LIBER QUARTERLY 18 (1), April 2008 – ISSN: 1435-5205. P39-48
http://liber.library.uu.nl
Igitur, Utrecht Publishing & Archiving Services

Automating Registration of Digital Preservation Copies: The Place of Registries in the Digitization Workflow

William Carney

Content Manager, Business Development Division, OCLC,
6565 Kilgour Place, Dublin, Ohio 43017, USA, carneyb@oclc.org

Introduction

I would like to thank LIBER and EBLIDA for inviting me to present this paper on the role of registries in the digitization workflow.

During the past 18 months, OCLC has been working on a project to synchronize WorldCat with mass digitization projects, which we will begin to pilot shortly. The concept is to educate WorldCat about the millions of new digital manifestations for print items being produced. During the past 35 years, librarians have built a comprehensive representation of print materials and holdings in WorldCat item-by-item. However, as we move into a more digital world through the production of born-digital materials and the high-volume reformatting of our print heritage, it is impractical to catalog these new manifestations via traditional workflows.

The OCLC eContent Synchronization program is one example of how OCLC is moving to address the need to ingest metadata representing digital works on an industrial scale. Through strategic alliances with key digital content producers and automated processing, new digital surrogate records will be created to increase the visibility of and access to content at the point of need.

While the eContent Synchronization program is an important initiative for visibility and access, of equal importance is the process of registering the existence of the preservation copies of these digital items.

The Role of Registries

Registries have enjoyed a prominent role in libraries since the time of the early Assyrian Kings (and likely before). Assurbanipal (668–631 B.C.) ordered that the palace collection of clay tablets be registered in a subject catalog of sorts. 'These tablets include entries giving titles of works, the number of tablets for each work, the number of lines, opening words, important subdivisions, and a location or classification symbol.'[1] The card catalog, the OPAC and indeed WorldCat itself are more recent examples of registries describing items held by libraries. But the number and role of registries is expanding beyond that of describing items held. For example, the new WorldCat Registry describes 'the institution itself: its identity, electronic services, relationships, staff contacts and other pertinent data that informs the processes and systems driving [the library] … enterprise.'[2] The objective of the OCLC Registry of Copyright Evidence, currently in the early development phase, is to reduce the time and costs associated with investigating the copyright status of potential digitization items.

When considering the role of registries in the digital preservation process, it is perhaps best to start with a clear statement of purpose, as ably provided by the Digital Library Federation (DLF): '… Staff engaged in digitizing efforts should be able to discover whether a specific item has already been digitized, and if so whether the digitization has been done at an adequate level such that another digital copy is not required.'[3]

At its most basic level, the need can be translated into one of economics. The costs associated with digital preservation can be staggering when one considers the resources needed for selection, logistics, scanning, OCR, quality control, access and preservation storage. Cooperation at the earliest possible time is crucial to eliminate redundancy and to focus scarce resources where they are most needed.

Few would argue the validity of the concept. In fact, the idea for such a registry is not a new one. 'Keyes Metcalf, [of] Harvard University, first proposed establishing a national register of microform masters in 1936.'[4] The European Registry of Microform Masters (EROMM) and the National Register of Microform Masters (NRMM) provide successful examples of collaboration to reduce duplicate conversion of print items to microforms. WorldCat contains nearly 600,000 NRMM records and more than 400,000 EROMM records.

Armed with a valid concept that has stood the test of time, it would seem logical to assume that a registry of digital masters would enjoy similar success. However, the evidence suggests otherwise.

The DLF/OCLC Registry of Digital Masters

The DLF/OCLC Registry of Digital Masters first became available as a searchable resource in 2004. Guidelines for standard surrogate records were created by a DLF working group in July 2004 and reissued in May 2007.[5]

In 2006, LIBER, OCLC and EROMM announced the intent to load all of the EROMM digital preservation records into the Registry of Digital Masters. The process to load these records is in the planning stages, with the intent to apply updates to these records in an automated process following the new DLF guidelines.

Renette Davis, Head, Serials & Digital Resources Cataloging, University of Chicago Library describes why they are contributing to the registry: 'We are putting records into the Registry of Digital Masters now for everything which we digitize in the library and that has MARC records. We want to notify our colleagues in other institutions that we intend to preserve these digital resources so they don't have to spend their money digitizing the same thing. We also see the Registry of Digital Masters as an additional way of exposing our digital collections. We spend a significant amount of money digitizing unique materials from our Special Collections Research Center, and we want to make these available to as wide an audience as possible.'[6]

Unfortunately, with a few notable exceptions, the registry has seen little regular usage, and contains just slightly more than 4,100 records. This low rate of contribution leads to a general perception of little value, which in turn becomes a self-fulfilling prophecy as users decline to come back after their first visit. In response to a query from the author, Oya Rieger, Interim

Assistant University Librarian, provided her thoughts on this issue: 'Although we were one of the first libraries to contribute records to the registry, I think we have not started using it systematically due to two key reasons: (1) Unclear/unproven nature of the institutional and community benefits behind contributing records to the registry. I am afraid the registry is still being perceived as a test bed. (2) Unknown nature of what is involved in the process from an institutional perspective – resources and time required to contribute records to the registry.'[7]

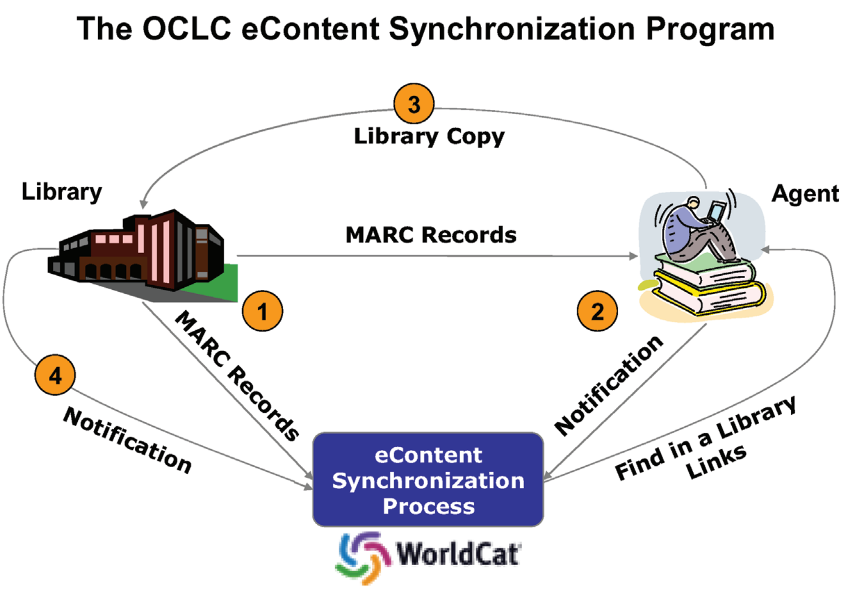
Susan Westberg, OCLC product manager for the registry, offered these additional thoughts: 'Reasons for slow adoption may be finding the right project to start; analyzing workflows; making changes to workflows; other priorities; lack of resources, such as staff; changes in direction; or the impetus for joining has changed. Be that as it may, the registry now has some 4,100+ records and the number of participating institutions has expanded from the original five institutions to nine institutions.'[8]

A general change in the environment since the concept for the registry was introduced may also be a contributing factor. Prior to the advent of mass digitization projects such as Google Book Search and Microsoft Live Search Books, transformation from paper to digital or to microform occurred at a much slower pace and was focused on brittle or special collections. Today, entire library collections are being scanned with incredible speed resulting in volumes unimagined even three years ago. The registry was built with the idea that organizations would individually select materials, alert colleagues of their intent to digitize and preserve via the registry, and then update the record when the task was completed. Mass digitization projects have accelerated the time between selection and transformation to such a degree as to make registration of intent superfluous. The manual process of updating the digital record in a timely fashion is impractical. However, these environmental changes are precisely the reason why the registry will thrive in the near future.

A Bright Future for the Registry

Given the minimal usage of the registry to date, one might be surprised by the title of this section. However, the need for a registry is stronger today

Fig. 1:



than it was in Metcalf's time. Not all preservation efforts fall within the realm of mass digitization. Many smaller, more selective projects are taking place. But resources are scarce; even more so now that many librarians are retiring, making manual surrogate creation less realistic.

How then can the registry be successful in its mission? The answer lies in building a critical mass within the registry quickly. This will lead to an increase in the value and usefulness of the registry, and in removing the test bed stigma pointed out by Oya Rieger. This can be done as a direct result of the OCLC eContent Synchronization program. As described earlier in this paper, the program is designed to create digital manifestation records for mass digitized content, with links to the host Web page describing the item. Certainly, these manifestations were never intended to be preservation copies, and indeed they are not. However, in nearly all cases, mass digitization agents are returning to participating libraries copies of the transformed items that can be preserved. While not all libraries are choosing to preserve their copies immediately, many are. The final stage of the eContent Synchronization process will create a digital manifestation record for the library copy; and for those libraries that are preserving the item, the process, when in production, will *automatically* apply the updates based on the DLF guidelines.

As shown in Figure 1, phase one of the process harvests MARC records from participating libraries representing the print item. The records are matched against WorldCat to create a table of identifiers and OCLC print record numbers. After receiving notification that an item has been digitized and is available on the host site, phase two of the process creates a digital manifestation record in WorldCat. Finally, phase three creates a Registry of Digital Masters-compliant record in WorldCat following notification that the library has received its digital copy and is preserving it.

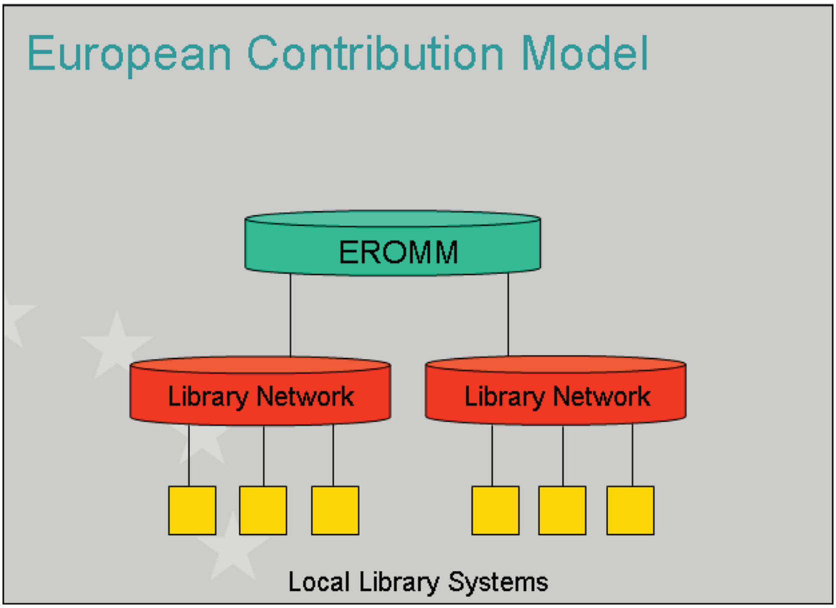
The result will be a large influx of records representing digitized and preserved items. Organizations planning preservation projects will now be able to avoid digitizing items that have already been transformed, enabling them to focus resources on other items.

A Suggestion for a New Model

The collaboration among LIBER, OCLC and EROMM is an example of the library and cultural heritage communities thinking globally to address the need to eliminate duplication in preservation efforts. Janet Lees, Community Liaison, OCLC describes the implementation of the initiative in her 2005 paper as follows: 'EROMM … act[s] as a coordinating centre for the registration of digital masters contributed from European libraries working within their own national library networks. These records … [are] exchanged with the DLF/OCLC Registry and possibly others to create global coverage.'[9]

Once the records are contributed, EROMM forwards the records to OCLC for ingestion into WorldCat and the Registry of Digital Masters. As noted earlier in this paper, EROMM has indeed provided OCLC with over 400,000

Fig. 2:

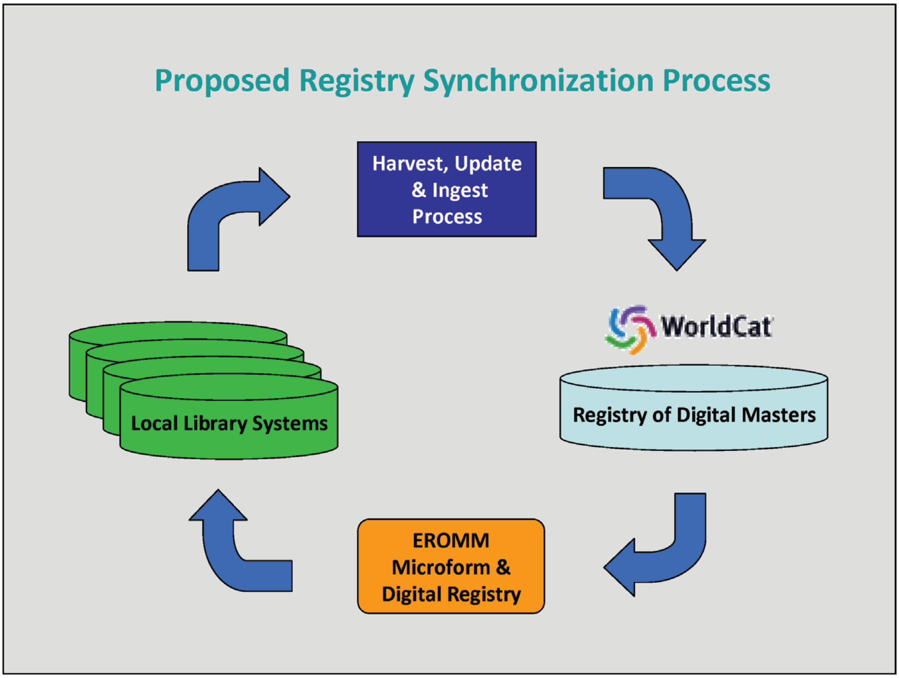


European Contribution Model

records and they are in WorldCat today. However, the 9,000 records representing digital items are not currently compliant with the DLF guidelines. In an effort to move these records into the registry quickly, OCLC and EROMM are discussing a process that will update the records in a similar fashion to the last phase of the eContent Synchronization process. Once complete, the records would be discoverable in the registry as well as in the EROMM database.

Going forward, this model of contribution from local libraries to regional networks, then through EROMM to OCLC can continue, but a model similar to the eContent Synchronization process could introduce additional efficiencies. By inverting the model, a process could be implemented to harvest the metadata directly from libraries and transform the records to meet DLF guidelines. Once complete, the records would be discoverable in

Fig. 3:


Proposed Registry Synchronization Process
Harvest, Update & Ingest Process
WorldCat
Registry of Digital Masters
EROMM Microform & Digital Registry
Local Library Systems

the Registry of Digital Masters and in the EROMM registry through the normal EROMM/OCLC record exchange.

Conclusion

As the pilot for the eContent Synchronization program begins in early 2008, users of the Registry of Digital Masters will see an exponential growth in the number of records available within the registry. The program is open to all parties interested in synchronizing their digitized collections with

WorldCat. As more collections are added, the value of the Registry of Digital Masters will continue to increase, saving resources and enabling preservation activities to be focused on those items unsuited for mass digitization, e.g., brittle, oversized or very valuable items. The registry concept will once again be validated as it was for microforms years ago.

Websites Referred To In The Text

DLF, Digital Library Federation, http://www.diglib.org

DLF/OCLC Registry of Digital Masters, http://www.oclc.org/digitalregistry

EROMM, European Registry of Microform Masters, http://www.eromm.org

NRMM, National Record of Microform Masters,
http://www.arl.org/preserv/presresources/Microform_masters.shtml

OCLC, Online Computer Library Center, http://www.oclc.org

WorldCat, http://www.worldcat.org

Notes

[1] Harris, Michael H.: *History of Libraries in the Western World*, Metuchen, NJ: Scarecrow Press, 1995, pp. 19-20.

[2] http://www.oclc.org/worldcat/registry/.

[3] Digital Library Federation: *Registry of Digital Masters Record Creation Guidelines*, May 2007, http://www.diglib.org/collections/reg/reg.htm (accessed August 16, 2007).

[4] Reed-Scott, Jutta: 'OCLC RetroCon project for preservation microfilm masters completed', *OCLC Newsletter* (233): May/June 1998, http://digitalarchive.oclc.org/da/ViewObjectMain.jsp;jsessionid=84ae0c5f824088e0f34bd83c4f1eb3f1aa4fa3f80bce?fileid=0000001714:000000043682&reqid=46 (accessed October 11, 2007).

[5] http://www.diglib.org/collections/reg/DigRegGuide200705.htm.

[6] Excerpt from e-mail correspondence between the author and Renette Davis (Head, Serials & Digital Resources Cataloging, University of Chicago Library), October 16, 2007.

[7] Excerpt from e-mail correspondence between the author and Oya Y. Rieger (Interim Assistant University Librarian, Digital Library and Information Technologies, Cornell University Library), October 12, 2007.

[8] Excerpt from e-mail correspondence between the author and Susan Westberg (Product Manager, OCLC Online Computer Library Center, Inc.), October 10, 2007.

[9] Lees, Janet: 'OCLC Registry of Digital Masters – Opportunities for European Cooperation', *LIBER Quarterly* 15 (2005) 3/4, http://liber.library.uu.nl/publish/issues/2005-3_4/index.html?000139 (accessed October 16, 2007).