year for four libraries or 46,875 volumes transferred annually per library. This represents an increase in transfer rate of over 40 times (40.97) from CRL. Such large transfer rates are common. The first module of the California Northern Regional Library Facility was filled with 3.1 million volumes in 6 years. With four large member libraries transferring materials that comes to 129,167 volumes annually per library. Although the transfer rate declined while filling the second module, 2.37 million volumes over ten years for a 59,250 volume annual transfer rate per library, that is still well above a 40:1 ratio with CRL. Incidentally, both California Northern and the Ohio Southwest repository just mentioned are each well along on planning their third repository module. In short, these repositories are not a one shot, one time solution but an ongoing way of life for the foreseeable future.

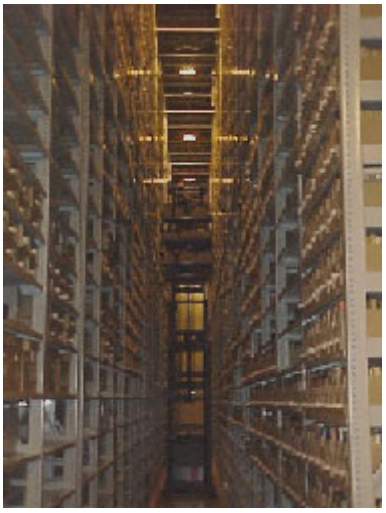
The traditional/opportunistic style is a very loose category and can involve either new physical plant construction or renovation of existing buildings. It can be exemplified by the California Northern Regional Facility constructed in 1980 and the Buhr Shelving

Facility, a former manufacturing plant which was purchased and remodeled as a storage facility by the University of Michigan in 1981. This model is characterized by a fairly traditional building, a major use of compact shelving as well as the common repository elements already mentioned. The attraction of retrofitting an existing building is generally the initial cost which is less than constructing an entirely new structure. There are two main problems with the approach, however. The first is that the retrofit can end up being more expensive than expected. Floors, even factory floors, may not be strong enough for compact shelving, power and telecommunications lines may be inadequate or out of date, and, of course, there are limitations and inefficiencies in a physical layout not designed specifically for print storage. The second problem is that preexisting buildings are not designed to be added to in any kind of modular way. This makes substantial expansion of the facility either impossible or hugely expensive. This model continues to be used. For example, the Ohio Southeast (not Southwest!) Regional Depository just began retrofitting a used car dealership building to use as their repository. New construction along traditional lines also continues as witnessed by the newly constructed California Southern Regional Library Facility which has recently come online at UCLA. Located right on campus and making heavy use of compact shelving in three tiers or floors, it is expected to hold some 7 million volumes. Nevertheless, the traditional/opportunistic model just described no longer represents the library repository mainstream in the U.S.

That mainstream is now dominated by what has come to be called the Harvard model. Pioneered by Harvard in 1986, this model is characterized by a highly specialized and somewhat radical storage building. Requiring new construction, the storage area of the repository is basically a large air tight cube with a super flat floor, 30 foot high stacks, and retrieval by an electric 'cherry picker'. This last item is an electric forklift with a large, extendable arm and platform which allows materials to be retrieved from anywhere in the oversized shelving. The Harvard model approach to repositories includes over 20 present buildings across the U.S. with at least three more under construction.

Since this is a somewhat unusual design, I thought that some pictures would be helpful. Harvard is traditionally reticent about self display, so these pictures come from Yale when they were constructing a similar repository several years ago.

Probably the first thing to note is that when I say the Harvard model is 'radical' I am not referring to the external architecture. This is clearly not the Bibliothèque Nationale or the new British Library! And they all look very much the same whether in New England, Ohio, Texas or elsewhere. The radical nature of the Harvard model is a fanatical, possibly obsessional, concern with storage efficiency, not graceful building design. The basic footprint of the building is a delivery dock, processing area and staff offices attached to a large storage cube. One wall of the cube is constructed with a large breakaway section so a substantial service door can be added when additional modules are constructed. This allows each new module to be easily serviced by the same, original service and processing area.



The core of the design is the unique stack area - the big cube. It consists solely of stationary shelving running 30 feet (10 meters) high with vertically adjustable shelves. The stacks, double sided, run the whole length of the storage cube, typically 175 feet (58 meters) in length plus a small space in front for the cherry picker to maneuver in. The shelves are 3 feet (1 meter) deep, being designed to hold cardboard shelving trays rather than individual books. The storage cube needs a super flat floor. Such a floor is smoothed using laser measurements to create a surface that has no measurable irregularities or slope. As you can imagine, the combination of the shelves' height and the fact that they are fully loaded all the way to the top means that it is critically important that there be not the slightest leaning.

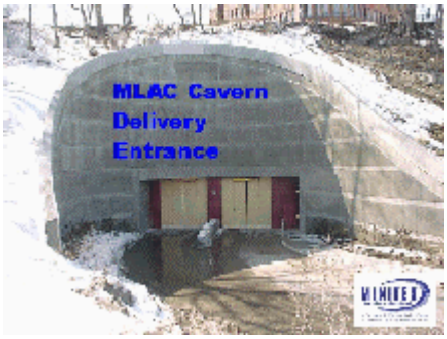
The processing area involves a loading dock for truck deliveries, a place to clean the delivered materials (they arrive amazingly dirty), and work space for sizing, bar-coding and loading books in their boxes. The sizing is particularly important since shelving books by size is critical for maximum storage. Typically, books are sorted into five different sizes plus oversized and loaded into cardboard containers of 20-30 books a piece. Added to the book's item record (and to the book itself) is its box number. Box numbers are assigned like American street addresses, by geographical coordinates, e.g. 1st stack, second section down, 10th section up. The book's location is not identified further than its box number. This presents no problem in retrieval.

The most fun part of this arrangement, if you're into adrenaline rushes, is the cherry picker. This is an electric fork lift with a large extendable arm ending in a small caged platform. The platform has remote controls so that employees can stand on the platform and maneuver the vehicle up and down the range and the arm up and down the 30 feet of shelving. This allows for the fairly convenient placement and retrieval of materials. While swaying around at the top of a fully extended cherry picker arm is not everyone's idea of fun, it does appear to be extraordinarily safe since in almost 10 years I have heard of no accidents in any of the 4 Harvard model Ohio repositories or in any other repositories for that matter.

A particularly nice touch when retrieving materials, a process which can mean visiting all parts of these extensive ranges, is pick list software. If, on a particular day, 10 books are requested from various locations, the list is run through the software which organizes the pick list in the most efficient manner for retrieval. No thinking or planning necessary for the student or staff member. Just print out the list and go get the books in that order.

It is perhaps, unfair, to leave you with the sense that it is entirely impossible to associate the Harvard model of repository with no creativity at all. At the University of Minnesota the Minitex consortium has put its repository not in a cube of a building, but in an underground limestone cavern. Opened in January 2000 the Minnesota repository, one of the U.S.'s more innovative library repositories, is buried 82 feet underground in one of the many limestone caverns located along the Missouri River. The cavern, fortuitously, is located underneath the main library of the U of M's West Branch Campus in

Minneapolis. Those of you acquainted with Mark Twain's Tom Sawyer will recognize and appreciate such a location.



Although the basic design is maintained, the stack height is only 18 feet (6 meters). Still, such shelving requires forklifts ("cherry pickers") as described earlier, shelving by size, etc.

Clearly the underground nature of the facility represents a huge convenience and savings in terms of maintaining optimal environmental conditions - particularly in a severe northern climate.

In addition to providing a storage location under optimal conditions for print materials, these modern day repositories provide several other major advantages. The first is the efficiency of the shelving. Shelving by size in oversize stacks dramatically increases the number of materials which can be stored. Minnesota studies indicate a 40% gain in storage capacity and the bigger installations (higher and/or longer) will increase that number even more. A second important point is cost. An informal review of comparative costs by Orbis, a coalition of academic libraries in Oregon and Washington, indicated

that the construction cost per volume was $3.75 for a high density facility compared to $13.39 for traditional campus library construction. Yale reports an even high rate of savings, calculating that off site storage is 1/10th as expensive as traditional library open stacks housing. For what they do the modern repositories are relatively cheap to build and very cheap to maintain. A third important point is that these repositories are basically local, serving a relatively small number of nearby libraries. This is a very important feature when the library tries to convince local faculty to agree to let 'their' materials be moved to another location. Take it from me, 'down the street' is a much easier sell than 'across the country', although neither is a walk in the park. And, of course, retrieval speed is also enhanced with a local, rather than national, facility.

Let me conclude this section by noting that modern repositories are perfect partners to deal with the growing number of titles and back issues of JSTOR journals. The repositories archive the seldom used print copies, seldom even bound any more, just shrink wrapped, while quick and convenient digital access handles the lion's share of the content use. And, of course, while books and journal runs continue to be sent as artifacts to the requester, journal articles are usually faxed or scanned and sent as digital attachments to email. All in all, the system works very well. For those of you interested in more follow-up on the subject of print repositories, take a look at the guidelines for storage of books on the West Virginia University Book Depository website (Nitecki & Curtis, 2001).

DIGITAL ARCHIVING AND LIBRARY DILEMMAS

This leads us to the issue of digital repositories? To begin with, in what ways do they allow us to address our three library dilemmas with a different set of tradeoffs and solutions. For all three of our library archiving dilemmas digital repositories represent an interesting set of new advantages and problems. In terms of use versus destruction a digital repository almost provides a solution to this dilemma. Clearly, digital use does not wear out the zeros and ones of which this information is constituted. This would be wonderful were it not for the fact that other issues have risen to complicate matters. There is the issue of acceptance by users, digital rights, and above all, that no clear path has yet been identified and tested for how digital materials are to be reliably and permanently archived. OCLC, LOCKSS (Keller, 2003), partnerships between libraries and publishers to create dark archives, are really only still experimental. In fact, we have in one sense only changed the terms of the dilemma. We don't need to worry about wearing out digital material, but we do have to be concerned that it may simply disappear. This may not be progress.

In terms of growth versus storage room, digital repositories represent as close to a solution to the dilemma as we're likely to get. Using JSTOR to replace vast sections of shelving devoted to bound journals is a major step in reducing stack overcrowding in the

U.S. And, in terms of digital storage itself, it just keeps getting cheaper and more abundant. There may well be problems down the road when we begin to transition old digital formats into more recent ones, but for now we are in a very sweet space as far as storage is concerned.

The final dilemma, artifact versus information, remains a tradeoff. We can digitize the content and throw away the artifact (newspapers), or store a single copy of the artifact (last remaining copy approach), or store all copies of the artifact (usual practice!). It is also possible to digitize the text using text based software which allows the text to be searched and manipulated and link such text to image representations of the page so that layout, founts, marginalia, are revealed. By having both an ASCII, PDF or other version of the text matched with an image version of the text, you can to a degree approximate retaining the artifact. Nevertheless, this is an area where difficult decisions still must be made.

This is also the place to point out that the increasing number of born-digital materials means we will have to come to terms with the archiving of digital materials whether we wish to or not. The preservation strategies which have worked so reliably for so long in the print on paper world and even the hybrid world of print/digital, will simply not work in a born digital world. 'Digital artifacts' cannot be archived using a print on paper approach without the risk of major information loss.

The problem in talking about digital repository initiatives in the U.S. is that almost every American academic library in the country is involved in creating some kind of digital repository. And we are not just talking about text repositories, but images, videos, sound files and born digital as well. The range of experiments with digital repositories in the U.S. is tremendous and to adequately deal with them and their issues would require at least a conference, not a part of a presentation.

Let us begin to sort through this rich confusion with some clarifications. I would first distinguish between libraries who are creating true scholarly repositories and those who are essentially creating advertisements for their print collections. The majority of digital materials put up by libraries fall into the advertisement category. Wright State University in Ohio, for example, has put up a collection of early photos of Wright Brothers flight experiments. They're quite interesting for a casual searcher or even possibly useful for teaching children about this important development in American history. But the image resolutions are quite low and the associated bibliographic material sketchy so that they are useless for scholars to use online. What they can and do show is what is available in the traditional archive for scholars who would want to make a visit and see the actual artifacts.

Another clarifying point which needs to be made is that the main focus of many of the initiatives, particularly many of the bigger ones, is not first and foremost preservation.

Rather the main goal is on increasing access or providing a cheaper business model to the material of scholarly communication. There is certainly and emphatically nothing wrong with this but it is important to be clear which part of the library mission they are primarily addressing. It is important to keep in mind because, as noted earlier, there is no universal agreement or established track record on how digital materials will be preserved. In short, from a preservation point of view what we have is either a huge mess or a wonderful opportunity. It's the Italian Renaissance all over again and Benvenuto Cellini is waiting around the corner. Whether he will create an artistic masterpiece of a salt cellar or cut our hamstrings and cripple us for life (he did both in his days) remains to be seen.

So, as this drama plays out, let me identify for you some of the key players and web locations to keep an eye on. The first place to keep an eye on are major U.S. research libraries. A summary of the activities of 60 libraries describing 150 digital publishing projects as of last summer can be found on the ARL (Association of Research Libraries) web site (Collections, 2002). Providing brief institutional background data (including name of appropriate contact person), individual summaries describing and explaining each project are provided. These are library rather than institutional projects and provide a fascinating and thought providing overview of the digital collection activities of some of the most significant U.S. library players.

Second, I would suggest reviewing the information (and collections) at the Library of Congress on digital repositories. As one of the earliest and certainly one of the biggest players in digital preservation and repositories and a compulsion to tell and document everything, it is almost too much of a good thing. Not only does it represent a wealth of examples in the digital collections represented in text, audio, video, image and various combinations, but it is almost overwhelming in the amount of background information provided - digital standards, best practices, scanning techniques, workflow procedures, metadata issues and the like. Begun as the American Memory Project in the 90s, LC's digital repository efforts were expanded and broadened to born digital earlier this year through the National Digital Information Infrastructure and Preservation Program. Still in terms of digital preservation and archiving this is one of the most serious sites you could visit.

Third and fourth are the big institutional programs at the two ends of the country - MIT and California. The most impressive by far is MIT's DSpace which came online in November, 2002. Developed as a joint project of MIT Libraries and the Hewlett-Packard Company, DSpace consists of two key elements. The first is a free database of articles, preprints, working papers, technical reports, books, theses, data sets, computer programs, simulations, etc. It is a database of digital works from MIT professors, students and staff which is mounted for web access and fully bibliographically described. It is distributed via the web and it is intended as a long term archive, and can be searched. The second key element is the supporting infrastructure which is also available to any educational institution. For free! It cost millions of dollars to develop but is being made freely

available. This is a Marshall Plan for higher education and clearly has as its goal a fundamental transformation of scholarly communication. The vision statement is simple and compelling: "A federation of systems makes available the collective intellectual resources of the world's leading research institutions". Since November over 2,000 institutions have downloaded the infrastructure software. A core group of major research libraries has been formed to serve as the core of the Federation. Developments here bear watching.

At the opposite end of the country is the California Digital Library's eScholarship Repository. Although called a 'library' and functioning as a digital library, it is an initiative of the UC central system rather than any library groups. Less radical than the MIT approach, the EScholarship Repository simply provides a central location for faculty and the various institutional elements (institutes, departments, etc.) of the UC system to deposit digital material in the form of working papers, technical reports, research results, datasets and even peer-reviewed series. The material is made freely available to anyone in the world. Begun in April 2002, the repository reported 1,200 papers as of March 2003 with about 60,000 full text downloads of repository materials.

Less well known, but perhaps more creative, are three other serious scholarly digital repositories you might want to track. The first is the University of Virginia Electronic Text Center. Although combined with commercial databases such as Chadwyck Healy a large number of UVA materials have been scanned in as well. There are some 70,000 humanities texts in 13 languages with over 350,000 associated images. The quality of the digitizing and the supporting metadata is at a scholarly level. The second repository worthy of brief notice was started at a neighboring university Virginia Tech and that is the Networked Digital Library of Theses and Dissertations (NDLTD). Begun as a pilot project at Virginia Tech in the mid 90s to provide free access to digital versions of its students' theses and dissertations, it has expanded to become a network of 172 libraries and library consortia world wide providing digital access to this core scholarship of their institutions. While the repository is distributed at the present time, preliminary linking has already been established and progress is already being made on a one-stop shopping solution. The third repository - Television News Archive collection at Vanderbilt University - is only in the process of being digitized, but it is a most interesting one. Through a strange fluke of circumstances, Vanderbilt University in the U.S. received permission in 1968 to record the evening news broadcasts of ABC, CBS, NBC and CNN plus their special reports, e.g. Watergate, 9/11 and the like. Each broadcast is extensively cataloged and the repository holdings can be searched online. Searching is free to anyone though registration is required and a modest charge is made if you wish tapes made of a particular segment. Funding has been received to digitize this archive so its resources should be soon available over the web.

I think there are two major questions surrounding these and other digital repository initiatives. The first has to do with repository content generated by the member scholars.

Are these repositories going to become the location for the latest, top drawer research or turn into the digital storage attic of possibly useful but really second tier research, committee reports, and so on while journals continue to be the 'repository' for the important stuff. The second question involves the sustainability of the business model. These are very expensive propositions to put into place and not generate any income stream. I remind you that in the U.S. the university community has traditionally given away little for free, especially recently. Even the internal subsidy which universities provide to their presses (and they do not give their books away for free) has come under fire in a number of institutions. The University of Arkansas and Northwestern come to mind. And U.S. libraries charge each other for ILL use. Providing free information access to the world academic community is unusual. The essence of a library archive is reliable permanence. The question is, can a completely free archive be permanent? We are, in short, watching a very bold and, frankly, very idealistic experiment in progress.

CONCLUSION

Responsible archiving has always posed problems for librarians - from Biblos and Alexendria to the present. What I have hoped to show today is that U.S. librarians are actively and creatively involved in seeking solutions. These solutions must look both backward at the massive amount of print on paper materials we have accumulated and continue to accumulate. And they must look forward to encompass how a new digital reality offers a new set of tradeoffs and partial solutions to traditional archiving issues and poses as well a whole new set of problems and tradeoff decisions. But then, we're librarians and that's what we do.

NOTES

1. The talk was to be presented at the 2003 LIBER annual meeting but wasn't, due to the illness of the speaker.

REFERENCES

1. Baker, N. Double Fold: Libraries and the Assault on Paper. Random House, 2001.
2. "Collections & Access for the 21st-Century Scholar: Changing Roles of Research Libraries". ARL, issue 225, December 2002. http://www.arl.org/newsltr/225/

3. Keller, Michael A., Vicky Reich and Andrew Herkovic. “What is a library anymore, anyway?”. First Monday, 8(2003)5. http://www.firstmonday.org/issues/issue8_5/keller/index.html

4. Nitecki, Danuta A. and Curtis L. Kendrick. Library Off-Site Shelving: Guide for High-Density Facilities. Englewood, CO: Libraries Unlimited, Inc., 2001

5. Rider, F. The Scholar and the Future of the Research Library. New York : Hadham, 1944. 236 p.

WEB SITES REFERRED TO IN THE TEXT

American Memory Project. http://memory.loc.gov/ammem/ammemhome.html

Association of Research Libraries (ARL). http://www.arl.org/

Center for Research Libraries (CRL). http://wwwcrl.uchicago.edu/

DSpace. http://www.dspace.org/

California Digital Library’s eScholarship Repository. http://repositories.cdlib.org/escholarship/

Library of Congress. http://www.loc.gov/

LOCKSS - Lots of Copies Keep Stuff Safe. http://lockss.stanford.edu/

National Register of Microform Masters (NRMM). http://www.arl.org/preserv/nrmm.html

Networked Digital Library of Theses and Dissertations (NDLTD). http://www.ndltd.org/

OCLC Online Computer Library Center, Inc. http://www.oclc.org/home/

Television News Archive. http://tvnews.vanderbilt.edu/

University of Virginia Electronic Text Center. http://etext.lib.virginia.edu/uvaonline.html

West Virginia University Book Depository. http://www.libraries.wvu.edu/depository/process/index.htm

The Wright Brothers in Photographs. http://www.libraries.wright.edu/special/wright_brothers/dmc.html