Facilitating preservation through collaboration: Harmonising practices across university departments

Received (in revised form): 24th April, 2017

Heidi Kelly

is the former Digital Preservation Librarian at Indiana University, where she focused on developing the Libraries' infrastructure for long-term management of digital assets. She is currently Digital Asset Manager at Saatchi & Saatchi LA. She received her MLIS from Wayne State University.

Saatch & Saatchi LA, 3501 Sepulveda Blvd., Torrance, CA 90505, USA
E-mail: heidi.kelly@saatchi.com

Nicholas Homenda

is the Digital Initiatives Librarian at Indiana University Bloomington Libraries, where he manages digital projects, services and initiatives in the Digital Collections Services Department. Nick has a master of science in information studies from the University of Texas at Austin and previously worked as a music librarian and an orchestral clarinetist.

Indiana University Libraries, 1320 E 10th St, Bloomington, IN 47401, USA
Tel: +1 812 855 2449; e-mail: nhomenda@indiana.edu

Abstract

In 2015, the Indiana University Communications Department partnered with IU Bloomington Libraries to collect and preserve born-digital institutional heritage. While this department utilises Widen as its enterprise digital asset management system, storage limitations necessitated the partnership in order for the objects to persist. Starting with 5TB of content, IU librarians collaborated with the University Archives in order to analyse the digital objects' metadata and determine a workflow for ingestion and long-term preservation of digital objects. This partnership expanded to include the Indiana University Kelley School of Business as well, leading to new challenges in terms of aligning practices across various units. The partnership between IU Communications, Kelley School of Business and the IU Libraries has posed many challenges to the existing infrastructure. It has required both social and technical solutions in order to establish a collaborative workflow that will in turn provide the basis for similar future projects with other schools and departments.
KEYWORDS: digital asset management, digital preservation, universities, metadata, collaboration, workflows

Introduction

Indiana University's Communications Department is responsible for collecting and managing a wide array of multimedia materials that document daily life at the university. These materials primarily consist of images and videos captured by Indiana University (IU) photographers, videographers and independent contractors. These materials are often used for marketing and promotional purposes. Beginning in the early 2000s, the Communications Department saw the shift of its multimedia materials from analogue to born-digital, which necessitated the use of a digital asset management (DAM) system. By 2015, the department had outgrown the functionality of its first DAM system, NetXposure (NetX — http://netx.net/), and was preparing to switch to a different DAM system, Widen (http://www.widen.com/).

The Communications Department approached the IU Libraries with a problem: how can terabytes worth of digital assets be preserved and made accessible outside of the DAM system after they have fulfilled the purposes for which they were created? Numerous campus departments have arrangements with the University Archives, a Libraries department, to deposit their records after a period of originally intended use; however, born-digital materials are more awkward to deposit in this manner. Throughout the 2000s, departments and individuals burned files onto optical media,

first CD-Rs, then DVD-Rs once the 650MB storage capacity of CD-Rs made them too small a vehicle for increasingly larger born-digital files. Once campus photography became almost exclusively born-digital and digital video became a common asset to manage, even DVD-Rs became too tedious a medium for transferring these materials to the Archives. By 2015, digital assets no longer immediately useful to IU Communications had grown to several terabytes' worth, with more and more added each year.

Concurrent with these discussions, representatives from another campus presence, the Kelley School of Business, approached the Libraries with essentially the same question. The selection archivist at the Kelley School was preparing for a similar DAM system migration from NetX to Widen and was wondering how the Libraries could provide assistance with configuring the new DAM system with sensible metadata fields from an appropriate schema to facilitate future ease of access and preservation. This case study describes the asset content analysis and discussions that followed these initial meetings, shedding light on the workflow, metadata and platform challenges present when approaching heterogeneous born-digital multimedia asset preservation and access at a large university.

Literature review

Although the business of managing information resources has long been the purview of libraries and archives, the concepts of digital asset management, descriptive digital image metadata, digital asset preservation and digital curation are relatively new, with focused work occurring over the last 15 years to establish policies and technologies to support procedures, workflows and best practices associated with these materials. The corporate world began understanding the necessity and utility of DAM systems for materials produced for marketing and promotional purposes, for example, although early solutions to this problem required extensive research and development compared with the numerous turnkey and customisable DAM systems available today.[1] At the University of Michigan, the Living Lab developed a DAM system to manage materials from numerous academic departments and support units but noted the lack of implementation information available despite the high level of interest in these sorts of projects.[2] There is also interest in leveraging libraries' expertise in managing materials housed in institutional repositories to expand their services. In 2009, Nicholas Joint from the University of Strathclyde outlined his vision for universities to implement a DAM system to house all university electronic materials in 'a single coherent framework', but noted that platforms that could handle the complex needs of this endeavour did not yet exist.[3] More recently, studies have investigated DAM tools as well as DAM system themselves. At the Los Alamos National Laboratory Research Library, librarians noted the shift from print to DAM due to the exponential growth of digital material, surveying data curation and digital repository assessment toolkits and describing using the Assessing Institutional Digital Assets (AIDA) toolkit at their institution.[4] Non-academic institutions are also concerned about the effective management of digital assets. In 2014, for example, Neumeier noted that a lack of personnel, institutional support and financial resources were common barriers to deploying a DAM system.[5] Libraries that want to implement a DAM system or migrate to a new one must approach the decision carefully, examining platforms according to specific criteria including metadata capabilities, administrative functions, content discoverability and usability, as chronicled by Wu *et al.*[6]

In addition to managing digital assets, research reflects libraries' concern with preserving them. Calculations in 2000 indicated that 93 per cent of the world's yearly intellectual output was in digital form.[7] Nearly two decades later, this percentage is undoubtedly higher. In the years following, strategic investigation into digital asset preservation was reported by the projects funded by the Joint Information Systems Committee, UK (JISC) Digital Preservation and Asset Management (DPAM) programme. These projects included preservation strategies for corporate and institutional

digital assets and produced surveys, tools and training programmes towards this initiative.[8] The construction of the Netherlands' Koninklijke Bibliotheek e-Depot initially aimed to preserve national library digital assets received in the 1990s but grew to become the trusted repository for electronic journal publishers including Elsevier and BioMed. Van Wijngaarden notes the need to be proactive: 'An archive cannot just wait for government organisations to turn over their records 20 years after their creation, because, by then, they may have become inaccessible already'.[9] In the USA, efforts to preserve digital assets include electronic government documents as well.[10] Moreover, digital asset preservation requires a robust technological infrastructure at the enterprise level, which is more easily attainable for some institutions than others. Dougherty advocates for a consortial approach enabled by the MetaArchive Cooperative Private LOCKSS (Lots of Copies Keep Stuff Safe) Network (PLN).[11]

Concerns about digital asset management often extend more broadly beyond DAM systems and digital preservation to related fields such as data management, data sharing and reuse, and researcher behaviours, and all of these can be collectively encapsulated within the discipline known as digital curation. The concept of digital curation has been discussed since 2001 as the umbrella over actions such as digital preservation and digital asset management, and in Dallas' examination of the literature, the 'wild frontier' myth of varied practices and approaches is questioned due to practical needs in today's environment.[12] Poole expands upon the conceptual landscape of digital curation through recent literature offering definitions of key ingredients such as curators, data, researchers, scholarly communication and infrastructure.[13]

Finally, relevant to this study is the issue of born-digital image assets and their metadata. The Smithsonian Institution's system-wide DAM system contains digital image assets with embedded descriptive metadata with fields defined by the International Press Telecommunications Council (IPTC) and using the Extensible Metadata Platform (XMP) format. Their Embedded Metadata Working Group created required and suggested fields for embedding in images, and the Still Image Working Group, as part of the Federal Agencies Digital Guidelines Initiative, endorsed these suggestions.[14] Reser examines these suggested fields, entertaining the possibility that certain embedded metadata might be apt to change, but affirming the usefulness of having any embedded descriptive metadata.[15]

Initial analysis of born-digital images

As the first step of the project, the Selection Archivist from Kelley School of Business provided Libraries staff with a 5TB external hard drive of IU Communications and Kelley School of Business content pulled from the old DAM system, NetX. Analysis of the content involved various steps in order to gain a comprehensive understanding of the assets and their unique needs for long-term curation.

The first step of analysis was simply to gain an understanding of the assets and their organisation. Using a dedicated digital archiving machine with BitCurator software (http://www.bitcurator.net/), the external drive was mounted with both hardware and software write-blockers to avoid unintentional manipulation of the metadata. The initial survey of the drive made it clear that there were a variety of file formats, unintuitive structuring of assets within folders and varying quality of metadata associated with each item. To develop a deeper understanding of the NetX assets, staff determined that the next step would be undertaking a full inventory of the drive's contents in order to capture essential information. This was done using a bash script (https://github.com/IUBDPresLab) originating from the Museum of Modern Art that was then modified to fit the IU Libraries' needs. This script captures six main points of metadata about each object: original filepath, filename, MIME type, MD5 checksum, checksum date and date last modified. Capturing this information about each asset at the earliest stage possible in the migration process was vital to

ensure its integrity, as system migrations and other events can often alter certain metadata. The second step of in-depth analysis involved running ExifTool (http://www.sno.phy.queensu.ca/~phil/exiftool/), a command line program that generates spreadsheets of information based on embedded metadata associated with image files, on the entire drive. While the results of ExifTool can be extensive depending on the options specified, this step was crucial in identifying trends such as lack of identification of subjects or locations and the varying level of technical metadata. It also pinpointed a key challenge in working specifically with NetX content, which was the issue of several large system glitches, which at various points overwrote the date of creation for all assets in the system at the time. Finally, the third step of analysis aimed to collect as much available descriptive metadata as possible to gain a sense of all of the metadata available for these assets. Metadata dumps from NetX and Widen were compared with and combined with the aforementioned bash script and ExifTool outputs.

Figure 1: NetX assets from Kelley School of Business by MIME type

Once initial analysis was complete, it was determined that bringing all stakeholders together would move the project forward and help in defining high-level priorities. Staff from the Indiana University Libraries, University Archives, Kelley School of Business and IU Communications were present at the meeting in February 2016, where next steps were established. Three major points were identified: holding a metadata summit to align practices in description; developing a specific repository solution in the Libraries based on the identified needs of the project content; and then establishing a workflow for moving content from Kelley and IU Communications into the Libraries' repository environment.

Aligning metadata practices

Indiana University's Metadata Analyst, Juliet L. Hardesty, developed and led a metadata summit for key staff at IU Communications and the Kelley School of Business. The aim of this was to ensure that these units were capturing necessary metadata as early as possible, so that long-term sustainability and discoverability of assets could be assured. Hardesty defined the minimum required descriptive metadata for discoverability as subject, location, date and key subject tags according to internal standards developed for digital collections materials and services developed and managed by the IU Libraries. The sheer volume of assets held by the two units necessitate this focus on targeted tagging of overarching subjects, as discoverability of these assets will likely differ from those of digitised image collections. Users for this type of content are often looking for marketing images, rather than searching for images related to a specific individual or event, so using tag ontologies that serve this purpose enables better discoverability. The summit also extolled the need to capture technical metadata from photographers' cameras, and to ensure the ongoing authenticity of such metadata as assets are migrated to different systems. As the date of creation was altered in NetX, there is a clear need to capture and maintain the metadata as a separate file, rather than as embedded metadata within the asset, as well as a need to ensure the integrity of assets and representation information through ongoing fixity checks.

Figure 2: Complete and incomplete metadata fields for all NetX assets

Born-digital marketing content curation needs

Indiana University Libraries has a long history of contributing to and leading open source development projects, especially in terms of repository solutions. The Libraries has a robust technical environment, with repositories developed to support the needs of digitised images (www.dlib.indiana.edu/collections/images/), born-digital and digitised audiovisual (https://media.dlib.indiana.edu/), and born-digital and digitised paged media content

(https://pages.dlib.indiana.edu/). The marketing image collections from Kelley School of Business and IU Communications, however, presented challenges to this existing infrastructure. From initial analysis of the NetX content, it was clear that the needs of born-digital marketing content are quite divergent from those of a majority of the content currently managed by the IU Libraries. Kelley School of Business and IU Communications were much more likely to rely on JPEG for their main file format, where the Libraries' standard, largely because of a focus on digitised assets in the recent past, has been TIFF. For marketing images with storage ceilings on their DAM system, TIFF is not a viable format due to its sheer size. However, preservation standards generally rely on the format for long-term archival purposes because of its losslessness and high quality. The differences in metadata related to each asset also have broad implications on discoverability within a repository solution, as the sheer volume of assets with little to no descriptive metadata would overwhelm any current repository and render all of the images more or less undiscoverable.

As determined from the initial research in this project, the proliferation of different formats and qualities, along with the varying quality of descriptive and technical metadata, require a different management system. The system will have to support more file formats and varying levels of quality, and would ideally have the ability to normalise content from proprietary or non-preservation formats into preferred formats. It would primarily rely on high-level search terms and tagging of images to render them discoverable. While these needs have been clearly identified, the next step of developing a repository for this content has yet to begin. This need will continue to grow, however, as more and more units within Indiana University develop marketing content that is necessary to preserve long-term. It is clearly a problem that exists beyond just Kelley School of Business and IU Communications, so this project aims to build out a solution that can be implemented widely. For this reason, the development of a migration workflow has been identified as the third key piece of the project.

Developing migration workflows

While Indiana University does not have a campus-wide licence for a DAM system, it is clear that departments will be increasingly utilising such systems to organise and maintain their various digital assets. For this reason, the IU Libraries needs to build out workflows to retrieve assets from a proprietary DAM system like Widen in order to move them into long-term repository solutions for preservation. The most important challenges will be exporting items and their corresponding metadata in a way that can then be easily ingested into a new repository. Initial tests with Widen materials suggest that this should be relatively easy; however, the scalability of export and reimport is unclear. This workflow development piece will be an area of continuous development as DAM becomes more widely addressed across Indiana University.

Conclusion

Several key findings have helped to guide Indiana University Libraries' development in the area of digital asset management. The first is that early education of other departments and colleges within the university is key to successful long-term preservation. This especially has an impact on the quality and amount of descriptive and technical metadata being captured and maintained along with the assets. This iterative approach is also especially important as the Libraries builds toward developing a repository solution for the sorts of digital assets that this project has focused on, as staff in other areas of the university can define their needs as both content creators and users. Actively working with these units will have a positive impact on the development of final repository solutions as well as the migration workflow from a DAM system to a future Libraries' system.

The second finding is that there is a great need beyond the Kelley School of Business and IU Communications for DAM solutions on campus. Developing an institutional licence for a system and providing training and best practices to various units will help in maintaining a comprehensive institutional memory in the digital age. As part of this project, the Libraries has started to connect various units to each other as a first step, as information sharing across the campus will assist in better management practices overall.

The final point is that the Libraries' must act as a leader for the whole institution in terms of digital asset management. By providing information about metadata best practices and connecting various units to each other, the Libraries empowers staff across the institution to manage their digital assets in better ways. Many units so far have had the appropriate staff and resources in place, but are lacking the knowledge of best practices and the ability to trust their own judgment.

On the part of the IU Libraries, this project has been both a great challenge and a great opportunity to take on a new role within the larger institution. It is clear from this project that practices in the long-term preservation of digital assets must evolve to meet the needs of various content types and creators. Memory institutions like libraries and archives should be thinking about building out pathways and connections instead of relying on boxy best practices and onerous requirements. The ability to cope with constant new challenges in the digital realm and work iteratively will situate these sorts of institutions to better preserve items for future generations.

References

1 De Lancie, P. (2002) 'Refreshing media management: Coca-Cola turns archives into assets', *E Content Mag*, May, pp. 18–22.

2 Kim, Y., Ahronheim, J., Suzuka, K., King, L. E., Bruell, D., Miller, R. and Johnson, L. (2007) 'Enterprise digital asset management system pilot: lessons learned', *Information Technology and Libraries*, Vol. 26, No. 4, pp. 4–16.

3 Joint, N. (2009) 'Practical digital asset management and the university library', *Library Review*, Vol. 58, No. 2, pp. 89–96.

4 Miller, L., Blake, M. and Sorsby, M. (2012) 'Evaluation of the Assessing Institutional Digital Assets (AIDA) Toolkit', *Science & Technology Libraries*, Vol. 31, No. 1, pp. 92–99.

5 Neumeier, L. K. (2014) '"The mess is there": digital asset management system use in non-academic archival repositories', unpublished master's paper.

6 Wu, A., Thompson, S., Vacek, R., Watkins, R. and Weidner, A. (2016) 'Hitting the road towards a greater digital destination: evaluating and testing DAM system at University of Houston Libraries', *Information Technology and Libraries*, Vol. 35, No. 2, pp. 5–18.

7 Kenney, A. R. and McGovern, N. Y. (2003) 'The five organization stages of digital preservation' in Hodges, P., Bonn, M., Sandler, M. and Wilkin, J.P. (eds) 'Digital Libraries: A Vision for the 21st Century: A Festschrift in Honor of Wendy Lougee on the Occasion of her Departure from the University of Michigan', Michigan Publishing, University of Michigan Library, Ann Arbor, MI.

8 Pennock, M. (2007) 'Supporting institutional digital preservation & asset management: a summary of the JISC DPAM programme synthesis', *New Review of Information Networking*, Vol. 13, No. 2, pp. 119–132.

9 van Wijngaarden, H. (2007) 'Long-term preservation and permanent access: how to ensure the long-term reuse value of your digital assets' *Journal of Digital Asset Management*, Vol. 3, No. 2, pp. 102–109.

10 Best, R. (2009) 'Preserving and providing access to digital state publications in Alabama: a case of cooperation', *The Southeastern Librarian*, Vol. 57, No. 3, pp. 3–15.

11 Dougherty, W. C. (2009) 'Preservation of digital assets: one approach', *Journal of Academic Librarianship*, Vol. 35, No. 6, pp. 599–602.

12 Dallas, C. (2016) 'Digital curation beyond the "wild frontier": a pragmatic approach', *Archival Science*, Vol. 16, No. 4, pp. 421–457.

13 Poole, A. (2016) 'The conceptual landscape of digital curation', *Journal of Documentation*, Vol. 72. No. 5, pp. 961–986.

14 Embedded Metadata Working Group — Smithsonian Institution (2010) 'Basic Guidelines for Minimal Descriptive Embedded Metadata in Digital Images', available at: http://www.digitizationguidelines.gov/guidelines/GuidelinesEmbeddedMetadata.pdf.

15 Reser, G. (2012) 'What not to embed: is it better not to embed certain cultural heritage metadata in images?', *VRA Bulletin*, Vol. 39, No. 1, Article 5.