Securing Digital Content

by Sally Hubbard

Digital Images in Museums

The Getty Research Institute is an operating program of the J. Paul Getty Trust in Los Angeles, which also includes the J. Paul Getty Museum, the Getty Conservation Institute, the Getty Leadership Institute and the Getty Foundation. This paper will briefly examine the steps taken at the Research Institute and at the Getty as a whole to ensure that its digital content is managed responsibly. This effort is very much a work in progress, as one of the complications of developing strategies for digital content is that the strategies may require innovative approaches that challenge existing practices, as well as organizational structures and hierarchies that require an openness from all parties and an acceptance that change is likely to be gradual.

Digital assets have been making inroads into the collections and production of the Getty for several years, as elsewhere. At the Research Institute, again as at many institutions, this new direction at first manifested itself in a series of projects aimed primarily at making surrogates of selections from analogue collections available over the web, but otherwise dissimilar. These early forays had varying concepts of digital stewardship where they had any at all. The master images files that these projects produced, in some cases almost as a side effect, were stored in the first instance under a hapless staff member's desk, held in several boxes of CDs containing one copy of each TIFF formatted image. In more recent years an awareness of the inherent vulnerability of digital assets and records has broadened and with it the realization that some sort of curation for such content was required.

Curation is understood here in the broadest terms. Without entering into any discussion about life-cycle versus continuum approaches, the term is used to indicate a recognition that there can be no benign neglect of digital content, and that it requires continual care, ideally from before the moment of creation (in terms of being made in accord with certain standards and policies). At first this awareness was rather unsupported, as few tools were available to aid in the nurture of digital collections. The Research Institute's first attempts with those boxes of CDs involved five simple, if laborious, steps:

1. Make duplicate CDs;
2. Separate the two sets geographically;
3. Store them at appropriate temperature and humidity levels;
4. Transfer the images to a server with tape backup;
5. Capture at least minimal technical metadata based on the now ancient Research Library Group (RLG) preservation metadata elements (description metadata for the items represented in those images were held in separate catalog systems).

This process hit several snags, including a lack of server storage space and the bandwidth limitations of the Getty campus. These problems have continued to dog the digital asset management process as it has become more sophisticated, since demand has tended to outpace improvements in infrastructure – though we are now at a stage where this imbalance is less of an issue. At the same time the Research Institute moved to implement minimum imaging guidelines and a file naming protocol, both based upon the work of the California Digital Library, with the aim of standardizing current and future production. These specifications, at least theoretically, were adopted as cross-programmatic standards.

As it became obvious that the sort of largely manual processes described above were not scalable and that the situation was likely to become more

Sally Hubbard is digital projects manager at the Getty Research Institute. She can be reached at shubbard<at>getty.edu.

serious as "born digital" material and media other than still images entered collections, the Getty undertook a lengthy cross-program search for a digital asset management system (DAMS) beginning in 2002. While uniting the programs made sense in many ways, both economically and philosophically, it also complicated both the search for and the still unfolding implementation of the chosen DAMS [1]. Each program inhabits a particular information universe into which the DAMS has to be inserted, with different applications, different metadata traditions and different professional mores, procedures and needs. For instance, the initial image metadata model of the DAMS had to be shared across programs and, therefore, had to accommodate many different needs. Perhaps inevitably it has a certain jack-of-all-trades quality. The information architecture of the Research Institute was perhaps the most complicated with the core library management system surrounded by various additional applications at assorted life stages (some being phased in and some lingering on only because they held essential data). These major systems were in addition to a plethora of desktop and paper-based systems, holding data conforming to various shared and local standards. The museum, by contrast, essentially had a single application to deal with – their collection management system (CMS), which held impressively unified data. The complexity of the situation at the Research Institute is to some extent reflective of its very much larger – in terms of the number of individual items – and more diverse collections, many of which are only cataloged at the collection level, meaning that item-level documentation is quite likely to happen only after, and to be prompted by, digitization.

A DAMS was intended not only to provide digital curation functions, but also to assist in various day-to-day business tasks, such as order fulfillment and sharing assets across programs and departments. The first department to implement the system was Trust Communications, whose requirements included providing electronic press kits for exhibitions and other events, photographs of the Getty site and staff for various purposes, etc. The DAMS was customized to allow "packaging" of images and caption metadata and invoicing for those packages, if required. Though this model was useful for the other programs it could not be adopted wholesale. The packaging specifications, for instance, for the Research Institute are slightly different, and invoicing occurs externally. The initial museum implementation involved importing existing images from local storage and metadata from the CMS and linking them within the DAMS, a process that in itself required a good deal of time and analysis. The Research Institute, with fewer existing assets and less unified existing metadata, moved more quickly to incorporate the DAMS into its production process, that is, images from the Institute's photography studio were uploaded directly into the DAMS, and item-level cataloging occurred within the system itself. This latter procedure was not ideal, as the DAMS was not designed as a cataloging utility, but in practical terms it was the best solution available. The DAMS is also the first place where item-level rights metadata is captured.

Assets ingested into the DAMS are all made to certain standards. They are tagged to a greater or lesser extent with descriptive and administrative metadata and therefore made locatable both in the sense that they can be retrieved and that they can be associated with a given collection or parent object. They can also be transformed on export as required and distributed easily via FTP and email, and they are subject to good business backup procedures. They are therefore significantly better managed and more useful than they were previously. However, no currently available DAMS provides a complete digital preservation solution, though some provide more pieces than others. The DAMS implementation has perhaps paradoxically involved a greater understanding – and a greater effort to communicate to the wider Getty community – that a DAMS should be regarded as one part of digital curation, and that key functions – preservation and others – will occur elsewhere. For instance, our current DAMS is a staff-only utility that does not provide any public access to our collections. Also, it is not "metadata agnostic" and so cannot ingest any object without metadata loss, and it does not provide persistent identification or fixity checking, both core preservation functions. Moreover, even the best backup procedures are intended to bring systems back tomorrow in a state analogous to the one they were in today, not to be able to preserve assets and records over years or decades and re-manifest them in what may be entirely different technological and indeed historical contexts, obviously a much more difficult and complex problem.

During the period of the Getty's DAMS implementation several

significant developments have occurred, both internally and externally. The year 2002 saw the release of *Trusted Digital Repositories: Attributes and Responsibilities* [2], which laid out the characteristics of any trustworthy digital depository. The Reference Model for an Open Archival Information System (OAIS) [3], developed to "create a consensus on what is required for an archive to provide permanent or indefinite long-term, preservation of digital information" (OAIS does not deal with digital stewardship issues prior to ingest into an archive), was approved as an ISO standard in 2003. The 2005 *Audit Checklist for the Certification of Trusted Digital Repositories: Draft for Public Comment* [4] was updated to the *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist* in 2007 [5] and provided a tool for the self- or external audit of repositories. Version 1.0 of PREMIS (Preservation Metadata: Implementation Strategies) data dictionary [6] was released in 2005, with the next version due for release in 2008. The InterPARES 2 project [7], which examined issues of reliability and accuracy during the entire lifecycle of records, concluded in 2006. These and other examples indicate that the level of understanding of digital curation and preservation issues has grown by leaps and bounds in the last five or so years, and perhaps even more crucially tools and applications that begin to make preservation a practical proposition have started to appear.

Internally, the Getty's institutional capability for digital stewardship has grown, though likely not to the extent that will eventually be required. A Getty institutional archives has been established and has developed records management policies and procedures. The archives has collaborated with the Research Institute's departments of digital resource management (also relatively new) and information systems, the central information technology program and other key departments across the Getty, such as the photography studios, to analyze existing practices and priorities and propose future strategies and solutions. Various initiatives that fall under the broad rubric of digital curation have emerged from this analysis and are currently either under way or at least under consideration. Many of these initiatives are not primarily about technology. Tools and applications are of course vital when dealing with digital material, particularly in any kind of scalable way. However, other factors are perhaps even more crucial:

- Policy development
- The nitty-gritty of workflow and dataflow, especially when dealing with several applications across different programs and departments
- Dedicated staff and budgets
- A sustained institutional commitment.

The Research Institute is currently engaged in a functional analysis of its different applications with an eye to developing the most efficient and responsible workflow and allocation of responsibilities among systems. This ongoing process has been recently energized by the arrival of another potential, library-oriented, digital asset management application [8], one that could not replace all the functions offered by the established DAMS, but does provide additional utility. These enhancements include a public interface and ability to perform preservation tasks such as file format identification and validation, persistent identification functionality, fixity checking to help ensure the authenticity of objects, technical metadata extraction and PREMIS compliant "event" tracking of changes to files. Again, this software would not represent a complete preservation solution, but it would fill some existing gaps. The Research Institute and the institutional archives are also exploring the potential of the iRODs system [9] to round out the preservation functionality available and plan a pilot project utilizing the system. It should be noted that the expectation is that any final preservation strategy will be Getty-wide, not least because it is not economically feasible to conduct effective preservation on a program-by-program basis, but exploration is occurring on a smaller scale.

Other Initiatives and Proposals

Several other proposals are being considered:

- The reconciliation of all policies and procedures relevant to digital content (and it may not always be obvious which those are), together with the development of comprehensive policies and procedures for the creation, ingest, maintenance, preservation and dissemination of digital content in accordance with current best practices (which will always be a moving target)

- The establishment of a technology archive that could provide in-house access to obsolete hardware and software. Such an archive could not hope to be comprehensive, but would be useful even if limited in scope
- The establishment or maintenance of allied analogue preservation strategies where practical, for instance, the generation of archival microfilm copies of digital content
- The testing of policies and procedures through a pilot project and eventually the recommendation of a particular trusted digital repository infrastructure and solution.

This last activity is obviously one of the most important and difficult tasks, with a host of variables to consider such as open source or proprietary, local or consortial. Perhaps fortunately this field is still maturing, and we have a great deal of work to engage us before any choice is required. In the meantime, we can hope for further resolution of what exactly is involved in archival storage and management or, to use OAIS terms, in creating an AIP (archival information package).

Among other significant aspects of a comprehensive digital strategy not examined in this paper are the following:

- Efforts at the Research Institute to adopt a more purposeful decision-making process as to what is digitized, hitherto largely dictated by ad hoc orders and upcoming exhibitions and similar events
- Improvements to the digital capture process
- The vitally important metadata strategy
- Provision of public access to content, though this is a core part of our mission.

This paper has focused on asset management, rather than record management, simply because this is the area the author is most familiar with. I will try to speak to all of these areas when I say that digital curation, broadly understood, is a collaborative and iterative process that will involve most staff in some way at some point in time and should be standards-driven, policy-based and strategic. ■

Software and Text Resources Mentioned in the Article

[1] The application then known as TEAMS was created by Artesia, now owned by OpenText. Author's note: This article should not be regarded as either an endorsement or a repudiation of any particular application.

[2] RLG (Research Libraries Group) & OCLC (The Online Computer Library Center). [2002, May]. *Trusted digital repositories: Attributes and responsibilities* (An RLG-OCLC report). Mountain View, CA: RLG. Retrieved February 19, 2008, from www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf. Author's note: The two organizations have since merged.

[3] Consultative Committee for Space Data Systems. [2002, January]. *The Open Archival Information Systems (OAIS) reference model* (CCSDS-B-1) (ISO 14721). Washington, DC: CCSDS Secretariat. Available February 19, 2008, from public.ccsds.org/publications/archive/650x0b.pdf. The OAIS is the primary international standard pertaining to the long-term preservation of information.

[4] RLG (Research Libraries Group & NARA –National Archives and Record Administration). [2005, August 31] *Audit checklist for the certification of trusted digital repositories: Draft for public comment.* Washington, DC: RLG.

[5] RLG-NARA Digital Repository Certification Task Force. [2007, February]. *Trustworthy repositories audit and certification (TRAC): Criteria and checklist* (Version 1.0). Chicago, IL: CRL, Center for Research Libraries; Dublin, OH: OCLC Online Computer. Retrieved February 19, 2008, from http://bibpurl.oclc.org/web/16712.

[6] PREMIS (Preservation Metadata Implementation Strategies). [2005]. *PREMIS data dictionary (Version 1.0).* Dublin, OH: OCLC.

[7] The InterPARES (International Research on Permanent Authentic Records in Electronic Systems) 2 Project ran from 2002-2006. For further information see www.interpares.org/

[8] DigiTool is created by Ex Libris. Author's note: This article should not be regarded as either an endorsement or a repudiation of any particular application.

[9] iRODS is a data grid software system being developed by the San Diego Supercomputer Center (SDSC) Storage Resource Broker (SRB) team and collaborators. The system (http://irods.sdsc.edu) allows the implementation of policy by translating it into rules and state information and providing a rule engine that dictates the system response to requests and conditions.