The Digital Object: Hidden, Exposed

Stephanie Krueger
Associate Director for Library Relations
ARTstor
stephanie.krueger@artstor.org

INFORUM 2008: 14th Conference on Professional Information Resources
Prague, May 28-30, 2008

Introduction

In the current networked age, libraries are negotiating complex information property terrains formerly traversed only by publishers and commercial entities. Many kinds of digital objects now touch the hands of librarians, ranging from the so-called traditional (books and serials) to a wide range of "new media" types. Stewardship of these objects, in many cases already digitized before they reach a library or other institutional repositories, involves detailed description, presentation and – in some cases - preservation.[1]

While the challenges of description and preservation are great, as will be outlined in this presentation, the question of *presentation* of objects can be one of the most difficult for libraries to solve. To whom and for what purposes can objects be presented? Within a defined academic circle (i.e., campus or other institutional community), or to a larger audience? What rights are tied to each object, and what does this mean for their availability on the open web? Does it make sense for one's library to manage such rights, or should this function of rights clearance – so foreign to the culture of many libraries – be outsourced to other entities? If so, to whom? To commercial organizations or to other rights clearance specialists (non-profit and other)?

Advocates of local solutions often see the role of library as publisher to be crucial to a definition of a new kind of library; others look towards centralized partners to aid local object holders. Digital object repositories platforms, such as open-source Fedora[2], can work towards both aims, allowing a range of local services that can, if deemed necessary, interact with and leverage more centralized, rights-managed environments (e.g., ARTstor[3]). Organizations such as Creative Commons[4] are working towards solutions to some rights issues, but these will never be universal, particularly with differing national understandings of copyright and because of differing media types, some with potentially large commercial potential, such as music and video, and others with more limited use in commercial hands- at least in educational settings[5].

---

[1] Lynch, C. (2005). "Where Do We Go From Here: The Next Decade in Digital Libraries." D-Lib Magazine. Retrieved 17 March 2008 from: http://www.dlib.org/dlib/july05/lynch/07lynch.html

[2] www.fedora.org

[3] www.artstor.org

[4] http://creativecommons.org/

Open-source platforms, while theoretically free, in fact do impose costs upon libraries, in terms of technical expertise and infrastructure required to maintain them. In the short-term, the open-source benefits can seem great; however, libraries must also think through their repository problems with a long-term view and calculate the costs of implementation and possible future migration over time. These calculations must include consideration of the levels at which the objects will be stored and presented. Hidden objects ("dark archives"[6]) may have different levels of maintenance than those exposed ("light archive"[7]/"open access"[8] archives) and it is shortsighted to think objects now behind institutional barriers may not, in the future, be found desirable beyond these borders. Grades of presentation/access rules may, in some cases, be necessary when storing objects; for others, where rights situations are clear, such requirements within the repository will be less necessary, and these will impact any repository implementation.

Open-source/open access solutions, while theoretically desirable at many institutions, may simply not always be possible, and libraries must be prepared, when storing digital objects, to deal with other contingencies, as long as institutional technology realities and national copyright regimes remain diverse and – perhaps most importantly – if libraries hope to remain relevant.

Description

Objects Themselves

The variety of digital objects presented for storage in repositories is broad, and such a wide spectrum of materials poses a great challenge in terms of description of these objects. Metadata and cataloging standards vary across disciplines, and initiatives intended to standardize metadata schema such as Dublin Core[9] do not always meet the needs of certain communities. The US-based Visual Resource Association has, for example, created its VRA Core 4.0[10] in an effort to better-describe image-based content, noting how:

> VRA Core is more robust than Dublin Core for describing art images and metadata in this format is consequently more powerful but more expensive to create. VRA Core contains both 'work' records describing an actual art object, and 'image' records describing representations of views of that object (slides, digital images, etc.) held by an institution. Best practice in creating VRA Core

[5] Wagner, G. (2007). "Sharing Visual Arts Images for Educational Purposes: Finding a New Angle of Repose." EDUCAUSE Review. Retrieved 17 March 2008 from: http://connect.educause.edu/Library/EDUCAUSE +Review/SharingVisualArtsImagesfo/45225?time=1205805451

[6] http://www.cdlib.org/inside/diglib/glossary/?field=term&query=dark+archive

[7] http://www.cdlib.org/inside/diglib/glossary/? field=institution&query=CDL&action=search#L

[8] http://en.wikipedia.org/wiki/Open_access

[9] http://dublincore.org/

[10] http://www.vraweb.org/projects/vracore4/

> records is to populate fields using appropriate controlled vocabularies such as ULAN and TGM, and the rules described in Cataloging Cultural Objects <http://www.vraweb.org/CCOweb/>.[11]

For repository frameworks like Fedora, rules have been established for the ingestion of objects in its Fedora-specific extension of the XML Metadata Encoding and Transmission Standard (METS),[12] and a Dublin Core record is optional. Even METS is "too restrictive" for the purposes of placing objects in Fedora:

> Since METS was designed to be very generic and support a variety of uses, the rules of the METS Schema are very general-purpose. Fedora objects must conform to other rules that are beyond the scope of what is expressed in the METS schema. Therefore, the Fedora Object XML submissions will also be validated against a set of Fedora-specific rules that are expressed using the [Fedora-specific] Schematron language. Internally, the repository will use Schematron to enforce these rules on incoming XML submission packages.[13]

Any library or institution wishing to pursue organized object repositories will at some point face these descriptive challenges. The challenges only grow when considering linkages between objects and to other repositories through Open Archives Initiative (OAI) harvesting.[14]

Object-to-Object and "External" Relationships

Relating objects within any repository is often desirable, particularly when grouping objects into particular collections.

Fedora refers to such relationships as "object-to-object relationships", encoded in XML using the W3C's Resource Description Framework[15]:

> Fedora digital objects can be related to other Fedora objects in many ways. For example there may be a Fedora object that represents a collection and other objects that are members of that collection. Also, it may be the case that one object is considered a part of another object, a derivation of another object, a description of another object, or even equivalent to another object.[16]

---

[11] http://wiki.dlib.indiana.edu/confluence/display/INF/VRA+Core. ULAN is the Getty Union List of Artists Names, http://www.getty.edu/research/conducting_research/vocabularies/ulan/, and TGM is the Library of Congresses' Thesaurus for Graphic Materials, http://www.loc.gov/rr/print/tgm1/

[12] http://www.fedora.info/download/2.1.1/userdocs/digitalobjects/rulesForMETS.html

[13] http://www.fedora.info/download/2.1.1/userdocs/digitalobjects/rulesForMETS.html

[14] http://www.openarchives.org/OAI/openarchivesprotocol.html

[15] http://www.w3.org/TR/rdf-primer/

[16] http://www.fedora.info/download/2.0/userdocs/digitalobjects/introRelsExt.html

Such linkages are complex and often standards-based – IFLA's Functional Requirements for Bibliographic Relationships can, for example, be employed.[17]

Object-to-object linking takes effort, but so too does getting information about objects to other systems. In some cases, as in ARTstor's current hosting pilot, where objects from a variety of institutions are "ingested" into our system - a lack of metadata standardization – even between objects which have been catalogued using the same metadata schemes but interpreted in different ways – means necessary manual mapping to ARTstor's own data standards ("ARTstor Core"). Automating this process is, of course, the only way to scale ingest of objects to greater levels. The actual automation itself is a definable problem - the key hurdle, again, is the lack of standardization between descriptive formats. And standardization has a high cost, in terms of staffing.

Preservation (or: *Purpose*)

In this talk, I wish to focus more on access to objects than to delve into archival issues, and to focus most on the question: What is the purpose of any digital object initiative? I believe this question is often neglected. Taking the time to clearly articulate the purpose of any initiative can be crucial to its long-term success.

I am an avid reader of the lib-license listserv, and a recent posting (29 April 2008) illustrate attempts at defining the purpose of repository efforts. In the first, Yale's Ann Okerson refers to a 2002 White Paper[18] arguing the case for institutional repositories (IRs):

> A recent reading of Raym Crow's 2002 SPARC White Paper on IRs reminded me that he gave two principal reason for setting them up:
> 1. Serve as tangible indicators of an institution's quality and to demonstrate the relevance of its research activities, thus increasing the institution's visibility, status, and public value (what one could term administrative aggregation).
> 2. Provide tools to assist universities in re-shaping the scholarly communications process (what one could term a repair function).

Note how the second point raises the issue of the institution as *publisher*. I will return to this topic when discussing presentation issues, because this illustrates a shift in burden, to the institution or to a central hosting authority, in terms of managing rights issues.

Before considering those issues, however, let us return to the questions of purpose and preservation. How will the integrity of the source objects be maintained, and how will the objects be migrated through future technology changes? Will objects be available to a certain community while they are maintained ("light archive"), or will they be housed away from public view ("dark archive")?

[17] http://www.ifla.org/VII/s13/frbr/frbr.pdf
[18] http://www.arl.org/sparc/bm~doc/ir_final_release_102.pdf

A 2006 Tufts report on its work with Fedora illustrates the complexity of preservation decision making – note the emphasis on the environment in which the tool is implemented:

> In serving as the repository application of a preservation system, a Fedora instance (or instances) would be only one of many components that comprise a preservation system. Large portions of ingest and access activities and all preservation planning decisions, among other activities, would occur outside of the Fedora instance. Even though some preservation policies many be articulated and managed in Fedora, an institution still has to formulate these policies—they are not preset in Fedora. Rather than serving as an out-of-box repository solution, Fedora is a repository architecture upon which an institution can build a repository in many different ways. As a result, the suitability of Fedora as the basis of a preservation system depends significantly on its implementation. The question we should have asked was: "Can a Fedora repository, surrounded by the proper preservation policies, tools, and Fedora services, serve as the basis of a trustworthy preservation system?" We feel the answer to this question is yes.[19]

Even if an organization selects not to call itself an archive, it must still make decisions about migrating digital objects through time. At ARTstor, for example, we are not an archive – we are, however, a digital library committed to providing appropriate access mechanisms to our content over the long-term, in an economically-sustainable manner.

Consider the last phrase when thinking through your own implementations: how will you fund your activities in the long-term? What are the obvious and less apparent costs associated with any digital object project? Project-based funding is a risky business with limited time horizons, so make sure your purpose is clear.

Presentation

Finally, we have reached the area of presentation. Will your objects be hidden, or will they be exposed? Are there shade of gray along this continuum? Who will determine which objects are available, to whom?

I believe these issues are more complex – and potentially more costly – than the descriptive and technical challenges posed by any digital object initiative.

Take, for example, this excerpt from a news item about a specific Fedora implementation:

> 2007 marks the centenary anniversary of the National Library of Wales (NLW). What began as an experimental Fedora project at NLW over three years ago has become an ongoing program at this national library dedicated to preserving Welsh culture, heritage and knowledge for the people of Wales. Paul Bevan explained

[19] http://repository01.lib.tufts.edu:8080/fedora/get/tufts:UA069.004.001.00011/bdef:TuftsPDF/getPDF

> that NLW uses Fedora with VITAL. Challenges include ingest scalability of large collections that may include 1,000,000 images. NLW continues to look at ways to streamline overall processes to get more content 'in', especially now that other teams at NLW have discovered what the repository can do for them.
>
> *Rightscom* specialises in the provision of solutions for the management, trading and protection of intellectual property rights and digital content in the network environment.[20]

A commercial entity has, in this example, been employed to resolve rights issues. Who, at your library or institution, has the ability to manage issues across a variety of content types? If your objects relate to one another or are "harvestable", have you included the necessary rights information in your initial metadata descriptions? OAI does allow such rights encoding, but many decisions must be made about the actual rights statement which is applicable – at a repository level, or at an item level, or perhaps combinations thereof:

> If the expression of rights statements in OAI-PMH leverages semantics in existing metadata formats, a harvester will need to determine which rights statement in which metadata format is the applicable one. For example, it is entirely feasible that a selected Item might disseminate multiple metadata Records including multiple and possibly conflicting rights statements. Is only one applicable? Which one is applicable? Is some combination applicable, with questions again arising about difficult or nonsensical semantic combinations? To what extent should the OAI-PMH enforce rights consistency in records?[21]

At ARTstor, rights issues are central to our efforts, and we make a distinction between our central library objects (in which our legal team manages rights issues and where we, as an organization, take on the burden for use of objects entrusted to us) and hosted content. In the latter case, we function as an Internet Service Provider as defined by US Digital Millennium Copyright Act.[22]

Even determining which objects are theoretically available in the public domain can be daunting and involve multiple layers of investigation.

There are no simple answers to copyright issues posed by the storage of varying kinds of digital objects, which are perhaps even more complicated in your countries than in ours.

Conclusion

In closing, I wish to once again reinforce the complexity involved in managing digital objects. Open-source/open access solutions, while theoretically desirable at many institutions, may simply not always be possible, and libraries must be prepared, when

[20] http://www.fedora-commons.org/about/news.php
[21] http://www.openarchives.org/documents/OAIRightsWhitePaper.html
[22] http://thomas.loc.gov/cgi-bin/query/z?c105:H.R.2281.ENR:

storing digital objects, to deal with other contingencies, as long as institutional technology realities and national copyright regimes remain diverse and – perhaps most importantly – if libraries hope to remain relevant. Thank you.