Planning for a University of Chicago Digital Repository

A Discussion Paper

Submitted to the Library Administrative Committee

Winter 2012

By

Charles Blair

Director, Digital Library Development Center

University of Chicago Library

Table of Contents

Planning for a University of Chicago Repository

The following is a framework for further discussion focusing on three constituents of an institutional repository at the University of Chicago: the University of Chicago Library Digital Repository, research data, and a publication repository.

The University of Chicago Library Digital Repository

Size and Scope

The University of Chicago Library Digital Repository (DR), which is managed by the Digital Library Development Center (DLDC), consists of over 6 terabytes (TB) of materials (almost 1.6 million files). They include photographs, musical scores, manuscripts, maps, administrative records, electronic theses and dissertations, oral histories, symposia, personal papers, including electronic mail, circulation records, performances, scientific datasets, one small website, etc. Materials are deposited in a variety of digital formats and are of different types: still images, sound, video, text, metadata, datasets, etc. Some are born digital and some are retrospectively digitized. Most of these materials are part of the University Archives.

Expected Growth

The DR is expected to grow at a fast pace, but it is not easy to quantify the rate of growth. An addition of 20TB has been projected for FY12, but some of this represents a backlog of digitized materials which had been vended out for reformatting. Accurate predictions can be made only by observations over several years. However, we can confidently predict that the rate of growth will be non-linear for at least the following reasons.

The Special Collections Research Center (SCRC) is responsible for managing deposits of University records. Daniel Meyer, University Archivist and Director of the Special Collections Research Center, has been asked by David Fithian, Secretary of the University, to perform a survey of records created by academic and administrative units of the University. Because these records are increasingly generated and maintained electronically, the survey will reveal the nature and extent of the need for electronic records management at the University of Chicago (digital files, database records, email, etc.). One such deposit has already been accessioned into the DR; it will be mentioned below.

The Digital Collections Steering Committee is actively pursuing a policy of identifying collections for digitization and making them available online; this activity is supported by the Preservation Department. Some of the digitization is performed in house; some is outsourced. The pace of this activity is projected to increase and, even at current staffing and funding levels, guarantees a steady flow of digitized documents into the DR.

Description, Discovery and Access

Many materials in the DR belong to collections which also have analog components. These collections are described using electronic finding aids and contain links to their digital components. These materials cannot be treated as stand-alone resources for two reasons.

First, archival materials are rarely described at the item level; description stops at the folder level. Item-level description does not exist and will not be created for most of these materials. Therefore, the finding aid provides descriptive context, or the information required for discovery.

Secondly, in the case of collections consisting of both analog and digital components, the complete context of any component is provided by the whole collection, not simply the digital component.

Some materials not born digital are digitized as part of the Special Collections Research Center's large-scale digitization program, which brings researchers directly to digital documents from finding aids. These finding aids are accessible through the Library's electronic manuscripts and archives finding aids system; some are also available through UNCAP. All are also potentially available to any discovery service that can import and display information from finding aids, such as Lens. Examples include the *Fielding Lewis Papers, 1783-1900*, which includes records from an ante-bellum slave-owning estate, and the *American Recipes, 1855-1905*, originally from the Crerar collection.

Thirty collections currently contain links to digitized materials. The largest of these is the *Chicago Committee of Fifteen Records, 1909-1927*, the digitized component of which is 400 gigabytes (GB) in size (almost half a terabyte).

There will be at least three ways to discover and access archival materials which have digital masterfiles and use copies in the DR, because all guides will have at least two methods of discovery: the dedicated finding aids website and Lens; for some, UNCAP will make three. Because finding aids can be collected and aggregated by third-party entities, such as OCLC's Archive Grid, some collections may have four ways of being discovered today.

Some materials deposited in the DR are exposed for discovery and access by means of collection-specific websites, such as the University of Chicago Photographic Archive, Century of Progress World's Fair 1933-1934, Chopin Early Editions, etc. Records for these may be found in the Library catalog, Lens and WorldCat; Chopin Early Editions is an example.

Collection-specific websites typically provide specialized functions such as purpose-built browse lists. For example, the browse lists for the Photographic Archive and Chopin Early Editions look very different from each other: the former will not contain a browse list for genres or dedicatee; the latter will not contain browse lists for photographers or student activities. Specialized functions targetted at audiences for particular collections will not be provided by a one-size-fits-all generic repository interface.

In the very near future one can imagine systems specialized for streaming audio or video, sometimes using authorization and authentication systems to deal with rights management. This approach—a common repository with specialized access systems built around the core—is gaining ground, because "no single system can provide the full range of repository-based solutions for a given institution's needs, yet sustainable solutions require a common repository infrastructure." (http://www.clir.org/dlf/forums/fall2010/08Hydra.pdf)

Keeping data and metadata in a repository, but providing discovery and access through other systems with specialized functions, is also a good strategy for system migration, because instead of migrating from an old system to a new system, with the inevitable conversion headaches, one can simply populate the new system with data and metadata from the repository. This strategy has been referred to as "durable objects, ephemeral applications." (ibid.)

In the old, mainframe days, "ephemeral applications" was not entertained as a concept; building applications was expensive and slow. Today, complex presentation systems built from web frameworks, such as Ruby on Rails or Python's Django, considerably simplify the manipulation of the increasing number of parts which comprise modern systems; therefore, while building applications still has a cost, the expense is significantly less than it was, and applications are not expected to last forever, if only because of the rapid pace of technological change.

Combining the functions of a system designed for long-term preservation and one designed for immediate presentation may have potential financial implications, because systems optimized for the latter consist of more expensive components (disk, memory, CPU, caches, clusters) to support faster access and immediate failover capability than those that do not. Tape is considered to be the most reliable medium for long-term storage—unlike disk, a tape can last twenty or more years without degradation, but tape access may take two or three minutes, which is an unacceptably slow response for many applications.

Lists of materials (data and metadata) deposited in the DR are available for inspection by depositors by means of an electronic inventory listing created at the time of deposit as part of the automated accessioning process. Deposited items are accessible by the same means. For example, if one deposits a TIFF image into the DR, one can click on a link to the TIFF image in the inventory listing for the deposit and retrieve the image from the DR in real time. This is a restricted, staff-facing function: registered depositors can view the materials which they or others in their group (e.g., SCRC or the Preservation Department) have deposited.

A simple, public-facing repository front-end presents a listing of the deposits in the DR, for example:

```
Lewis, Fielding. Digital Collection
Dec. 6, 2010, 9:26 a.m.
Kathleen Arthur
Preservation Department
ID: acr1p4v5c1j

American Recipes. Digital Collection
March 29, 2011, 2:38 p.m.
```

```
Kathleen Arthur
Special Collections Research Center
ID: ac62b2x158k
```

Discovery and access to fully processed items whose accessibility is not restricted is provided by public-facing websites, such as the electronic finding aids site, collection-specific or purpose-built websites, Lens, UNCAP, the Chicago Collections Consortium Portal if and when it is built, and WorldCat, as mentioned above.

Constraints (Rights and Permissions)

Not all materials in the DR are available for access.

Some materials are embargoed for long periods of time. These include electronic administrative records such as the Advance Meeting Materials and Minutes for the University Board of Trustees meetings, deposited by the Secretary of the Board of Trustees. These materials, though they may be discoverable, may not be made accessible except to the Secretary and his delegates, because access is intentionally restricted; a system to allow restricted access to these materials has been created. Part of the DR is therefore intentionally a "dark archive" with respect to public access, if not discovery. The dark archive also contains materials for which the right to disseminate from the archive has not yet been secured, for example, some oral histories.

Some materials are not meant for permanent retention. Although the archivists in the Special Collections Research Center accession all deposited materials as part of their workflow, it is not their policy to retain all deposited materials indefinitely: some materials from accessioned collections may be removed during the processing of those collections if not deemed worthy of long-term retention. The DR will therefore contain some materials which are being made available to archivists as part of their workflow but which are not necessarily intended as material to be made permanently available to researchers from the DR. Part of the DR is thus meant as working space for archivists. The notion of a "work space" is familiar from large-scale scientific data repositories, which may have a "scratch space" and a "work space" (or "project space") in addition to an archival space for digital preservation.

Some materials are duplicates of copies held elsewhere and intended to be discoverable and accessible from those locations, for example, electronic theses and dissertations, or electronic journals to which the Library subscribes but which are not archived in Portico. In these cases, the DR's copy is simply an archival preservation copy, access to which is intentionally restricted.

The Oriental Institute (OI) is about to deposit images and associated metadata from the Persepolis Fortification Archive project into the DR. Some of the images are digital masterfiles and are meant as archival preservation copies. They are expressly not intended to be made accessible directly from the DR. The OCHRE system, which the Library maintains for the OI, is designed to provide discovery and access for metadata and use copies in a search and retrieval context designed specifically for and by the primary research audience, which consists of archaeologists, philologists, etc., world-wide.

Summary and Next Steps

It is fair to view the DR at this time as concentrating on materials most of which are locally created and which, if not curated by the local institution, one cannot expect anyone else to curate. A few materials may not be locally created, for example, rarely held objects, such as some maps, the *Woods Hole Laboratory Reference Documents*, deposited by Susan Kidwell, William Rainey Harper Professor in the Geophysical Sciences, and some licensed resources (e.g., electronic journals not archived elsewhere), access to which we want to ensure in perpetuity. However, as a rule, the DR does not intentionally duplicate the functions of other, well-established repositories, such as HathiTrust or Portico. The DR may therefore be considered to exist in an ecosystem of other similarly purposed repositories. The DR also does not at this time duplicate discovery and access functions to fully processed materials discoverable and available by systems which the Library maintains or to which it contributes and publicizes to its users, as discussed above. The current model is therefore similar to that of the Stanford Digital Repository (SDR).

Overview
The Stanford Digital Repository (SDR) provides digital preservation services for scholarly resources, helping to ensure their integrity, authenticity, and usability over time. Services are focused on protecting against data loss and mitigating long-term risks to accessing digital information in ever-evolving technological contexts. To support these services, the SDR system is built to be flexible, secure, and sustainable.

Background
...
In operation as a production system since 2006, the SDR currently contains more than 60 terabytes of text, manuscripts, images, maps, GIS data, and audio-visual content. It serves as the preservation repository for significant collections from Stanford University and beyond, such as the National Geospatial Digital Archive, the Parker on the Web digital manuscript project, historic recordings from the Monterey Jazz Festival Collection, and Stanford's own digitization efforts.

Services Scope and Definition
The SDR services are available for content of any scholarly discipline, regardless of data type. The system's access and security model can also accommodate a range of needs, from preservation of open access content with no licensing or security restrictions, to private or sensitive content that must be kept "dark" for finite periods of time.

Within the larger context of Stanford's digital library, the SDR serves as an underlying layer, designed to prioritize content integrity and bulk operations over immediate and granular accessibility to ingested contents. As such, it is not a back-up system, but rather designed to serve a "back office" preservation function, while separate but complementary digital library and information technology systems provide end-user discovery, delivery, and access environments. (Stanford Digital Repository)

The Director of the Digital Library Development Center and the Director of the Special Collections Research Center have discussed some of the staffing needs of the DR. The need for a digital accessions specialist has clearly emerged. This position would function like the Digital Collection Manager but with responsibility for

managing electronic acquisitions. A position description has been drafted. If, after consideration and review, it is approved, it is intended that the position be hired and supervised by the SCRC, but coordinate activities with the DLDC. The accessioning activities currently being performed by the DLDC would then have permanent staffing and full attention devoted to them. The DLDC would provide the tools to automate workflows; the digital accessions specialist would use these tools to perform the work.

The need for a specialized professional position in SCRC for digital archiving has also been discussed, but it is has been deemed too premature to take action at this time: a precise definition of the position will depend on the outcome of the records management survey mentioned above. The DLDC may need more staffing to support the growth in digital archiving, but it is too early to say anything specific at this time.

Research Data

Size and Scope

Data vary in size.

Some data, such as the *Woods Hole Laboratory Reference Documents*, are small and human-readable. Some, such as the Persepolis Fortification Archive, are large and human-readable. Others, such as the Sloan Digital Sky Survey Data Archive Server (SDSS DAS) are large and machine-readable. Some are created by hand in Excel spreadsheets. Others are created automatically by sky telescopes or particle accelerators. Still others are derived by analysis from other data.

Data vary in type.

Some data are scientific; some are not. For some researchers, images, sound recordings or video are their data. The Persepolis Fortification Archive consists of high-resolution digital photographs of inscriptions.

Data vary in purpose.

Some data are archived for long-term safe-keeping. It is not expected that they will often if ever be referred to in the near term. (Oral communication from the University of Chicago Research Computing Center.)

Some data are required to be preserved and shared as part of data-intensive science projects. See, for example, Yale University, Office of Digital Assets & Infrastructure: NSF Data Management Plan.

Some data are intended as supplemental materials to journal articles. Supplemental materials include "multimedia—or text, tables, and figures that would occupy too much space or would interrupt the flow of the narrative in a traditional print article—as well as data and computer programs. These vary in importance to supporting the article's conclusions. Some may be absolutely essential, whereas others may be useful, but not critical." (Recommended Practices for Online Supplemental Journal Article Materials) Recommended best practices for the

publication of supplemental materials have yet to be finalized; see the NISO/NFAIS Supplemental Journal Article Materials Project. There is a need to expose data to support published articles both in the sciences and in the humanities. One of the earliest requests for the Library to house such data came from the Society for the Study of Early China (SSEC).

Some data may be acquired from elsewhere to be used in local research. Examples might include full-text data from Google Books or HathiTrust.

Data vary in value.

Some data, once recorded, cannot be recreated. For example, the University of Chicago is currently curating the Sloan Digital Sky Survey (SDSS) DAS (75 TB) and CAS (22 TB) data together with the Johns Hopkins University Library. The "... MOU [between the University of Chicago and The Astrophysical Research Consortium (ARC), which expires at the end of 2013] proposes a fixed-term agreement ... for practical reasons, but this does not suggest that responsibilities for the curation of the data are expected to terminate in the foreseeable future. Rather, at the end of the agreement, ARC will reevaluate the status of the data archive and determine how best to construct a new agreement between the partners." (From the MOU)

Some data can be recomputed inexpensively, in which case it is not necessary to archive the data, but simply to redo the computation.

Expected Growth

It is expected that data as supplementary material to published work will continue to be produced. It is expected that the amount of these data will be modest. It is as yet unclear how much "big data" will be produced and expected to be archived or made accessible at the University. The Research Computing Center is reluctant to make predictions before observing faculty use of its high-performance computing cluster. Estimates could only come from interviewing the faculty themselves, if they are able to predict their needs, together with surveying institutions with comparable activity.

Description, Discovery and Access

As with published articles, access to supplementary data can be achieved using the Digital Object Identifier (DOI) System. The DOI Data Model contains details on what and how metadata are recorded. These can be crosswalked to other standards for inclusion in their discovery and access systems. Cataloging according to this standard is potentially a role for a metadata librarian.

The Library is in the process of acquiring DOI minting capability for articles. The immediate use is for technical reports published by the Computation Institute, but DOIs can also be applied to data: "For creations, the abstract nature of the content of a referent, irrespective of its creationStructuralType, is typically described by creationType, which may be extended as needed to include format and genre elements (for example: audio file, scientific journal, musical composition, dataset, serial article, eBook, PDF)."(Ibid., under referentType).

Constraints (Rights and Permissions)

The same considerations apply to data as to any other types of materials: some are open for discovery and access, some may be discoverable but not immediately accessible (embargoed), some are constrained by governmental regulations or local policy, etc.

In addition, not all research is grant-funded. "Some scientists dread data management plans because they are worried that others will guess what they're researching." (From a presentation given at the Preservation and Archiving Special Interest Group [PASIG], 2012).

Therefore, unless the Library or University has an open-access policy for research data in any locally managed repository, which is far-fetched, necessary constraints must be implemented to ensure data security as needed.

Summary and Next Steps

Small data can be handled much as other digital resources are handled, whether for preservation or for discovery and access. Some data can be delivered as files; others may need specialized systems for delivery.

It is reasonable to expect that the Library or the University will partner with others for big data. It is reasonable to decide not to manage all large-scale data curation unassisted, but to consider a consortial approach at least in part. For example, curation of the SDSS DAS and CAS was not undertaken by one institution but by three (UofC, JHU and Fermilab). HathiTrust is another example of a consortial arrangement. At petabyte scales, such arrangements seem reasonable. Even smaller datasets may have their proper home in open access data repositories.

In any kind of shared arrangement which is managed in part by the Library and in part by other units of the University, or groups outside the University, the division of labor would have to be both rationalized with respect to who is best able to staff what function, and also coordinated so that operations function smoothly.

The Research Computing Center sees the Library as its partner for data archiving. It does not want to do data archiving itself, but foresees a need for data moving from its computing cluster to an archiving facility. It cannot quantify how much data will need to be archived, but others report that "only 30 percent of the researchers tick the box that says they care about saving the data." (PASIG, 2012) Thirty percent is a useful figure for planning only if one knows thirty percent of what. However, it does indicate that the archiving needs for data might be significantly less than the data produced.

It is clear that there is an important metadata component to data archiving. It is not simply getting researchers to fill out a form, but also checking that it is filled out completely and correctly. This is potentially a role for a data archivist. It is also about deciding what elements are needed in the first place. This is another role a metadata librarian might play.

Publication Repository

Functional Overview

Many universities have a need to publicize or locally publish their research output. The mechanism for doing so usually goes by the name of institutional repository, which is the term we will use in this section for the sake of consistency with current usage, though we are arguing that an institutional repository at the University of Chicago will have a broader scope, and that an institutional repository in the customary sense is really a publication repository, hence the title of this section.

Following is what the University of Michigan has to say about its institutional repository (IR), "Deep Blue." It should serve to outline just what functions a typical IR purports to support, and so help introduce the topic for discussion.

Deep Blue provides access to the work that makes the University of Michigan a leader in research, teaching, and creativity. By representing our faculty, staff, and student scholars as individuals and as members of communities, Deep Blue is where you will find and the Library preserves the best scholarly and artistic work done at Michigan.

What you get when you use Deep Blue:

Visibility
Making your work accessible via Deep Blue will ensure more of your peers can find it (in Google Scholar, for example) and will cite it.

Permanence
Deep Blue uses special technology that assures the stability of your work's location online, making the citation to it as reliable as a scholarly journal, while as accessible as any website. No broken links!

Comprehensiveness
Deep Blue supports a variety of formats, and we encourage you to deposit not just the finished work but related materials (including data, images, audio and video files, etc.) to create a "director's cut" that gives context to that work and promotes further scholarship.

Safe storage
This goes hand-in-hand with permanence. Deep Blue ensures that you only have to deposit the work once. From then on the Library takes care of backups, compatibility, and format issues. There are some technical limitations to the formats we can support indefinitely, but our commitment to preserving the integrity of your work exactly as you deposit it is 100%.

Control over access
Deep Blue allows you to limit who can see various aspects of your work for a given time, if you need to. This is difficult to do on a personal website without hiding the work completely.

Context
Beyond what is described above, Deep Blue provides context in two additional ways. First, UM is a destination for the best researchers and scholars, and Deep Blue places you in the larger context of the UM environment, side-by-side with the scholarly and artistic

```
contributions of your colleagues and students. Second, as other
universities, institutions, and organizations begin to provide this
service for their work as well, we will collaborate with them to
create discipline-specific services.

The University Library provides this service free to you as part of
the UM scholarly community. Further, Deep Blue is designed to meet not
only today's demands but also new ones as they evolve. It will
continue to grow and evolve to reflect current publishing needs and
norms identified by UM faculty, staff, students, and the communities
you form.

Your work: cited more, safe forever. Deep Blue makes it simple.
```

Deep Blue supports three functions: preservation; discovery and access; a work space that supports constraints on sharing.

Ideals, the IR of the University of Illinois, supports similar functions, but makes it clear that there is a strong preference for open access:

> IDEALS collects, disseminates, and provides persistent and reliable access to the research and scholarship of faculty, staff, and students at the University of Illinois at Urbana-Champaign. Faculty, staff, and graduate students can deposit their research and scholarship - unpublished and, in many cases, published - directly into IDEALS. Departments can use IDEALS to distribute their working papers, technical reports, or other research material.
>
> By default, items in IDEALS have no access restrictions, that is, they are openly and freely available via the World Wide Web. Open access to deposited items encourages a primary mission of IDEALS: the distribution, dissemination, promotion, and use of research and scholarship produced at UIUC. The University Library and CITES strongly encourage depositors not to place access restrictions on deposited items. However, there may be some situations when depositors need to restrict access to items in IDEALS. For example, a publisher may allow deposit of published articles into an institutional repository (such as IDEALS), but require an embargo of six months before the article may be made publicly accessible. Such a postprint might be deposited into IDEALS, but no access would be allowed for a period of six months. ... Access restrictions may be set to never expire or may be set to expire after a specified period of time. If access restrictions are necessary, the Library and CITES urge depositors to only put in place the minimum level of restriction necessary.

Technological Overview

Both Deep Blue and Ideals use DSpace for their repository management systems. An important part of repository management for an IR is the "ingest" function, or how data get into the system. In the DR, this is normally handled by repository administrators, because most of the deposits are very large and are not amenable to a "Click here" to upload solution: an IR typically deals with small items such as PDF

files representing an eprint or a small supporting dataset; the DR often deals with large, multi-item deposits. For example, the *Chicago Committee of Fifteen Records, 1909-1927*, consists of 26 volumes representing over 13,000 files and totalling 400 GB in size. Like Stanford's SDR, Chicago's DR is designed to handle "bulk operations." Turn-key IR systems such as DSpace and EPrints will provide an ingest function for simple document types such as PDF files (as does the DR).

DSpace and EPrints are not suitable if heavy customization is needed. For heavy customization, Fedora is the more flexible solution, but its drawback is the lack of a user interface out of the box, its greater complexity, and consequently a greater need for support.

DuraCloud DfR (DuraCloud for Research), which should be available later this year, is designed for the storage of large research datasets "in the cloud." This is another option to be evaluated, but one also needs to consider whether one is comfortable storing one's data in a cloud not administered by one's local institution or a trusted third-party. However, because DuraCloud DfR is being offered to the community as a solution for large research datasets, it does warrant consideration by this community.

One of the most impressive solutions today is the open-source SobekCM digital content management system, developed by the University of Florida Libraries with input from the University of Florida Digital Library Center. It has sophisticated ingest functions, portal and "skinning" options, and supports a number of document types and metadata formats: MARC, MODS, Dublin Core, VRA Core, Darwin Core, EAD. The administrative interface and overall functionality are very sophisticated.

Unlike DSpace, SobekCM is not simply an IR solution, but also a content or asset management system for digital collections. Unlike Fedora, it is not a toolkit, but a fully functioning product. Part of its ingest mechanism is to create METS objects from deposits; these can then be deposited downstream into a preservation repository. In Florida, this is DAITSS ("Dark Archive in the Sunshine State"), run by the Florida Center for Library Automation (FCLA) under the direction of Priscilla Caplan, Assistant Director for Digital Library Services. This mechanism allows a preservation archive to remain relatively dark, in the sense that it is not required to support sophisticated discovery and access functions because these functions are provided by the IR/digital collection front end.

SobekCM supports tailoring of the interface, including branding, to specific communities of users.

Staffing Overview

Repository management typically consists of 1 FTE librarian to manage front-facing end-user interaction, and up to 1.5 FTE technical staff for solutions such as Fedora. DSpace and Eprints installations may be managed with 1 FTE technical support staff. Digital Commons provides a remotely hosted IR ("software as a service") requiring no local technical staff support (though the cost of the service will include that support). Technical staff are programmers, not system administrators, though they are called "system administrators" in IR-speak. What that means is that they administer the IR system, not the underlying operating system and hardware. Adding that kind of

system administration support will add a fraction of an FTE to the cost of a locally hosted IR, and at least 0.5 FTE if there is a preservation archiving component, because preservation archiving requires ongoing monitoring activities that are specific to a preservation archive. Thus repository management may be expected to consist of at least 1 and as many as 3 FTE, depending on the solution.

Needs Overview

Here are some known needs at the University.

Law School faculty are interested in both long-term preservation and discovery and access for their publications. The desire is to integrate discovery and access with the existing Law School website.

The Computation Institute is building a site to publish a working papers series. It has expressed the need for services such as DOIs and archiving.

The Early China website provides discovery and access for its resources from its own purpose-built website. Its welcome page states:

> We hope that the site will be more than just a place to publicize the activities and publications of the SSEC, but that it can be a home to various types of research projects small and large that are more suitable to online rather than traditional print publication. (Early China: Welcome)

The Database of Early Chinese Manuscripts is hosted directly on this site.

The Divinity School site links to its publications on the goodreads site. The Humanities Division website points to the Research section of the UChicagoNews site from its "Faculty Research" link. The Department of Economics lists Selected Publications, which link directly to the sites various sites in which they are published.

Each of the preceding examples represents a different ad-hoc solution to the same problem. Each has its own drawbacks.

> Clicking on the Serial Number for an item in the Database of Early Chinese Manuscripts returns the following error message: "We are terribly sorry, but the URL you typed no longer exists. It might have been moved or deleted, or perhaps you mistyped it. We suggest searching the site." *The Serial Number link is not persistent*, and does not survive back-end reconfiguration of the website.
>
> The Divinity School points to publications on an external site designed for general readers to share information about books they are reading with other users of the site; *the site was temporarily down* the first time I checked.
>
> The Humanities Division points to a University-run site that is *not specifically focused on that discipline*.

> The Department of Economics points to articles on *commercial websites with constraints on access*. This represents a serious obstacle for prospective students if they do not have access from their home institutions.

In short, the same problem is being addressed in different ways by different parts of the University with varying degrees of failure; none of them can be said to be truly successful.

Next Steps

A good case might be made for a unified approach to showcasing faculty publications and other research, such as the technical reports of the Computation Institute, at the University. If one enters Swift Hall on the way to the coffee shop, the first thing one sees is a big showcase for recent faculty publications. It is possible that an electronic showcase might be combined with things like VIVO (or the equivalent). Such a solution should supplement the considerable investment in departmental and other organizational websites, rather than compete with them. A uniform approach might be a welcome service for many departments. An effective approach would have to be well planned, well designed and well coordinated. An IR with even a relatively modest initial scope such as this might serve as a visible and valuable first step in an effort to coordinate a response to related needs.

It is not at all clear whether there is a need for an all-inclusive, general-purpose, omnibus IR at the University of Chicago at this time. It is unlikely that a one-size-fits-all solution to the University's information long-term preservation and discovery and access needs will prove desirable or even feasible. For example, there will always be a need for a *work space* for the University Archives after materials are deposited, or for a researcher who wishes to share research with a limited audience; there will always be *long-term preservation archiving* needs for many fully processed materials regardless of how they are best discovered and accessed; there will always be demands for *customized front ends* serving specific needs. Therefore, it seems reasonable to expect that a good, workable solution will be both tiered and faceted, tiered to reflect different requirements for availability and response times (high vs. low demand), and faceted to reflect the kind of materials involved (text, image, audio, multimedia, restricted, unrestricted, etc.), the communities most likely to want to work with them (e.g., Oriental specialists; the general public), and how they want to work with them.

A unified solution to the University's long-term preservation and discovery and access needs is desirable, but it is conceivable that the unification will result in a coordinated approach to a set of problems and solutions, including partnerships with other institutions or consortia for some materials, with a service layer that masks implementation details from the end-user, rather than a single monolithic technical solution. ("Monolithic systems tend to serve poorly." [PASIG, 2012]) The California Digital Library expresses this as "a shift in emphasis from systems to services. Technical systems are inherently ephemeral, their useful lifespan being constantly encroached upon by disruptive technological change. Rather than pursuing the somewhat illusory goal of long-lived systems, curation goals are better served by concentrating on persistent services that can evolve and be easily reimplemented as

necessary while continuing to provide necessary function. This change in emphasis is best exemplified by a concomitant deprecation of the centrality of the curation repository as place. Rather than relying on a conceptually monolithic system as a locus, curation outcomes should be the product of a set of small, self-contained, loosely-coupled, and distributed services capable of operating on content in situ without a necessary precondition of being transferred to a central point for processing." (UC3 Curation Foundations, p. 2) A coordinated approach should seek to identify similarities and address them uniformly, so as not to duplicate the same functions, but at the same time seek to respect difference where difference is important to the end-user, or because the type of media or legal restrictions or policy decisions require it.

One way to move forward, once we have understood which pieces of a solution we have today, as well as how the functioning of those pieces can be improved to make them better, is to understand what new pieces are of immediate concern, and to add those to our planning. This is a stepwise approach as opposed to a piecemeal approach. A piecemeal approach would be to pick some so-called low-hanging fruit "and just do it," whether or not there is really much interest in having it, without a lot of planning or thought given to long-term maintenance, and without knowing how it fits into the bigger picture. A stepwise approach would be to identify real concerns, whether immediate or longer-term, and see what steps are required to address them systematically.

An analogy is solving a jigsaw puzzle. To begin the puzzle at one corner and proceed by adding pieces only from the starting point means that the solution will take an unnecessarily long time. To begin in the middle without knowing what the puzzle looks like is another way to take a long time, if one does not give up in the middle. The puzzle is solved most effectively by working on several sections simultaneously while keeping the end-result in view: the picture on the box. Several people can work together in parallel on the same puzzle in this way. First one needs to identify the elements of the big picture before turning attention to how to fit the individual pieces together.

Summary

The preceding discussion might be summarized as follows.

Staffing

The University of Chicago Library Digital Repository
: The immediate need is for 1 FTE Digital Accessions Specialist (exempt staff). Future needs may include a Digital Archivist or Electronic Records Librarian (1 FTE).

Research Data
: Immediate needs will include a metadata librarian function and a data archivist function (see preceding discussion *ad loc.*). Because the actual work of these positions and the time involved have yet to be adequately defined, the Library might begin by assigning existing staff to the responsibilities involved, though the case for 1 FTE metadata librarian can be made for reasons that go well beyond research data.

Publication Repository

The first steps here are planning and design, and a plan for ongoing maintenance. Staffing will require 1-3 FTE. In any scenario, 1 FTE librarian will have overall responsibility for managing the task.

Storage

Storage costs are harder to estimate for a number of reasons.

- We do not yet know the full extent of the problem, though we can take steps to estimate it.
- Cost is affected by the approach to the solution: wholly local; shared; outsourced.
- Cost is affected by whether compression is applied: some materials compress nicely; some do not; some would argue against compression entirely.
- Cost is affected by what we choose to keep, how long we choose to keep it, and how many copies we store.
- Cost is affected by the type of solution: disk-based solutions can be cost-effective for smaller problems; tape-based solutions are cost-effective for larger problems; sophisticated hierarchical storage management solutions add both complexity and cost.

Serge Goldstein, Associate CIO for Academic Services, Princeton University, has argued that "if you replace the storage every 4 years and the price drops 20%/yr, you can keep the data forever for twice the initial storage cost." (http://blog.dshr.org/2011/02/paying-for-long-term-storage.html) David Rosenthal, LOCKSS Chief Scientist, has challenged this, partly on the grounds that the exponential decrease of storage costs in past years will grind to a halt, because the physics of squeezing ever increasing amounts of data onto hard drives will hit a wall, with no research being done on marketable technologies to replace the current technology. However, IBM has recently announced a technological break-through that might break that barrier before it is reached, even taking into account the time from lab to market.

The initial cost of archiving approximately 30 TB in the DR at today's prices is approximately $15,000. Because the DR keeps two copies on disk, and one on tape for disaster recovery using IT Services TSM system, which is currently a free service, then, using Goldstein's algorithm, it would cost the Library $30,000 to archive 30 TB forever on disk. If we think, not in terms of forever, but of the next twenty years, then this is $1,500 per year. By way of comparison, licensing WebExpress is $10,000 per year, licensing MarkLogic is $5,336 per year, and the annual cost to mint DOIs will be $2,500.

Continuing to archive the Persepolis Fortification Archive past the first 3-5 years, which is being paid for by the Oriental Institute from grant funding, and assuming a final size of approximately 20 TB, would be even cheaper per TB, since the Library would not have to start paying for the first 3-5 years (i.e., before the end-of-life of the technology).

The current MOU for SDSS expires at the end of 2013. Goldstein's algorithm results

in $3,715 per year for 20 years to archive one copy of the DAS (75 TB) to disk uncompressed, or $1,486 compressed. Compression is assumed to be 40 percent, the result of a few tests on FITS files. ("Flexible Image Transport System (FITS) is a digital file format used to store, transmit, and manipulate scientific and other images. FITS is the most commonly used digital file format in astronomy." *Wikipedia*. To be certain about the compression that can be achieved, a random sampling of the data would need to be compressed.) The cost to archive the CAS would be less, since it is smaller than the DAS.

It is not clear at this time whether SDSS DR7 (the version currently supported by the University of Chicago) will need to be accessible after the MOU expires:

> "Our responsibility is to continue DR7 (and earlier) operations through Oct 31, 2013, after which time responsibility will shift (likely to JHU astrophysics for the active servers, they already are serving DR8), while the U of Chicago and JHU libraries will have the 'static archives' DAS and CAS for the long term." Brian Yanny, Fermilab, 8 Feb 2012

If not, then the cost to archive one copy of the DAS (75 TB) to tape for 20 years would be $3,050 total at today's prices for tape media. Again, the cost to archive the CAS would be less.

By way of contrast, a hierarchical storage management solution for between 100TB and 250TB of data (capacity depends on how well the data compress), with failover capacity to a remote location, consisting of low latency disk arrays and higher-latency tape libraries, could cost between $425K and $500K depending on the configuration. $100K would be for software licenses for the cluster and storage archive manager.

Large research datasets are another concern: some of these might be of a scale not easily accommodated without making special provisions. However, we do not yet know what we might be receiving, and we are as yet unable to predict what this University faculty might do.

For grant-funded research, we should not just "provide this service free to you as part of the scholarly community." Costs for long-term storage of data need to be and are today written into grants. This is part of data management planning. In the case of preserving and presenting small-scale faculty publications and supporting materials, we might just provide it as a free service.

Outreach

For any kind of digital research material, there is a need and an expectation for Library involvement to educate researchers with respect to best practices for metadata creation, file formats for digital files, and rights and permissions. There is also a need for the Library to better understand the type and scale of the problem that a repository is meant to solve, while at the same time alerting the campus that the Library is there to help solve it.

It is at this overarching level that a unified approach to the production, acquisition

and retention of digital research material at the University makes sense, to deal uniformly with the common problems associated with time, effort, cost, long-term preservation, and discovery and access. It is at the overarching level that the Library should begin its planning, because ad-hoc, piecemeal solutions do not scale, are often incomplete, overlapping or competing, and sometimes cease to be maintained. They should not be allowed to proliferate. Instead, a systematic, stepwise approach is called for.