Registering digitized monographs and serials
John Price Wilkin
4 May 2001, DLF Forum

Definition

The term "registry" is used in a variety of ways in the digital library community, all of them legitimately of course, but with such great differences that there is often confusion. In the DLI1, we talked about registries of services, in OAI we talk about registries of repositories, and elsewhere we have discussed registries of collections. The type of registry that I am talking to you about today is a registry of digital objects, a database of objects that an institution has digitized or intends to digitize. The purpose of this sort of registry is, simply put, to declare where we will now, or will in the future, invest our digitization resources, and especially our digitization resources for that class of materials we hold redundantly.

For some time, we have asked each other, and have been asked by others outside our community, how we know what has been digitized or what is in queue to be digitized in order to avoid duplication. It has been easy to offer facile responses. We know because we know each other, what we are doing, and where any of us would invest our resources. And this was often true. We could avoid duplication of effort by simply asking our small community of similar institutions involved in similar activities. When, at the outset of the first Making of America project, Cornell planned to digitize a large monographic series, they asked and easily determined that the series had been converted for a commercial project—a nuisance, but not a disaster. And yet, as more institutions took on more, the likelihood of redundant efforts increased, and in the last year we have seen some notable instances of near misses, and major near misses in some cases. We cared intensely about coordinating our microfilming activity, even though microfilm does not promote remote access (or not in the same way), and yet we do not have in place similar mechanisms to prevent duplication of digital efforts. Rather than asking "why?" it seems appropriate to note that our evolution has made the question more compelling today, and perhaps a question that we are willing to address.

If the primary function of a registry is to declare the availability of digital objects, or object that we intend to digitize, are there areas or types of materials for which the registry is best suited? Presumably, a registry is most useful in those cases where there is a likelihood that our efforts could be duplicated. Discovering the digital availability of a uniquely held photograph or papyrus is indeed useful, but not nearly as useful as learning that Cornell has digitized a run of Harper's Magazine when I am considering that title in a conversion project. Consequently, discussions about the possibility of establishing a registry have typically focused on finding ways to declare information about books and journals, and especially commonly held books and journals.

This leads me to my last point about registries before beginning my discussion about the meeting that took place in early April: I would like to digress for a moment to explain one important example of what a "registry" is not. When discussing registries, we often find ourselves slipping
into parallel discussions about "discovery" (for users) or the development of cross-collection searching of our digital libraries. These issues are of course related, and we can easily see how they might be articulated as layers on top of a functioning registry--that is, a registry could facilitate "discovery." It is helpful and even important, though, to separate out these issues and treat each with its own imperatives. We are, as a community, finally making headway on metadata harvesting and finding ways to provide an aggregate sense of what is online. The requirements and functions of "discovery" are so different, though, that we need to put them aside, at least momentarily. A registry is about declaring the results of our conversion efforts to each other.

Report of the 11 April 2001 Meeting

Overview

In April, DLF convened a meeting of persons from institutions known to be active in large-scale book and journal digitization efforts. The institutions and organizations represented included:

- CLIR
- Cornell University
- Harvard University
- JSTOR
- Library of Congress
- The British Library
- The California Digital Library
- The CIC
- The University of Michigan
- The University of Virginia
- The University of Wisconsin
- Yale University

The meeting occurred over most of one day, and there seemed to be considerable consensus on the value of a registry as well as, it seemed, a common conception of how that registry might work. The increasing likelihood of redundant effort, in particular, seemed to provide a new impetus to a discussion that is certainly not new. The new interest seems to have grown out of the recognition of

1. the costs that libraries could potentially avoid if such registry services were in place (e.g. redundant digitization)
2. the new services and service functions that libraries could potentially supply by reallocating even a fraction of the avoided costs.

---

1 The summary of the meeting was borrowed overwhelmingly from Dan Greenstein's notes to the participants.
DLF’s expressed interest in hosting the meeting was to help develop a registry service to some prototype stage if a compelling case could be made for that investment either in terms of cost avoidance, new service, or similar benefits.

**High-level statement of aims**

In summarizing the meeting for the participants, Dan Greenstein offered a high level statement of aims that effectively captured the gist of the discussion. In articulating the aims of the registry he said:

A service that records information about digitized books and journals may be a key infrastructural part or utility in an evolving network of organizations and services that support the efficient and responsible stewardship of our cultural heritage, all formats, old and new, and the economical and effective development of high-quality scholarly collections.

**Chief characteristics of a registry service**

1. Records information about digital surrogates (whether in existence or about to be created) for books and journals (in all languages and on all topics), that is for objects that are collected redundantly by libraries.
2. By recording information about a digital object in a registry service, an individual or institution records their intention to ensure that the digital object persists.
3. Digital objects referenced in the service must be available to users, that is, accessible. The objects need not be freely accessible (thus not excluding JSTOR or responsible commercial entities).
4. Terms and conditions of access must be recorded for information referenced in the registry service according to some agreed mechanism.
5. Records in the registry service must include a persistent link to a “use-copy” of the relevant digital object. Where archival master copies exist, they will be indicated in the record but need not be accessible.
6. Rather than prescribe minimum requirements pertaining to the characteristics of digital objects that are referenced in a registry service, (e.g. formats, terms and conditions of use, etc), the service will simply implement agreement about how to record such information. For example, we might say that "Method X" consists of 600dpi bitonal TIFF images with G4 compression, associated page level metadata including pagination, sequence, and "function" (e.g., table of contents pages).

**Relationship of Registry to other services**

As we discussed it, we saw the Registry as a part of other key services. As I mentioned earlier, it is not intended to offer end-user services. We expect that its existence will potentially encourage the development of a wide range of end-user services that may include:
1. content services that aggregate or otherwise leverage off of existing digital content (e.g., using OAI)
2. print-on-demand services
3. copyright clearance services

Moreover, the registry (as a key service infrastructure) would ideally exist within and interoperate with other key pieces of service infrastructure including:

1. catalogues of books
2. microfilm registries
3. print repositories
4. digital repositories
5. digitization services

What we saw the registry serving as is a starting point. By separating out the functions and services into distinct components, we felt it would help to focus our efforts in a way that will allow us to assess key assumptions and technologies with a definable set of information content. While it is conceivable (and perhaps even inevitable) that such a service could be extended to include audio-visual and other non-unique materials, this starting point will begin to address some pressing needs and allow us to grow it and the surrounding services effectively.

The registry service would enable institutions individually and collectively to:

1. locate information and potentially access digitized books and journals
2. avoid redundant digitization effort
3. co-ordinate digitization efforts (e.g. by divvying up responsibility for digitizing a common body of materials)
4. co-ordinate print deposit/preservation effort
5. support economical institution-level collection development decisions viz
6. acquisition / disposition of printed materials
7. digitization of books and journals
8. support a range of end-user services
9. identify collaborative opportunities

Other benefits/uses envisaged for the service include its

1. support for incremental development/improvement of existing digital objects. (Consider, for example, the problem of missing pages in brittle book reformatting. In microfilming, an institution needs to hold out a copy of an item until it can be fully represented. Digitization will allow us to reformat and then declare missing pages so that an institution can volunteer that information, and thus complete the volume.)
2. formal disclosure of preservation practice as it evolves and support for ancillary community discussion and debate about what constitutes good practice. (There are certain minimums we
can more easily reach agreement on. Begin with these and build toward more challenging or complex issues.)  

3. cross-fertilization with commercial data producers and suppliers who, as contributors to and users of the registry would be sensitized to community awareness of needs, good practices, etc  

4. support for a range of end-user services as described above

I should also note that we saw the registry being used primarily by

1. Collection managers, and  
2. Service providers who would build end-user services that rely on the registry’s existence

**Research issues that remain to be investigated**

Along the way we of course encountered stumbling blocks. Dan helped to grease the skids by identifying them as "research issues" that can be dealt with in parallel to the main body of work, but that need not be addressed to move forward. Some worth noting include:

1. **Strategic issues**
   - Costs of building/maintaining a registry (comparable data may be available from other registry and cataloguing efforts)  
   - Costs that may be avoided by libraries and others if a registry service existed  
   - How the existence of a registry service would leverage existing investment e.g. in print collections, digitization, and digital and print repositories  
   - Models for organizing and sustaining a registry service

2. **Metadata issues**
   - Collection level descriptions; their structure and possible use in a registry service  
   - How to describe an intention to ensure persistence of a digital object that is referenced in the registry  
   - What granularity for registry entries (journals will be particularly challenging)?  
   - How will the registry make it possible to update records, for example to reflect changes in access or preservation copies of the digital content?

3. **Other issues**
   - How is information in the registry accessed?  
   - What inter-relationship exists between information about digital surrogates as recorded in the registry and information about print and microfilm editions as recorded in bibliographic and microfilm catalogues and registries respectively  
   - Does the registry include references to digitized newspapers?
Next Steps

In light of this consensus and of the potential benefits seen in a registry service, participants agreed to five further steps that should be taken to develop the registry and hoped the steps could be taken in an 8- or 12-week period. Probable stakeholders are noted in parentheses.

1. Develop a brief and compelling summary statement describing aims, goals, and potential of a registry service. The case can be used to generate support among key stake-holding groups. Some of the elements of a supporting case for a registry include:

2. Leveraging existing investment in digital content (libraries and their owners)

3. Helping to rethink collection development and management costs (library managers and their owners)

4. Helping more economically to re-think preservation / persistence strategies (library managers)

5. Maximum value for funds spent on digitization (funding agencies and others investing in digitization)

6. Key infrastructural component of national print and digital preservation strategies (information producers, information users, and repository managers/libraries)

7. Key support for a new generation of end-user information services (information users but also information providers and libraries)

8. Exploration of costs involved in doing nothing; that is, in continuing as we are (all)

9. Professional development/training and awareness-raising

10. Hold an expert workshop to develop a detailed functional specification for the registry service as defined above. This work should begin by reviewing Michigan's Making of America records as currently recorded in OCLC. We suspect the records are clearly inadequate to help determine what the corresponding digital object is, and in determining the short-comings, we may be able to more easily create those functional specifications (as well as determine what more OCLC would have to do to support the registry functions).

11. We should approach OCLC to discuss possible role for OCLC developing a registry service.

12. Hold an expert workshop to document the extended metadata set required by the registry, as well as the mechanisms and incentives for creating and supplying those metadata, and discusses issues of granularity, metadata updating, etc..

13. Hold an expert meeting to review existing preservation reformatting guidelines with a view to identifying agreed benchmark practices if possible. The review should include institutions with such guidelines in hand (e.g., the institutions represented at the meeting), as well as efforts like METS, presented elsewhere at this Forum.