

Research Article

Open Access

Peter X. Zhou*

Towards a Sustainable Infrastructure for the Preservation of Cultural Heritage and Digital Scholarship

<https://doi.org/10.2478/dim-2020-0052>

received October 18, 2020; accepted November 16, 2020.

Abstract: The digital lifecycle encompasses definitive processes for data curation and management, long-term preservation, and dissemination, all of which are key building blocks in the development of a digital library. Maintaining a complete digital lifecycle workflow is vital to the preservation of digital cultural heritage and digital scholarship. This paper considers digital lifecycle programs for digital libraries, noting similarities between the digital and print lifecycles and referring to the example of the Digital Dunhuang project. Only through a systematic and sustainable digital lifecycle program can platforms for cross-disciplinary research and repositories for large aggregations of digital content be built. Moreover, advancing digital lifecycle development will ensure that knowledge and scholarship created in the digital age will have the same chances for survival that print-and-paper scholarship has had for centuries. It will also ensure that digital library users will have effective access to aggregated content across different domains and platforms.

Keywords: data management systems, arts and humanities, data storage and integration, document integration and text processing, digital lifecycle, digital preservation, digital asset management, Digital Dunhuang, preservation of world heritage

1 Theoretical Framework of the Digital Lifecycle and Digital Preservation Processes for Cultural Heritage and Digital Scholarship

The continued existence of cultural heritage sites around the world is under serious threat from human activity, environmental change, natural disasters, and the limited effectiveness of conservation efforts. At the same time, the rapid development of digital technologies and digitization capability has provided unprecedented opportunity to digitally record, reproduce, preserve, and disseminate this heritage. Taking advantage of this opportunity, museums, libraries, and other similar institutions are actively engaged in exploring ways to preserve world culture and history through digital initiatives.

The essential first step in these initiatives, invariably, is construction of a sustainable infrastructure that will support the complete digital lifecycle, including digital asset management, digital preservation, and digital publishing.

Before laying the infrastructure, however, institutions must determine the objectives and desiderata that will define the design and implementation of the digital enterprise. For example:

- Balancing scalability with size and flexibility while staying within the confines of international standards, best practices, and overall compatibility.
- Being able to manipulate massive amounts of content (text, images and visuals, numerical data).
- Enabling rapid discovery and information retrieval.
- Creating immersive experiences which incorporate virtual reality or 3D modeling. Increasingly, immersive experiences are finding favor in large digital heritage projects for their ability to maximize user experience.
- Offering vivid data visualization, such as charts and graphs.

*Corresponding author: Peter X. Zhou, C. V. Starr East Asian Library, University of California, Berkeley, CA, USA. Email: pxzhou@berkeley.edu

- Providing infinite reuse of content (text mining), a powerful tool unknown in the world of print.
- Furnishing dynamic publishing and sharing platforms for the benefit of both the academic and nonacademic user.
- Aggregating content through linking, both internally and externally, in coordination with other content providers.
- Using AI (artificial intelligence) technology to facilitate machine interpretation and machine learning.

Among the challenges in developing and curating digital content for posterity is the issue of sustainability. At present, no one can say with certainty that what we digitally create today will be extant in a hundred years, or fifty, or even twenty-five. Moreover, special attention has to be given to orphan works, whose authors no longer exist or who no longer control the content they created — unique books, manuscripts, paintings, and other artifacts. The question arises as to who will digitally preserve these works in perpetuity. The enormity of the challenge is multiplied when we take into consideration the work of organizing and integrating content from different corners of the globe, in different languages and with different functionalities. Sustainability will require the establishment, implementation, and maintenance of an effective digital lifecycle program (DLP) in compliance with international standards and best practices.

The digital lifecycle encompasses definitive processes for content curation and management, long-term preservation, and dissemination, all of which are key building blocks in the development of a digital library. This is preceded by content creation — the selection, production, harvesting, and digital conversion of print or analog content. Created content is then organized by analysis, integration, aggregation (through linking), or interpretation, and recorded by creating metadata, tagging, and cataloging. Subsequent to this, content must be stored, by making derivatives or duplicates, verifying formats, performing checksums, data repair, or data migration. At this point, content may be pushed out to users, whose experience may be enhanced through facilitated navigation, discovery, and rights management.

While the digital lifecycle is merely decades old, China's print lifecycle dates back over a thousand years; a cursory examination of both of these processes reveals that the digital lifecycle actually repeats certain essential components of the print lifecycle, despite differences in the formats and platforms used now and in the age of the woodblock.

Paper was invented in China nearly two thousand years ago. While the origins of woodblock printing in China are obscure, it was in use at least as early as the ninth century and remained the most common method of printing text and images throughout East Asia into modern times (Needham, 1994). Movable-type technology, although invented in China around the tenth century and used quite successfully in Korea and Japan, failed to diminish the popularity of the woodblock, which receded only with the importation of Western print technology (which, ironically, owes its origins to Chinese movable-type printing) (Gies & Gies, 1994).

With the development and spread of printing in China came the establishment of processes and standards to ensure quality and the efficient use of published content. These standards were reinforced by commercial and institutional forces that worked together to ensure continuity in the print tradition. Examples of these enduring standards include leaf size, page layout, and font design, which for centuries have been implemented under the fine craftsmanship of the most celebrated calligraphers and printers.

This continuity did not preclude innovation and technological advance. The Ming dynasty (1368–1644), a period of stability and prosperity, witnessed the development of woodblock illustration, color printing, printing from metal plates, and further experimentation with movable type. The rise of commercial printing, moreover, led to an unprecedented volume, scope, and variety of printed material.

Collecting and curation developed at roughly the same time as the invention of paper. The first centralized library in China, with standard procedures and practices, as well as the first classification scheme for the organization of books, date to the Han dynasty, long before the introduction of woodblock printing. As collecting became more widespread, preservation and conservation entered the print lifecycle. The culmination of the final component of the print lifecycle, circulation, and dissemination, occurred in the twentieth century, when Western-style libraries began to serve the general public and academics alike.

The print lifecycle has endured for more than a millennium and is still vigorous, while the digital lifecycle only has a history of merely forty years or so. It is consequently reasonable to assume that as it develops, the digital lifecycle will follow the patterns and constraints of the print lifecycle, and that its optimal utility will have to be ensured through a careful examination of the processes, standards, and general practices, similar to the

manner in which the optimal utility of the print lifecycle was ensured.

Echoing the print lifecycle workflow, the digital lifecycle workflow moves from selection to creation, description, management, preservation, discovery, use, and reuse. All these processes can fit into the following functional stacks.

Workflow starts with data creation and ingestion, when the content is added to a digital object repository. The content is then cataloged, creating metadata, some possibly even at the point of ingestion. Data processing follows, with the creation of derivatives, registration of formats, and reconciliation as needed through media manager and transmission procedures. Next, all data,

metadata, and format registration are transferred to a permanent storage site, whether in the cloud and on an institution's proprietary space. Content can then be moved to a public-facing platform through application programming interface and linked data presentations. The public-facing tier contains collection management, search, and discovery portals, creating a link between the home institution's content and that of other providers via open space and open data solutions. Data content is managed by administrative protocols that direct and shape user and community services. In a ubiquitous world, content is accessible in various ways, including mobile systems. These processes are captured in the following chart.

This DLP program workflow is compatible with the widely accepted standard of OAIS (Open Archival Information System), which is a reference model for organization of people and systems to preserve information and make it available to a designated community. It is considered the optimal standard for creating and maintaining digital repositories over long periods of time. The model aims at ensuring the content can survive the impact of evolving technologies, new media, and data formats as well as the changing user communities.

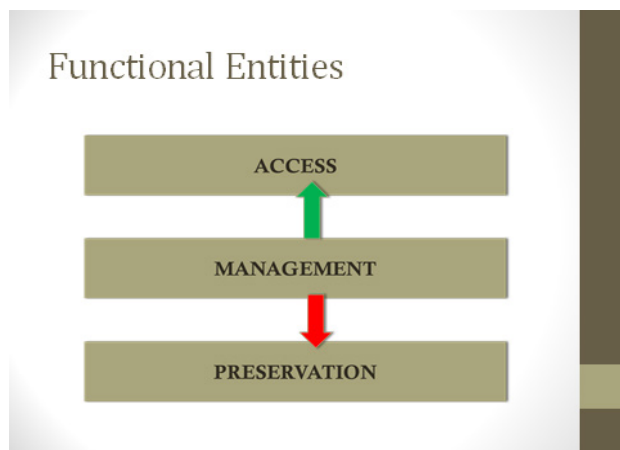


Figure 1. Functional stacks for the digital lifecycle.

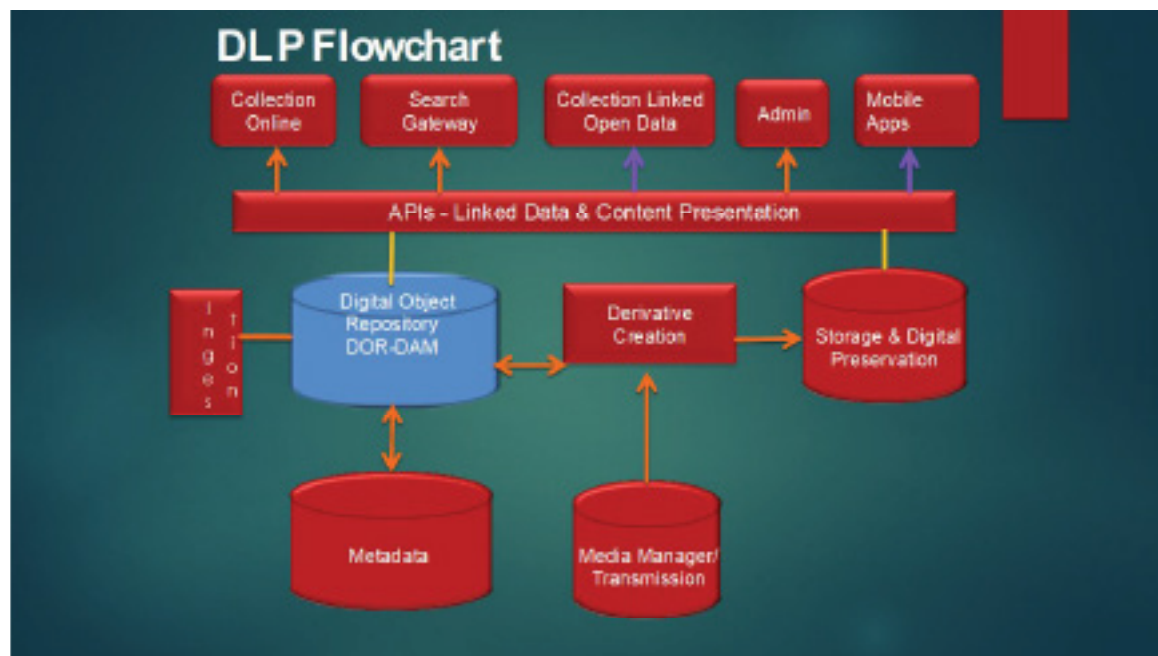


Figure 2. Digital lifecycle program workflow.

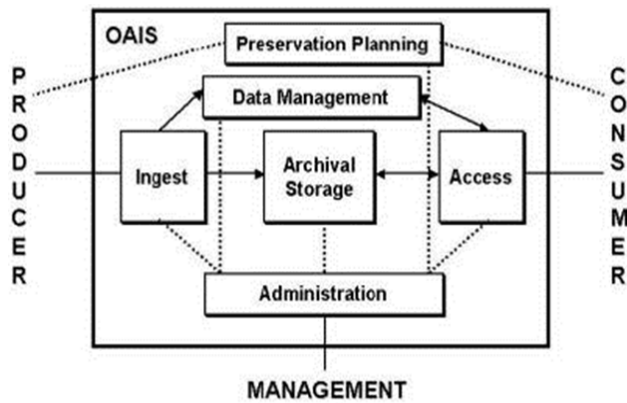


Figure 3. OAIS (Lavoie, 2000)

2 Digital Dunhuang: Solutions for Developing and Maintaining Digital Content to Enable Digital Scholarship

After examining the theoretical framework for a digital lifecycle program, it should be instructive to consider how the principles and processes of that framework have been applied in the Digital Dunhuang, one of the largest digital heritage projects in existence today.¹

Collectively, the 812 caves commonly referred to in the West as “Dunhuang” comprise a treasure house of art in Western China’s Gansu province. The major sites include the Mogao Caves (a UNESCO World Heritage site) and the Western Caves of the Thousand Buddhas, both in Dunhuang county, and the Yulin Caves in nearby Guazhou county. Aesthetically, commonalities in the artwork preserved in the caves have led art historians to identify a “Dunhuang style.” Over the last century, the caves’ excavation, study, and conservation have attracted global interest and inspired a new field of academic endeavor, which is known as Dunhuang studies.

Construction of the caves began as early as the fourth century and continued into the fourteenth. The 492 caves of the Mogao site, which has received the most attention, present a remarkable concentration of works to be preserved: 45,000 square meters of murals and 2,000 painted reliefs and sculptures. First revealed to scholars in 1900, Mogao’s Cave 17 originally contained nearly 50,000 early manuscripts, silk banners and paintings, embroideries, and other textiles predating the year 1000. These materials, now in collections scattered across

China, Japan, Europe, and North America, are loosely referred to as the Dunhuang Archive.

The complete digital lifecycle can be seen in specific cultural heritage projects, including Digital Dunhuang.² Digital Dunhuang enables long-term preservation of cultural heritage of inestimable value, while providing a platform for sharing all digital assets generated in the act of preservation. With the support of the Mellon Foundation, the Dunhuang Academy has explored construction of a permanent repository for all Digital Dunhuang assets. The academy concluded that the only way to ensure preservation of their digital assets for future generations is to integrate all content created in the past, created now, and to be created in the future, in one large digital repository that will combine the functions of perpetual preservation, effective digital asset management operations, and easy access in a systematic way.

The Digital Dunhuang project incorporates the following four key components:

- Capturing high-fidelity images;
- Restoring virtual reality through geospatial information, laser scanning, and 3D modeling;
- Creating an immersive experience, through either virtual exhibits or actual-size reproductions; and
- Harvesting and integrating massive amounts of information through micro-environmental monitoring.

An enormous amount of data has been created under the auspices of the Digital Dunhuang project, including, to date, photographs of over 200 caves and over 14,000 square meters of digitally captured murals. Together they total over 300 terabytes of data, over 87 cases in complete QuickTime virtual reality, and over twenty years of climate monitoring data for 87 caves. And the collection is growing continuously.

In addition to print and analog data which is created historically, much of this data was created and ingested daily inside the caves through photography and other means such as laser survey and scanning. According to the Digital Dunhuang lifecycle program, data feed of differing content is uploaded into a digital asset management system, or DAM, and integrated in a digital repository. The content includes a rich variety of raw images, QuickTime virtual reality, historical photos, videos, digitized texts, scanned manuscripts, sculptural artifacts, freehand reproductions, microfilm, interactive visuals, and conservation data.

¹ This case study is based on Tadic and Zhou (2012).

² For an earlier version of this DLP program, see Zhou (2011).

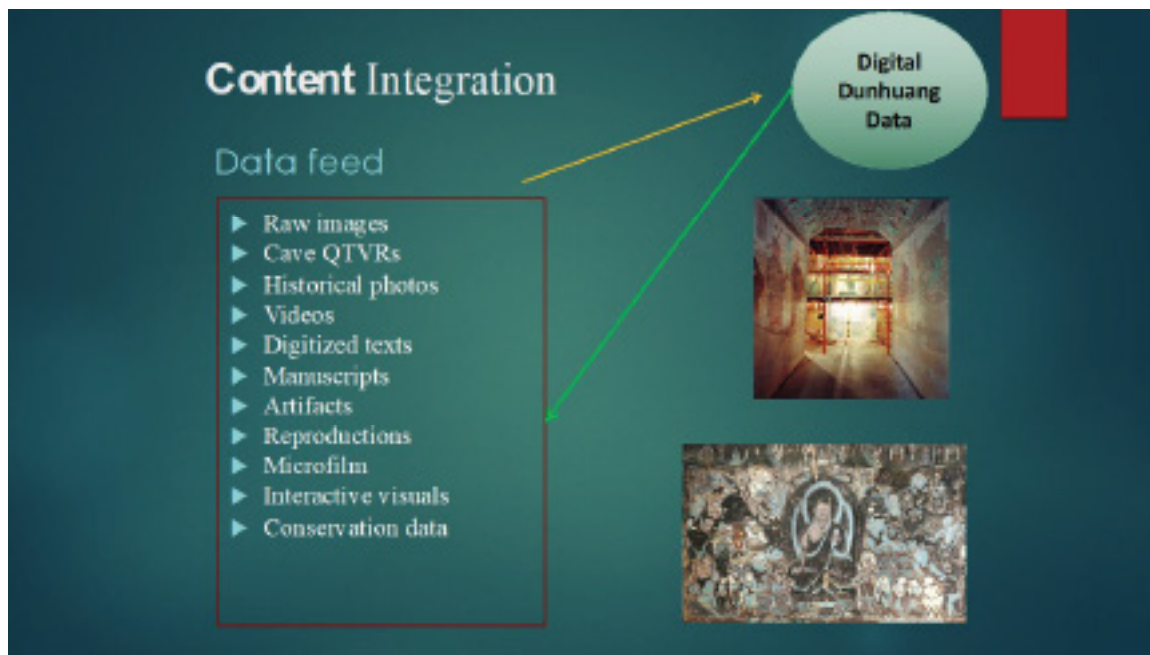


Figure 4. Digital Dunhuang data types.

Managing a diversified set of data types and formats requires a powerful and comprehensive DAM that can ingest data, group, classify, label, and organize collection files systematically under the broad framework of a digital repository. The repository facilitates DAM operations, and provides search engine and discovery tools through keywords, facets, thesauri, linked data, and terms from within metadata and texts. It also aims at interchange and sharing of non-textual data, such as images and visuals, in an IIIF (International Image Interoperability Framework) -compliant system, as shown above.

The Digital Dunhuang DAMS facilitates asset creation and cataloging of image, video, and text files; manages high-resolution master files and original documents; supports version control during the lifecycle of the digital assets; and tracks digital preservation actions and pushes metadata and content to the delivery platform.

The Digital Dunhuang program performs digital preservation actions on-site, while staff track and, at times automate, actions through the DAM. Managed digital preservation actions include creating checksums, validating files, and extracting technical metadata upon ingest; monitoring file format obsolescence; migrating file formats; and tracking copied files to LTO tape.

Digital preservation is a managed process (Tadic & Zhou, 2012). It focuses on preserving the file, not the medium it has been stored on. As formats become obsolete, the file format must be migrated forward, as well as to a new physical carrier. Because preserving digital content is a primary concern

at the Dunhuang Research Academy, digital preservation workflow is planned out and followed meticulously. All actions are tracked through the DAM, including:

- **Checksums.** Checksum algorithms are unique alphanumeric strings that check file integrity. When a file is ingested into the DAM, a checksum is run and the resulting string saved in the DAM. When the file is next transmitted — whether downloaded, uploaded, checked, or migrated — the checksum from the new inspection should match the original exactly. A failed match indicates file corruption or bit loss. Until recently, MD5 was commonly in checksum, but it is known now that it can be corrupted; consequently, many organizations, including U.S. agencies, have shifted to the more secure SHA-2 algorithm.
- **File verification.** Files should be verified to determine their exact file format and version. Java-based open-source verification tools such as JHOVE and JHOVE2 support many of the file formats created by the academy.
- **Preservation metadata.** This describes a digital object and its metadata, and in which applications, environments, the object, and data were created and are known to work. This is crucial information for file preservation. Much of the data can be automatically extracted from the file and its storage environment.
- **File locations.** The DAM should track the location of every file, including the file's folder or directory structure, as appropriate. All DVDs, HDDs, servers, LTO tapes, and other media must be included.

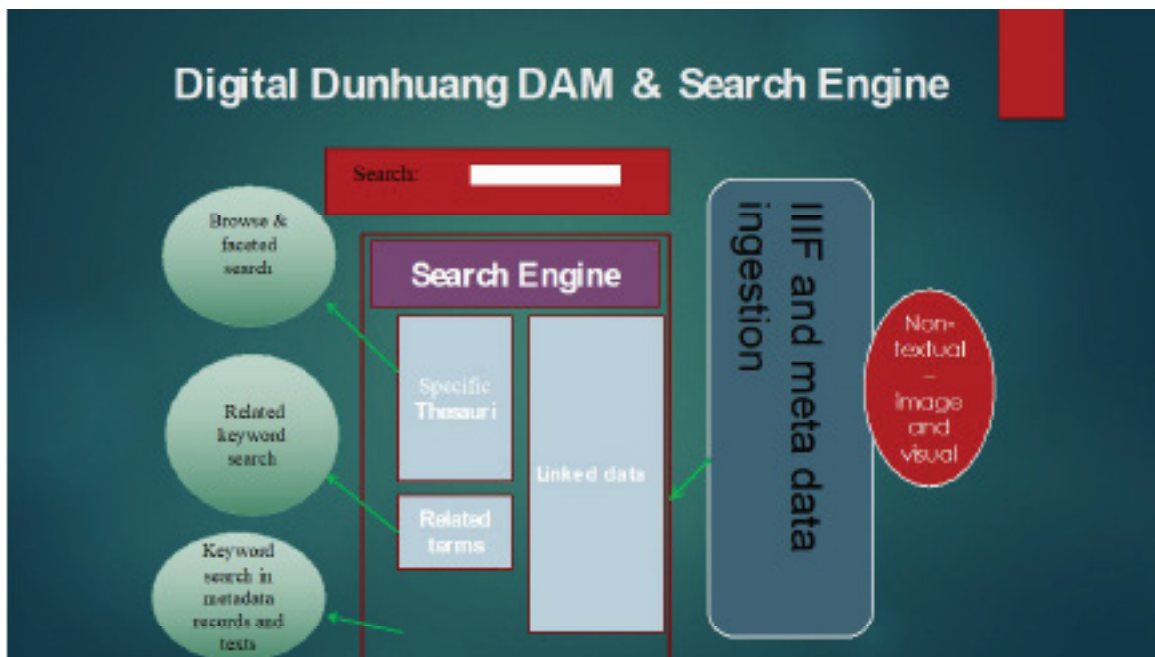


Figure 5. Digital Dunhuang DAMS and search engine.



Figure 6. Digital Dunhuang preservation workflow.

- Forward migrations. When files are rechecked and copied to a new medium (or transcoded to another file format), this action must be tracked. The file record should indicate the exact original source (whether a digital video tape or a file), what software and hardware was used to transcode it, and where the file is currently stored. Old files should not be discarded when files are transcoded, in case problems with the new file require that the work be redone; however, saving new and old files increases storage requirements. Using open, rather than proprietary, formats helps decrease format-forward migrations.



Figures 7. Pure land: Inside the Mogao Grottoes at Dunhuang. Images courtesy of Sarah Kenderdine, 2012.

It is important to note that each and every digital heritage project is unique. There are specific issues that affect the integrity of the Digital Dunhuang content as well as its preservation. One example is the detection of obsolete formats. Much of the Digital Dunhuang content has been converted from analog content created over the span of a century. While older formats must be searched out and converted, scholars and others familiar with the content must be consulted to ensure its integrity. Metadata must be updated to reflect preservation treatments undertaken. Content must be safely stored in multiple locations, preferably offline, to ensure data security.

The last stage in the Dunhuang project's digital lifecycle program calls for public engagement. There are various ways this can be done. In 2012, Sarah Kenderdine, a recognized expert in digital museology, used digital files to create an immersive environment that allowed museum-goers to virtually experience the Dunhuang caves in all their three-dimensionality.

In recent years, the Dunhuang Research Academy has been incorporating actual-size replicas, high-fidelity imaging, and 3D modeling in those exhibits which are forwarded to venues in China and the United States. Jointly, these tools and techniques allow museum-goers to have the virtual experience of looking at a Dunhuang mural or walking through a painted cave. The Dunhuang Academy has also mounted a number of mural images on its open-access platform, where they can be viewed by people at any time, from anywhere.

Lately, the academy has also introduced innovative approaches to content processing and management that

incorporate AI technology. One application uses facial recognition technology to search and link digital images of artwork to historical events and figures which are portrayed in the images and delivered by machine-assisted interpretation. AI and machine-assisted interpretation can both be expected to play a bigger role in Digital Dunhuang in the future.

3 Sustaining Digital Cultural Heritage and Digital Scholarship for Posterity

What is digital preservation? The American Library Association defines it in this way:

Digital preservation combines policies, strategies and actions to ensure the accurate rendering of authenticated content over time, regardless of the challenges of media failure and technological change. Digital preservation applies to both born digital and reformatted content. (Association for Library Collections and Technical Services, 2007)

In other words, digital preservation makes available a general conceptual framework for creating a sustainable infrastructure for digital heritage and scholarship projects. Unfortunately, many confuse digital preservation with the work of digitization or the publication of digital content.

Consider the misconception that digitization is the equivalent of digital preservation, an idea that is common even among seasoned professionals and overseers of conservation and the preservation of cultural content. But converting a print volume to digital format does not ensure its existence into the indefinite future. In fact, the guaranteed life span of a digital object is only a few years, after which there is no guarantee that it will maintain its original shape. This requires elaborate digital preservation work.

A variation on this theme is that digital preservation means making copies. But again, the simple creation of storage copies now does not necessarily translate into a true and useable copy in the future.

Another misconception is that digital content will last forever. This is little more than wishful thinking. How much digital material created twenty or thirty years ago is still here today? A lot has disappeared, gone bad, or is simply unrecoverable.

And then there's denial: "Digital Preservation is not our business." Again, this kind of thinking is common even among library and museum professionals, who should be building real and sustainable digital

preservation programs. In China, for example, we have yet to see any such programs created by the Chinese Library Association, the National Library of China, CALIS (China Academic Library and Information System), or any of the national museums in China, although at present the time is appropriate to begin the work of systematic and serious digital preservation. The National Science Library of the Chinese Academy of Sciences is leading an effort to preserve scientific journal publications and has achieved good results. However, the question arises as to whether anyone in the corporate world — for instance, Baidu, Alibaba, or Tencent — has created a digital preservation program for libraries or museums. The institutions holding and guarding cultural heritage need platforms to preserve their vast digital collections, but these platforms seem to be lacking in China.

Akin to denial is the notion of leaving the work for someone else to do. This too is common among library leaders in the U.S., although they do have an excuse: digital preservation is complex, expensive, and difficult to execute. Some institutions consequently prefer to rely on the Library of Congress, HathiTrust, or corporations like Google and Amazon to undertake the work of digital preservation through the cloud, for the benefit of current and future generations.

There are alternatives; the solution need not be left up to larger-than-life players. The California Digital Library (CDL), for instance, provides all campuses within the UC system with Merritt, which is a shared digital preservation system. Some academic libraries rely on professional, nonprofit, or commercial service providers such as DuraCloud, Portico, Chronopolis, and APTrust. Others are using digital preservation tools such as Preservica, MetaArchive, or Ex Libris Rosetta to engage in their own in-house digital preservation projects. What these approaches have in common, and get right, is undertaking the work now, at this critical moment in time.

This leads to the issue of trusted digital repositories. A popular idea making the rounds now is to create universal standards and certification for what are called trusted digital repositories — institutions that meet certification standards such as OAIS (Open Archival Information System) or TRAC (Trustworthy Repositories Audit and Certification) in the U.S., or the European Framework for Audit and Certification of Digital Repositories in the EU.

Establishing national and international certification standards is a good approach, but questions, such as the following, remain:

- Who will do the certifying — government agencies, institutions, consortia, private businesses, or all of the above?

- How is a unified standard to be applied when multiple copies are stored by different organizations in different parts of the globe?
- How can discovery and access be ensured across multiple trusted digital repositories?
- Will all trusted digital repositories follow the same standards and practices, or will national or regional standards vary?

Finally, there are the issues of security, content reliability, and authenticity. The digital assets of the Dunhuang Research Academy, for instance, are national treasures; the academy cannot afford to simply turn them over to any outside organization or repository for storage and preservation. It needs to know that the digital assets will be secure and stable through the entire term of deposit.

We must also consider the following threshold questions regarding digital preservation and the sustainability of digital cultural heritage and digital scholarship:

- Can we succeed to the same extent as our forebears did in preserving knowledge and published content?
- If we fail to keep in perpetuity content which is created today, how will future generations come to know us?
- What tools will ensure the preservation of massive amounts of digital content, and will they allow content created today to be accessible a century from now?
- How much do we want to preserve?

Since the objective of digital preservation is to ensure the existence of digital content for perpetuity or posterity, the digital lifecycle must be able to guarantee that knowledge and scholarship created in the digital age will have the same chances for survival that print-and-paper scholarship has had for centuries. In the context of global cultural heritage, digital preservation becomes a matter, not simply of national, but of international importance, whose significance extends beyond the boundaries of the academic community.

Acknowledgements: I would like to thank the following colleagues from the Dunhuang Research Academy: Ms. Fan Jinshi, the visionary who founded the Digital Dunhuang Project and spearheaded the academy's work in digital preservation for decades; Mr. Wang Xudong, the Academy's former director who was a driving force in the project with a comprehensive understanding of the opportunities and challenges of digital preservation; Mr. Wu Jian, Mr. Xia Shengping, and Mr. Zhang Yuanlin, who

led the digitization teams photographing the Dunhuang caves and organizing the content created with dedication. The success of Digital Dunhuang is due to their efforts and diligence.

I would also like to acknowledge Ms. Linda Tadic, recognized expert on digital preservation. Between 2011 and 2012, Linda and I studied the Digital Dunhuang project, co-authoring a report on the functional requirements of the Digital Dunhuang project. Both the study and report were commissioned by the Dunhuang Research Academy and the Mellon Foundation. Linda contributed enormously to the design and functional requirements of the Digital Dunhuang project, from data management to digital preservation. I am much in her debt. I have further benefited from discussions on digital preservation with some of the leading thinkers in the field, including Brian Schottlaender, Clifford Lynch, Larry P. Alford, and Luisa Mengoni. My thanks to all for sharing their wisdom and expertise with me.

Any errors are mine.

References

- Association for Library Collections and Technical Services (2007). Definitions of digital preservation. Prepared by the ALCTS Preservation and Reformatting Section, Working Group on Defining Digital Preservation. Washington, D.C.: American Library Association Annual Conference. Retrieved from <http://www.ala.org/alcts/resources/preserv/defdigpres0408>
- Gies, F., & Gies, J. (1994). *Cathedral, forge, and waterwheel: Technology and invention in the Middle Ages*. New York: HarperCollins.
- Lavoie, B. (2000). Meeting the challenges of digital preservation: The OAIS reference model. *OCLC Newsletter*, 243, 26–30. Retrieved from <https://www.oclc.org/research/publications/2000/lavoie-oais.html>
- Needham, J. (1994). *The shorter science and civilization in China* (Vol. 4). Cambridge and New York: Cambridge University Press.
- Tadic, L., & Zhou, P. (2012). *Digital Dunhuang: A repository system of Dunhuang murals and digital resources: Functional requirements*. Dunhuang: The Dunhuang Research Academy.
- Zhou, P. (2011). The development of a repository system of Dunhuang Murals and digital resources for Dunhuang studies. *Proceedings of the International Conference on Cultural Heritage and Digitization*, 317–323.