Indiana University Digitization Master Plan

April 2014

Introduction & Motivation

In his State of the University address on October 1, 2013, Indiana University President Michael McRobbie emphasized that universities have a critical role to play in the preservation of knowledge. In keeping with this goal, President McRobbie announced two new digital initiatives, the Media Digitization and Preservation Initiative (MDPI) and a charter for an Indiana University Digitization Master Plan (DMP).

Both initiatives target the digital preservation of media. The MDPI, with total funding of $15 million over the next five years, is a production operation that commences in July 2014 digitizing time-based media (audio, video, and eventually film) owned by the university. The DMP is to look beyond time-based media and formulate a university-wide roadmap to "digitize and store in some form all of our existing collections judged by experts and scholars to be of lasting importance to research and scholarship, and to ensure the preservation of all new research and scholarship at IU that is born digital."[1]

President McRobbie envisioned that the DMP would be developed in conjunction with academic leadership and faculty as part of the president's 2020 Strategic Plan. He charged Brenda Johnson, University Dean of Libraries, Jorge José, Vice President for Research, and Brad Wheeler, Vice President for IT and CIO, to oversee development of a DMP. They in turn charged Professor Beth Plale, School of Informatics and Computing and director of the Data to Insight Center, and Dean David Lewis, IUPUI University Library, to co-chair and lead a broad engagement with many stakeholders across all IU campuses. The engagement was to ensure that a forward-looking DMP would be based on substantive input from IU's faculty, staff, students, administrators, and appropriate external constituencies.

Key Questions for Digitization

The assessments completed by Professor Plale and Dean Lewis answered five major questions in planning IU digitization efforts:

1. What Content Should be Considered for Digitization?

[1] President Michael A. McRobbie, *State of the University Address*, Indiana University, 1 October 2013.

To determine the scope of the content to be considered for the DMP, Professor Plale and Dean Lewis engaged with stakeholders across all IU campuses, gathering information about the collections held. A detailed account of the findings from the stakeholder engagement process can be found in the appendices. Select highlights include:

- Collections held by academic units at IU.
    1. 36 collection units completed online surveys.
    2. 10 units instead submitted memos describing their collections.
- The collections included more than 45.9M items.
- The collections are very diverse,
    1. Size varies from <1,000 to1M+ items
    2. Contents include photographs, books, documents, physical objects, data, etc.
    3. Time period spans 19th to 21st centuries
    4. Metadata about the items varies greatly and is often incomplete
    5. Interest for use within IU and/or beyond IU
    6. Ownership regarding copyright or public domain
- Primary challenges to data digitization include the large number of items in these collections and the difficulties involved in compiling metadata.

The stakeholder engagement process identified three primary categories of materials.

1) *Established collections.* Established collections are materials that are housed in organizational units that have collection management as part of their mission. These include libraries, museums, and other units such as the Kinsey Institute. More than 41M items were identified in established collections.

2) *Hidden collections.* Hidden collections are materials that are housed in various organizational units where there is no formal collection management responsibility. This includes non-digital collections housed in faculty, labs, departments, and centers.

3) *Born digital research or scholarly content.* More than 365,000 items were identified as "born digital," with the likelihood of more material accruing over time. Digital format content that is emerging from scholarship and research led by IU faculty and researchers and is of lasting value. This content may find a home in a discipline specific repository, may be associated with publications or not, or may be preserved at Indiana University. Future time-based video and audio collections that are born digital are included here.

The stakeholder conversations also identified two other categories of content:

4) *Born digital instructional content.* Increasingly content that is developed for or is captured as part of instructional activities is digital. This includes syllabi and assignments in the Oncourse Learning Management System, and audio/video capture of classroom presentations and lectures.

5) *University records:* The collections include documents, photographs, and other materials – both analog and digital – that capture the life and culture of Indiana University and its student activity through time.

2. Why Digitize Now?

The answer for *when* to digitize is a balance among answers to three sub-questions:

1. What is the risk of loss of an item due to deterioration, inadequate preservation, or theft?
2. When do the technologies to digitize and preserve reach viable economic price points?
3. What is the scholarly and/or reputational value of being an early mover for providing digital preservation and access?

IU's inventory of time-based media and subsequent reports[2] [FOOTNOTE LINKS HERE] documented extensive risk of loss for many time-based media holdings of audio, video, and film. Risk of loss for other items varies by age, media format (e.g., photograph, university records, lab artifacts) and if it is currently part of any managed collection. Those items at greater risk of loss and that are valuable are obviously of greater urgency for nearer-term action. For example, the Kinsey Institute has objects that are quite fragile, so 3D scanning would create a renewed opportunity for research. Interestingly, born digital items may be among those that are the most at risk due to a lack of intentional management and frequent technology changes.

The economics of digitization, preservation, and access will continue to evolve with technical innovation. By 2014, there is already considerable maturity in digitization technologies and work processes – including the ability for 3D digitization of physical collections of artifacts. For example, 3D scanning has advanced such that the quality versus cost tradeoff has surpassed the tipping point in favor of preservation. A 3D scanned version of a collection is more discoverable when applied to select museum or artistic artifacts, such as sculpture or the Slocum Puzzle Collection. Another example is the IU School of Dentistry which is required to keep its dental impressions of young children for a decade. Now that 3D scanning is available and affordable, electronic versions could serve as the replacement for the physical impressions. Software systems and governance structures for preservation and management of digital content are rapidly evolving with

[2] Indiana University Media Preservation Survey (2009) http://www.indiana.edu/~medpres/documents/iub_media_preservation_survey_FINALwww.pdf

new solutions appearing regularly.[3] Technological innovations continue to make digitization a more attainable and affordable reality.

Providing open and widely available access to IU content and collections may enhance IU's reputation, in addition to creating new forms of research, and enabling IU's digital assets to be more easily accessed. This could in turn draw students, researchers, and research dollars to IU. The availability of robust services and infrastructure to manage digital content could provide a competitive advantage to IU researchers in their pursuit of funding and to IU students in their pursuit of educational and employment opportunities. It may also be the case that being an early adopter – and thereby developing unique expertise and infrastructure – might position IU to provide services to other universities. MDPI clearly has this potential. Furthermore, funded efforts already exist nationwide for ingesting and curating born-digital content, and for semantic and metadata rich discovery.[4] IU could potentially capitalize in terms of reputation and funding if it undertakes a large-scale digitization process now.

The converse is also possible. If IU does not prioritize digitization, collections that might otherwise find a home at IU could migrate to other locations where a stewardship model exists. The International Forestry Resources and Institutions database (Workshop on Political Theory and Policy Analysis) is one example. IFRI was founded at IU, and moved to the University of Michigan in 2006.

The risk of loss, the advances in the technological innovations, and the potential for reputational gain all indicate that the time to undertake a large-scale digitization effort is the present.

3. How Will Priorities for Digitization be Set?

In his State of the University address, President McRobbie noted that the Digitization Master Plan should consider "collections judged by experts and scholars to be of lasting importance to research and scholarship." While there is no doubt that digitization activities will need to be prioritized, the best mechanisms for prioritizing are not clear and will vary within and among collections. Some of the collections are at risk because they are fragile or subject to uncertainty, in the latter case due to repatriation or being stewarded by an emeritus faculty member. Some collections have a technical component to them (e.g., technically enhanced data such as the Chymistry of Isaac Newton project), so are subject to obsolescence. Some collections, such as those that enhance the reputation of IU, could be strategic priorities from the outset. Finally, it is

[3] Digital Preservation Network (DPN), SHared Access Research Ecosystem (SHARE), and Academic Preservation Trust (APT)

[4] Research Data Alliance (rd-alliance.org), DataOne (www.dataone.org), Sustainable Environments Actionable Data (sead-data.net), Digital Public Library of America (DPLA), ClearingHouse for the Open Research of the United States (CHORUS)

widely understood that decisions about the future value of a collection require input from the national research community closest to the collection.

It is clear that scholarly merit as judged by peers will be important, and some criteria will be essential to arbitrate beyond just preferences. This might include:

1. Uniqueness
2. Scholarly importance of collection
3. Breadth of interest or extensibility of collection
4. Risk of obsolescence, decay or other loss of collection
5. Potential for external funding to support digitization

The availability of resources will also be a significant determining factor. Content that can attract grant or philanthropic support will, in most cases, be digitized first. It also seems likely that in many cases decisions will be made at the school or department level. The content to be digitized is so diverse that university-wide decision-making will be quite challenging. For example, what is more important, herbarium samples, historical IU Foundation tax records, or astronomical star surveys?

4. What Resources are Needed for a Robust Digitization Program?

The digitization of paper-based content could be accomplished with high-volume 2D scanners operated by dedicated staff. This content would include manuscripts, unique print publications, drawings, slides, pre-digital photographs, university forms and receipts, and special classes of physical objects like pressed, dried plant specimens and microscope slides. Different types of content require different handling, but the required techniques and equipment are understood.

Physical objects possessing more bulk – including fossils, medical specimens, and paintings – will require enterprise-level 3D scanners, which now exist but are very expensive and require trained operators. At the current time, only the IU Bloomington and IUPUI libraries possess equipment, staff, and repositories with modest capabilities of handling small-scale, collection-level projects in a reasonable time. However, it is acknowledged that these units will not scale to the extent necessary. Regional campuses possess rudimentary resources, and their librarians say they cannot handle additional projects without additional resources.

Outsourcing may be desirable in some cases and is already a common practice for some library projects. We will learn more about outsourcing from the MDPI experience with Memnon. Some units – like the Kinsey Institute, the Lilly Library, the Mathers Museum, and the Glenn Black Lab at IU Bloomington, as well as the Medical School at IUPUI – face donor restrictions, HIPAA regulations, or insurance liabilities that require that digitization be

performed within the unit itself. Regional campuses collaborating with community organizations on important collections may also prefer to have the digitization work performed locally.

5. How will Digitized Collections be Maintained and Accessed?

Access to and preservation of digital collections requires the following:

1. A robust and flexible repository infrastructure with a very large storage capacity
2. Good metadata to facilitate discovery
3. Clear copyright, intellectual property, privacy, and access policies
4. A strategy for the long-term preservation of the digital content

Every collection requires an appropriate repository solution, but there should be generalizable workflow solutions that can be broadly applied. Our librarians and IT professionals have generally tackled these issues in a piecemeal way, but the DMP initiative poses the problem on an extremely large scale. This problem is being worked on by a number of universities nationally, and IU is a collaborator on many of these projects. IU's current solutions, ScholarWorks and the Scholarly Data Archive (SDA), are adequate. Nonetheless, they will need significant enhancement to meet the needs of DMP. The amount of storage required will be large. Estimates project that the MDPI will require 10 Petabytes by 2020, and the Medical School alone may need another 10 to 15 petabytes. It is unclear exactly how much additional capacity a fully implemented DMP will require, but it will challenge even IU's capacities, which are significant.

Access, discovery, and reuse of data depend on availability of metadata. Faculty members are reluctant to spend time manually adding metadata, and data curation specialists are in short supply and often lack sufficient area expertise. Tools and ingest processes that enable handoffs and collaboration between researcher and data curation specialist are needed. It will be necessary to determine the ownership of all of the material that will be deposited in a DMP repository (or repositories), and articulate any permissions or restrictions that may govern access to materials. Researchers are generally concerned about controlling access to their unpublished materials. If rights issues are managed correctly, researchers could have access to sensitive materials in protected data enclaves. The existence of such enclaves could also increase the competitiveness of IU funding proposals.

The Digital Preservation Network (DPN)[5] is a rapidly developing national solution to the *long-term* preservation of digital content., and IU is a co-founder and partner in this universities-led project. DPN appears to be a good solution for preservation of high value content, and the further development of DPN's business model will provide greater insight

[5] http://www.dpn.org

over time regarding its costs. To be clear, DPN is about long-term *preservation* even beyond catastrophic events – it will *not* enable access.

6. What are the Possible Sources of Funding for Implementation?

President McRobbie's call to "digitize and store in some form all of our existing collections judged by experts and scholars to be of lasting importance to research and scholarship, and to ensure the preservation of all new research and scholarship at IU that is born digital" is a bold one. A strategy for funding such an initiative will be multi-faceted. An IU strategy that curates and cultivates select digital data collections in science, informatics, and medicine has the potential to increase research funding to the university. This we heard clearly from groups such as the Indiana Clinical and Translational Sciences Institute (CTSI). Reputational gain to the university from digital assets that are broadly available could manifest itself in both increased student interest in IU and alumni giving. It will, however, be difficult to deploy the necessary resources to meet the call without a concerted effort by the senior leadership.

Costs are non-trivial in the aggregate. For example, the Lilly Library estimates an initial cost between $25 million and $44 million just to digitize (not fund access) for its collections that can and should be digitized. The software, hardware, and human resources needed to provide ongoing access to and preservation of the digitized content could also be millions per year, but those would become more efficient with scale.

Beyond access, the cost of *long-term* preservation of digital content is also not yet clear. One highly promising solution to the problem is the previously mentioned Digital Preservation Network. DPN is in its early days, but it has suggested that a one-time $5,000 per Terabyte charge for 20 years is a one cost-modeling scenario. Such estimates will likely change over time, but they provide a sobering example of the cost of digital preservation – just like the capital and operational costs of physical preservation in buildings. If the MDPI produces 10 Petabytes of content as is estimated, this would mean a one-time cost of $50 million to deposit this content into the Digital Preservation Network.

It is useful to think of the resources requirements in three categories:

1. Retrospective Conversion — This is a one-time cost of converting an artifact to its digital form. The bulk of the estimated costs from the Lilly Library fall into this category. In some cases, the content is fragile and risks permanent loss if not digitized quickly, but in most other cases the content is in a format, like most paper, that is relatively stable. It is likely that a retrospective conversion funding strategy that combines some base funding for the steady advancement of the goal with an aggressive opportunistic strategy of pursuing external funding through grants and philanthropy would be appropriate.

2. Services — A variety of software and staff services will be required to digitize, provide access to, and preserve existing collections. Software workflows for the intake and curation of born digital data need development. Data curators are needed to ensure high quality digital collections. Funding these technical and library science positions will require a combination of new funding and reallocation of existing resources. A new funding source for born digital data is emerging nationally, as federal funding agencies grow more receptive to data management costs being called out as line item funding.

   Most of the services will be provided by the IU libraries on all campuses and by UITS. This is where both reallocation and some new investment in positions will be necessary. Where new investments are required, assessments should be the preferred mechanism for providing the funding. In cases where exceptional services are required for particular projects, fee-based services would be appropriate.

3. Technology Infrastructure — An initial investment in technology infrastructure, particularly to create a robust repository and large-scale storage system, may be required, but going forward the technology needs to be base funded to accommodate appropriate lifecycle replacement. The mechanism for funding this infrastructure should be assessments, though some reallocation of research overhead might be appropriate. IU has been heavily involved in open and community source software projects, and it is likely that at least some of the software infrastructure required will be of this sort. In these cases, the IU investment may be primarily in contributed staff.

While not a sizeable component of a funding solution, our study found surprising IU community enthusiasm from librarians, faculty, and collection holders wanting to contribute their own effort to collection organization and digitization. We believe there is latent capacity for advancing the university's digitization goals with a smaller investment in mobile digitization equipment, and recommendations for and licensing of collections management software.

We heard of creative ways of funding digitization and preservation efforts that, for instance, sell subscriptions to a magazine that highlights recently digitized pieces of a unique collection. The magazine would need a theme to create sufficient interest for a long-term subscription, but online publishing reduces the overall publishing costs and makes theme publications easier.

We learned of the rich cultural life captured in numerous photos and documents at the campuses. This rich resource, when digitized, could be used to enhance the value of membership for an IU alum by means of social media and online engagement with IU and the IU experience, thereby increasing the potential for giving by alumni.

Recommendations

This Digitization Master Plan is a first step in formulating an executable digitization strategy for Indiana University. This work has enabled many conversations across the university to surface areas of similar need and opportunity for digitization. The recommendations that follow are based on those conversations, conducted in the first quarter of 2014, and insights from other national and international trends.

The path for IU will be an ongoing work-in-process. The recommendations that follow provide a way to get started and address three enabling first steps:

- *IU should make further development of the IU Digitization Master Plan the authoritative roadmap for digitization efforts for the university.*
- *IU should develop the technical and service infrastructure required to support digitization at scale.*
- *IU should develop services required to support a robust digitization program.*

More specific recommendations follow:

Recommendation 1: IU should develop a DMP framework including roles, responsibility, and authority to steer digitization for the university.

1.1 IU should appoint a senior level digitization "tsar" to lead and coordinate IU's digitization efforts.

Developing and implementing a comprehensive, forward-thinking digitization effort is too large and important a task to leave as an "add-on" to the other responsibilities or hope for multi-school/campus coordination. The university should establish a charge and authority to further develop and act on the DMP. The DMP tsar should report to appropriate level senior officers of the university to be able to exert influence and authority for DMP matters.

1.2 IU should create a detailed digitization plan with a full inventory of analog and digital content, as well as strategies and tactics for making significant progress on digitization and providing the infrastructure and services needed for development and support.

This DMP provides a start on the important questions and answers for a university-wide digitization effort, but further planning is required to avoid uncoordinated and haphazard investments.

1.3 The university should establish funding, consistent with its ambitions, for digitization and preservation. This may include fundraising efforts for digitizing some of the institution's most significant collections in the upcoming fundraising campaign.

Digitization efforts will work within the "Reality Triangle" of project management where scope of digitization and time are determined by resources. In the absence of university funding to further develop and operationalize a DMP, many uncoordinated and unsustainable efforts are quite likely to arise among collections and units of the university. Some ongoing funding for DMP is the best way to provide a coordination point and capability to ensure institutional digitization efforts achieve their longer-term goals.

1.4 University and campuses should identify the most important collections and consider direct funding for digitization. This should include established collections, hidden collections, and born digital content.

Recommendation 2: IU should develop the technical and service infrastructure required to support digitization at scale.

2.1 IU should establish a robust enterprise scale digital repository with a layered architecture that supports extensions for custom discovery and access interfaces with collection branding.

This capacity is the core infrastructure required for housing and providing access to digitized collections. UITS, IU Bloomington libraries, and IUPUI libraries, in collaboration with partners from other universities should lead this effort.

2.2 IU should develop a storage infrastructure that can accommodate its digital collections. We believe this will be on the order of 25 petabytes by 2020.

While it has substantial capacity, the SDA has neither the adequate structure nor the capacity that will be required. The cost of the storage portion of the recommended infrastructure is estimated to be $10 to $12 million per year. Once a source of funding is identified, UITS should develop/implement the storage solution.

2.3 IU should develop the capacity to digitize collections in standard formats and should develop relationships with appropriate vendors who can provide digitization services.

IU will require a variety of digitization capacities. In some cases, outsourcing will be the best alternative; in others, digitization equipment would best be supplied locally; in select cases, this will best be done at the collection location. The libraries at IU Bloomington and IUPUI have established centers for digitization, and this capacity exists in other locations as well.

2.4 IU should continue to develop the Digital Preservation Network (DPN) for the secure long-term preservation of its digital collections.

The DPN provides a secure, long-term solution for the collation and preservation of digital collections. Should the Digital Preservation Network not develop as expected or should its

economic model preclude use at scale by IU, other alternatives need to be established. The Digital Preservation Network is still in development, and it will likely be at least a year before its capacities and cost will be understood. While a robust long-term preservation mechanism for the preservation of digital content is required, no near-term decisions are required.

2.5 IU should work to integrate systems in a seamless manner and adopt standards for capture of metadata in order to facilitate content collation and discovery in the future.

Many of the current systems used to manage university business activities do not interface with the university archives. For example, the current systems used by university marketing departments to manage photos and video do not capture sufficient metadata or deposit their content into the university archives. The university archivists and the managers of these systems should be charged with developing appropriate structures to assure that the university's history is preserved. Archivists at IU Bloomington and IUPUI, as well as senior leadership from appropriate university offices can lead this effort.

Recommendation 3: IU should develop services required to support a robust digitization program.

3.1 IU should develop staff expertise locally to support IU researchers in adequately preparing research data such that it can be interoperable and discoverable in the future.

Born digital data that is given to IU for stewardship and preservation has degrees of variety and complexity that affect its ingest, access, and discovery. IU risks falling behind its neighboring research universities in its ability to handle digital research data. The IU digitization "tsar" should lead the effort to develop staff expertise, so specialized knowledge can be shared across all campuses of the university.

3.2 IU should develop legal expertise devoted specifically to digital content, including classroom materials.

IU has limited legal resources to support digital efforts where rights issues are often complex. It may be best to house this expertise in the libraries, as the position will need to provide education and advice as much as legal opinions. The position should be modeled on the one previously held by Kenny Crews at IUPUI. This recommendation may require coordination by the Office of the Vice President and General Counsel, Office of the Vice President for Information Technology (OVPIT), and university libraries.