DigItalia

Numero 2 - **2007**

Rivista del digitale nei beni culturali

ICCU-ROMA

Getting to the Web

Sally Hubbard
Getty Research Institute of Los Angeles

This article discusses the long chain of operations involved behind the scenes before, and after, cultural heritage collections make an appearance on the World Wide Web, focusing particularly on Web access and digital preservation. Many institutions are in the process of making the transition from "project-based" to "program-based" digitization, and are attempting to knit together and implement a fully integrated and coherent digitization strategy. Because digital technology has an inherent tendency to break down time-honored barriers and niches, this transition can be difficult. A digitization program is likely to have an impact in many traditional arenas: acquisition; collections conservation and cataloguing; description and access; distribution and exhibition; and intellectual property or digital rights management, in addition to requiring attention to digital capture itself and the management and preservation of digital objects. Digitization programs may therefore require greater consensus and cooperation across an institution, or between institutions, to be successful than was true of limited digital projects undertaken by one department or another. Even after the long upstream journey to a live Web site has been made, the accessibility of collections is a complicated issue with no single solution. A combination of traditional cataloguing; new data standards and protocols; social tagging; full-text availability; thesauri and ontologies; and perhaps eventually some forms of automated visual and or aural indexing may all be required to navigate intelligently through an increasingly massive complex of heterogeneous material.

The Getty Research Institute, an operating program of the J. Paul Getty Trust in Los Angeles (which also encompasses the J. Paul Getty Museum, the Getty Conservation Institute, and the Getty Foundation) has been engaged in digitization for several years, but has only comparatively recently attempted to formulate and begin to implement a comprehensive digital strategy, one part of which is an explicit goal of providing surrogate access to our collections over the Web. One implication of this goal is that the Research Institute will be required to move from its current artisanal, hand-crafted approach to a more scalable or industrial model. Achieving this even in a limited way may require a significant cultural and methodological shift within our institution.

At the time of writing, past projects and current production have left the Getty Research Institute with over thirty thousand digital images to manage – a tiny proportion of the total number of original items held in our collections. Where the legacies of past projects – generally surrogate images of special collections items, the master files generally in TIFF format but otherwise created according to varying specifications – are available over the Web, they are distributed across access mechanisms: some are available on the Research Institute's publicly available Web pages as online exhibitions or searchable "digitized library collections"; others are attached to an EAD finding aid, others to one of our photo study collections; and others to entries in our OPAC. There is, at the moment, no one place for a member of the public or a library patron to find all digital assets through the Research Institute site.

Some material has been made available externally; both the J. Paul Getty Museum and the Research Institute have contributed to ARTstor, in a test case utilizing the OAI Protocol for Metadata Harvesting (OAI-PMH). More recently, the Research Institute has become a participant in the Open Content Alliance (OCA), contributing public-domain books and periodicals: under this arrangement, Research Institute general collection, non-fragile material is sent to the University of California, Los Angeles to be scanned, with the intention that a scanning station will be set up at the Research Institute itself to scan special collections or fragile items. As of this writing, no local copies of material submitted to the OCA are kept.

The Getty is in the process of introducing a Digital Asset Management (DAM) system, which has involved a phased, program-by-program, implementation. While a shared application brings many advantages, there are also difficulties, such as fitting one application into the various information architectures and metadata traditions of the different programs. For the Research Institute, with its extremely large and diverse collection, the DAM is often the place where collection materials first receive item-level descriptive and rights metadata – although all materials held by the Institute have at least a collection-level record in our OPAC. That data is repurposed within the DAM – which has required enlisting the expertise of cataloguing and registrarial staff. This is only one instance of the complicated, department-crossing workflow and dataflow involved in the asset management process.

Approximately one third of existing images of Research Institute collection materials have made their way into the DAM at the time of writing, with the others currently staged for ingest. The DAM is for staff use only and provides no public access, though the intent is that in time it will generate access images for public and scholarly consumption. On a day-to-day level, the DAM provides a destination for the production of the in-house photography studio, largely generated in response to *ad hoc* orders from researchers and scholars, but also fulfilling larger-scale projects, exhibitions, etc. The expectation is that digital capture will become more systematic and occur on a considerably larger scale in the future, and will include audiovisual assets. If this indeed occurs, it may necessitate a review and reformulation of current production procedures, which offer limited scalability.

In the meantime, the hybrid strategy of outsourcing large-scale digitization projects while keeping smaller ones in house has made it possible for the artisanal and industrial production models to co-exist, but it is unlikely that such a separation will be sustainable.

A digital asset management system was from the beginning viewed as only one part of responsible stewardship of our digital collections, and the Research Institute is also exploring the best path forward for a digital preservation program. Digital preservation is this context refers not only to the digital surrogates that have to date made up the bulk of our digital collection, but also digital originals, which are entering our collections in larger numbers.

Because the technological/hardware path to digital preservation repository is still unclear, and the central OAIS[1] standard is actually more helpful as a tool to guide policy than as a blueprint for technological development, we are now concentrating on reviewing current practices and developing policy while keeping abreast of standards and other developments in the field. The Research Institute is also exploring a small testbed project using the iRODs system[2].

The foregoing description is a very brief and necessarily partial view of the current situation at one institution. It is probably not untypical, at least in the sense that it indicates that the process of pulling together the legacies of past projects and future ambitions into one coherent whole may be complex. An additional consideration is the fast-changing nature of the technological landscape, and while we are pursuing the basic goal of creating surrogates of our collections, "Web 2.0" developments such as wikis, social tagging, and other user-generated and interactive forms of content are changing the parameters for providing access. (The Research Institute intends to experiment with the implementation of PennTags[3], and recently set up its first wiki for a scholarly project that involved the digitization of the nine-volume *Cérémonies et coutumes religieuses de tous les peuples du monde: représentées par des figures dessinées de la main de Bernard Picard* [Picart], 1723 -1737).

A metadata strategy is a key component of any digitization program, and one that will obviously be vital to discovery and access to collections (and to other aspects of digitization). Its importance is underscored by a recently offered caveat:

[1] The Open Archival Information Systems Reference Model (ISO 14721), the primary international standard pertaining to the long-term preservation of information.

[3] 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 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, http://irods.sdsc.edu.

[3] PennTags is a social bookmarking tool for locating, organizing, and sharing online resources developed within the University of Pennsylvania, http://tags.library.upenn.edu/help.

> «The supposed universal library [...] will be not a seamless mass of books, easily linked and studied together, but a patchwork of interfaces and databases, some open to anyone with a computer and WiFi, others closed to those without access or money. The real challenge now is how to chart the tectonic plates of information that are crashing into one another and then to learn to navigate the new landscapes they are creating»[4].

By implication, one challenge for metadata will be to help users find their way. Another challenge will be, again, to find a scalable way of applying metadata to all digital assets, which may involve negotiation on the minimum requirements per asset. In this context it is interesting to note the work of the DCMI Kernel Metadata Community, which proposes a minimal set of four elements: *who, what, where*, and *when*[5].

Dempsey et al. have suggested that the metadata landscape is now extremely intricate and diverse, and that institutions may experience a list of complications that existing approaches may be unable to accommodate. "Multiple metadata creation and repository environments", referring to the plethora of both local and shared applications and interfaces many institutions now employ, such as library management, content management, and digital asset management systems. "Multiple metadata formats", formal standards such as MARC, Dublin Core, EAD, etc., coexisting with informal or "vernacular" ones, all of which will require mapping to allow metasearching or harvesting but which may or may not have been created using consistent rules. Dempsey points out that interoperability will have to occur at three levels: encoding or format, data structure, and data content values. "Multiple controlled vocabularies" may be needed, with different materials imposing different demands. "Automatic categorization and metadata creation" may cease to be merely interesting and become necessary "as the volume and variety of digital resources increases, and the economies of shared cataloging are not available in the same way as many of these resources are unique. This becomes increasingly the case as other kinds of metadata are also required: structural or technical for example". Finally, metadata will be needed to manage complex objects[6].

[4] Anthony Grafton, *Future Reading: digitization and its discontents*, «The New Yorker», 5 November 2007.

[5] The DCMI Kernel Metadata Community is a forum for individuals and organizations interested in lightweight representations of Dublin Core metadata aimed at maximizing utility and minimizing cost of creation, maintenance, and exchange of Dublin Core and other metadata standards that interoperate with it, http://dublincore.org/groups/kernel.

[6] Lorcan Dempsey – Eric Childress – Carol Jean Godby – Thomas B. Hickey – Andrew Houghton – Diane Vizine-Goetz – Jeff Young, *Metadata switch: thinking about some metadata management and knowledge organization issues in the changing research and learning landscape*, in: *LITA guide to e-scholarship*, Chicago: Debra Shapiro, c. 2005, http://www.oclc.org/research/publications/archive/2004/dempsey-mslitaguide.pdf.

Another established practice already being questioned is the breakdown of metadata standard adoption by institutional type within the cultural heritage sector, rather than by the type of material described. A more sustainable and practical model would seem to be the one suggested by Elings and Waibel and re-presented below[7]. However, it should be borne in mind that these standards are themselves likely to evolve: Witness the work of the Library of Congress Working Group on the Future of Bibliographic Control[8].

Material Type	Material Culture/ Visual Resources	Bibliographic	Archival
Data Structure	CDWA/VRA Core	MARC	EAD
Data Content	CCO	AACR2 (RDA)	DACS
Data Format	XML	XML/ISO2709	XML
Data Exchange	OAI	OAI, Z39.50, SRU/SRW	OAI

Rights metadata can be a particularly thorny issue for archival and special collections, where often the only intellectual property information in place is at a collection level and may or may not pertain at the item level. This may require that rights research be part of the digitization process, which can be an arduous task given the size and complexity that archival collections can reach, and one that can be difficult to reconcile with an attempt to create a "bare-bones" industrial approach to digital production.

Although certain limits and exceptions to copyright restrictions may apply – such as "fair use" in the United States and "fair dealing" in other jurisdictions – these still provide at best a contested legal space in which to conduct digitization for many materials. The safest route is often to choose to make available only material that is clearly in the public domain, or else clearly owned by the hosting institution or an agreeable third party. Another option is to restrict who can see what – for instance allowing only on-site viewing of certain material. However, such options might mean that access to the most interesting and unique materials is restricted; this particularly applies to so-called orphan works for which no owner can be identified or located. In the end, there is often a certain level of risk involved in posting collections to the Web, particularly when one considers international variations in law. For this reason, part of the policy and strategy regarding digitization should be determining the institutional level of risk aversion.

[7] Mary W. Elings – Günter Waibel, *Metadata for All: Descriptive Standards and Metadata Sharing across Libraries, Archives and Museums*, «First Monday», vol. 12, n. 3, 5 March 2007.

[8] The charge of the Working Group on the Future of Bibliographic control is to «present findings on how bibliographic control and other descriptive practices can effectively support management of and access to library materials in the evolving information and technology environment», http://www.loc.gov/bibliographic-future/.

Any discussion of access should also touch on the allied imperative of preservation, as preservation without access is pointless, and access without preservation will be brief. There is no real equivalent in the United States to European initiatives such as CASPAR (Cultural, Artistic and Scientific Knowledge for Preservation, Access and Retrieval)[9], DPE (Digital Preservation Europe)[10], or PLANETS (Permanent Long-term Access through Networked Services)[11], but there has been a good deal of activity and focus on digital preservation in recent years, particularly on digital repositories and the preservation of electronic records. The most notable of these programs are the NDIIP (National Digital Information Infrastructure and Preservation Program,)[12], based at the Library of Congress, which provides a collaborative research framework which crosses sectors and states; and the National Archives' ERA (Electronic Records Archive)[13], program, part of the Federal Records Management Project, which promises to create technology with the potential to be useful to institutions and business in many different sectors. Other important developments are the release of the PREMIS *Data Dictionary for Preservation Metadata* in 2005, and the revision and expansion of the *Audit Checklist for the Certification of Trusted Digital Repositories*[14] into TRAC, the *Trustworthy Repositories Audit & Certification: Criteria and Checklist* [15] in early 2007.

Digital preservation is something of a fuzzy term. One advantage of the Getty Research Institute's recent forays into establishing a digital preservation program has been the ability to more precisely delineate the difference between asset management, preservation, and business backup procedures for the wider Getty community. Not all applications are the same, and certain tasks may be allocated to different applications in various technical configurations, but it is important to be clear about both what functions are required for preservation, and which can and cannot be taken over by any application. An asset management system may simply and straightforwardly allow the storage, tagging, and transformation (on export) of objects. This functionality may be extended by customization; the Getty, for instance, has built some customizations for its DAM system that additionally facilitate order fulfillment, such as the generation of reports showing thumbnails and captions as well as cover and permission letters. Everything stored in a DAM system should be subject to good business backup procedures, largely

[9] http://www.casparpreserves.eu.

[10] http://www.digitalpreservationeurope.eu.

[11] http://www.planets-project.eu.

[12] http://www.digitalpreservation.gov.

[13] http://www.archives.gov/era.

[14] Originally developed by the Research Libraries Group and National Archives and Record Administration (RLG-NARA) Digital Repository Certification Task Force.

[15] Created by the Center for Research Libraries (CRL), the National Archives and Record Administration, and the Online Computer Library Center (OCLC).

aimed at disaster recovery. That is to say, backup procedures are intended to bring the system back up tomorrow in a state analogous to the one it was in yesterday or today, not to be able to preserve an asset over years or decades. A DAM system may be regarded as one component in a continuum of preservation, but one that assumes that key preservation functionality will occur externally.

Digital preservation functions, more likely to be found in or referenced by an expressly built preservation repository than in an asset management system, would include (but are not limited to): persistent identification, identification that can be used and maintained over the life of an digital object; digital object validation and identification, the ability to determine the file format and validity against standard format specifications; metadata support and registry, ideally "metadata-agnostic" in order to allow the ingest of any required object without metadata loss; file format support and registry – again, ideally the repository would be format-agnostic and able to accept any file submitted; fixity checking, as an essential component of ensuring authenticity; and migration/normalization capability, as it may be assumed that migration or transformation will be at least one of the preservation strategies employed. This would be in addition to or as part of the key functions identified by the OAIS reference model: ingest; archival storage; data management; administration; preservation planning; access.

Although the research of the last several years has made such a listing of preservation functionality possible, there is still some debate about precisely what is involved in archival storage and management, or in creating an AIP (Archival Information Package) in OAIS terms. It should be remembered that all digital preservation strategies and programs are speculative, in the sense that no one yet knows from experience how to preserve digital objects for decades or longer. The PREMIS metadata model has some promise in this regard. It acknowledges the complexity of digital ecosystems by using a model that describes *intellectual entities* (coherent sets of information, such as books or Web pages); *objects* (discrete units of information, such as photographs, which may constitute parts of an intellectual entity); *events* (any preservation action); *agents* (any actor – person, organization, or software – associated with an event); and *rights* (permissions pertaining to an object or agent); and the relationship between these things, rather than looking at digital objects in isolation. PREMIS recognizes that digital content may lose viability if separated from its technological environment, and therefore requires thorough documentation of technological context and dependencies. It also allows the documentation of encoding and encryption, and the recording of fixity and authenticity measures. However, despite being developed as a practical implementation of the OAIS model, PREMIS is still both rather ungainly and largely untested – though trials are underway – and it may be assumed that it will be subject to revision.

What is clear is that digital stewardship in the broadest sense cannot be merely an optional add-on to existing positions and programs, but requires on the one hand dedicated staff and budgets, and on the other a systematic analysis of existing practices and priorities. It has been noted that

«digital preservation is not an isolated process, but instead, one component of a broad aggregation of interconnected services, policies, and stakeholders which together constitute a digital information environment»[16].

The same could be said of a digitization program as a whole. Indeed, it may be helpful to view digitization as an iterative and collaborative process rather than an event, and one that requires a cultural or perceptual shift. The development of such a program should be driven by certain principles, not least that the program be standards-based, in the interest of promoting interchangeability of data (between individuals, institutions and over time) and enabling interoperability between systems and applications.

L'articolo discute la lunga catena di operazioni che si svolgono dietro le quinte e precedono e seguono l'apparizione del patrimonio culturale sul World Web Web, soffermandosi in particolare sui problemi dell'accesso Web e della conservazione digitale. Molte istituzioni hanno avviato un processo di transizione che vede il passaggio da una digitalizzazione "a progetto" a una digitalizzazione "programmata", e si stanno adoperando per mettere in pratica strategie di digitalizzazione pienamente integrate e coerenti. La tecnologia digitale tende per sua natura a rompere barriere e nicchie da tempo consolidate, e ciò può rendere tale processo di transizione più difficile. Ogni programma di digitalizzazione andrà infatti con ogni probabilità a coinvolgere anche sfere tradizionali, quali quelle dell'acquisizione, della catalogazione e conservazione delle collezioni, della descrizione e dell'accesso, della distribuzione ed esibizione, della gestione dei diritti di proprietà intellettuale o dei diritti digitali. A tutto ciò occorre poi sommare la cura che le attività di cattura digitale, gestione e conservazione degli oggetti digitali richiedono in sé. I programmi di digitalizzazione devono pertanto essere basati su un consenso e una cooperazione di più ampia portata all'interno dell'istituzione coinvolta o tra diverse istituzioni, rispetto a quanto necessario nel caso di progetti digitali circoscritti e realizzati dall'uno o l'altro dipartimento. Una volta compiuto il lungo e difficile percorso che porta alla messa in onda di un sito Web, l'accessibilità delle collezioni continua a essere una questione complessa e priva di soluzione unica. Per rendere possibile una navigazione intelligente attraverso una massa crescente ed eterogenea di materiali, occorre combinare forme tradizionali di catalogazione, nuovi standard e protocolli per i dati, servizi di disponibilità full-text, tesauri e ontologie, e a volte anche forme di visualizzazione automatizzata e indicizzazione vocale.

[16] Brian Lavoie – Lorcan Dempsey, *Thirteen Ways of Looking at... Digital Preservation,* «D-Lib Magazine», vol. 10, n. 7/8, July/August 2004, *http://www.dlib.org/dlib/july04/lavoie/07lavoie.html.*

Cet article traite la longue chaîne d'opérations qui se déroulent derrière les rideaux et précèdent et suivent l'apparition du patrimoine culturel sur le World Web Web, en s'arrêtant surtout sur les problèmes de l'accès Web et de la conservation numérique. De nombreuses institutions ont lancé un processus de transition prévoyant le passage d' une numérisation «à projet» à une numérisation «programmée», et travaillent à tisser et mettre en pratique des stratégies de numérisation parfaitement intégrées et cohérentes. La technologie numérique, par sa nature même, a tendance à briser les barrières et les niches consolidées dans le temps c'est pourquoi le processus de transition peut être plus difficile. En effet, chaque programme de numérisation touchera probablement aussi des domaines traditionnels tels que l'acquisition, le catalogage et la conservation des collections, la description et l'accès, la distribution et exposition, la gestion des droits de la propriété intellectuelle ou des droits numériques. Il faut encore ajouter à tout ceci le soin requis par les activités d'acquisition numérique et de gestion et de conservation des objets numériques. Les programmes de numérisation doivent donc se fonder sur le consentement et sur une coopération plus vaste au sein de l'institution concernée ou entre des institutions différentes, que celle nécessaire aux projets numériques circonscrits et réalisés par l'un ou l'autre département. Même lorsque le long et difficile parcours menant à la mise en ligne d'un site Web a été achevé, l'accessibilité des collections reste une question complexe, sans une solution unique. Pour qu'une navigation intelligente soit possible parmi une masse toujours plus grande et hétérogène de matériaux il faut combiner les formes traditionnelles, les nouveaux standards et protocoles pour les données, les services de disponibilité full-texte, les thésaurus et les ontologies, et parfois même les formes d'affichage automatisées et l'indexation vocale.