ARTICLE
EXCERPTED
FROM:

ISQ INFORMATION STANDARDS QUARTERLY

SPRING 2010 | VOL 22 | ISSUE 2 | ISSN 1041-0031

SPECIAL ISSUE: DIGITAL PRESERVATION

DIGITAL PRESERVATION METADATA STANDARDS

TRUSTWORTHY DIGITAL REPOSITORIES

UNIFIED DIGITAL FORMATS REGISTRY

AUDIO-VISUAL DIGITIZATION GUIDELINES

DIGITAL PRESERVATION EDUCATION

NISO
How the information world
CONNECTS

THE UNIFIED DIGITAL FORMATS REGISTRY

ANDREA GOETHALS

Why do we need a format registry for digital preservation?

If you diligently protected a WordStar document for the last twenty-five years, all of its original bits may still be intact, but it would not be usable to anyone. Today's computers do not have software that can open documents in the WordStar format. It's not enough to keep digital bits safe; to fully preserve digital content we must make sure that it remains compatible with modern technology. Given that the ultimate goal of digital preservation is to keep content usable, practically how do we accomplish this? Somehow we need to be able to answer two questions: (1) is the content I'm managing in danger of becoming unusable, and if so, (2) how can I remedy this situation?

Formats play a key role in determining if digital material is usable. While traditional books are human-readable, giving the reader immediate access to the intellectual content, to use a digital book, the reader needs hardware that runs software, that understands formats, composed of bits, to access the intellectual content. Without technological mediation, a digital book cannot be read. Formats are the bridge between the bits and the technologies needed to make sense of the bits. The formats of the bits are the key to knowing if there are technologies that can make the bits usable.

Returning to the question—Is the content I'm managing in danger of becoming unusable?—the question can be answered if we know the formats of the content we're managing, and additional information about those formats. We

CONTINUED »

Formats are the bridge between the bits and the technologies needed to make sense of the bits.

CONTINUED »

need to know if there are current acceptable technologies that support the formats, sustainability issues related to the formats, and how others in the digital preservation community have assessed the formats. If we determine that the content is in danger of becoming unusable, we can form a remediation plan if we have additional information about the formats. We need to know alternative formats for the content, supporting transformation or emulation tools, and as a last resort, enough documentation about the format to construct our own tools to transform or render the content.

All institutions engaged in long-term digital preservation need this same format information. The concept of the format registry is simple—pool and share the data so that each institution does not have to collect and manage this information for itself, and does not need in-house expertise for all the formats it needs to manage. Additionally, because the format registry would provide authority control for format names and identifiers, it would enable institutions to more easily share file tools and services, and exchange content.

History of the format registry initiative

The first planning sessions for what came to be known as the Global Digital Format Registry, or GDFR, were sponsored by the Digital Library Federation (DLF) in 2003. These meetings were attended by policymakers and technologists from various national libraries and archives, academic research libraries, universities, library organizations, and standards bodies. Out of these meetings came a clear rationale for a shared format registry, over thirty use cases demonstrating how the registry could be used in preservation operations, and preliminary designs.

Following those meetings, Harvard University agreed to seek funding for and host the first instance of the registry. The Mellon Foundation funded a two year project beginning in 2006, and the development was subcontracted out to OCLC. The project produced a very detailed data model, and a registry model based on shared governance, cooperative data contribution, and distributed data hosting. When the project ended in 2008, a proof of concept registry at Harvard containing a limited amount of format information was made available on the Internet.

Following the project, Harvard began to plan next steps for the registry. The proof of concept registry would need additional technical work to turn it into a full-fledged registry. In addition, there were a number of governance issues still to be resolved to make the registry sustainable. It would need long-term administrative, operational, and financial resources. The reality, however, was that the registry landscape had changed a great deal from when the GDFR project began. Now there was already in existence another format registry that was being used by many in the preservation community: PRONOM.

PRONOM, developed by The UK National Archives (TNA), was created to meet TNA's requirements, but the registry information was freely shared on the Internet. Like the GDFR, PRONOM contains information about formats as well as related software, hardware, media, documents, and organizations. It's not a coincidence that the GDFR and PRONOM data models are similar. TNA was a significant contributor to the GDFR effort and the GDFR and PRONOM teams shared data model information so that they would be compatible. The intention was that PRONOM would become a node in the GDFR network of format registries when GDFR became fully operational. However, in 2008 when Harvard started to look at next steps for the GDFR, it was clear that PRONOM was further along technologically and in terms of use by the preservation community. But because PRONOM is owned and maintained by a single institution, it was not possible for other institutions to contribute information to the registry, and the community had become reliant on a single institution for sustaining an essential piece of preservation infrastructure.

This was the dilemma: neither GDFR nor PRONOM alone was fulfilling the long-term requirements for the digital preservation community. The community needed the format information and services already provided by PRONOM but also wanted the shared governance, cooperative data contribution, and distributed data hosting promised by GDFR.

Progress: UDFR established

In early 2009, the National Archives and Records Administration (NARA) hosted a format registry planning meeting, which included members of the GDFR and PRONOM teams. In this meeting it was agreed that it would be advantageous for all to combine the PRONOM and GDFR initiatives into a single registry—the Unified Digital Format Registry. UDFR would include the services and data of PRONOM and support the shared governance, cooperative data contribution, and distributed data hosting of GDFR.

The work required to establish the UDFR falls into two general categories: governance and technical work. The governance work includes designing and implementing the plan for ongoing UDFR governance, funding, and operations. The technical work includes the design, development, and testing of registry software and processes needed to exchange registry information with tools, services, and repositories. To address this work, an interim governing body and a technical working group were formed consisting of members from national and academic libraries, universities, and national archives who had participated in the earlier registry initiatives. These groups formed a plan to quickly put into place an operational first version of the registry, while working in parallel to replace the interim governance body with a permanent governance structure for UDFR.

Working from documents created for the GDFR and PRONOM projects, the technical working group compiled the requirements that should be implemented in the first version of the UDFR:

- A publicly accessible web-based user interface that can be used to search, browse, display, and download registry records
- An API for tools and services to query, retrieve, and export registry records for use in local repositories or applications
- Ability to export information to DROID, a format identification tool created by TNA
- Automatic tracking of the history of registry information changes
- Population of the registry with all of the PRONOM content

Near the end of 2009, the governance working group submitted a proposal to the Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP) to fund the one-year program of technical work needed to establish the first version of the UDFR. Under the proposal the work would be conducted at the University of California Curation Center (UC3) of the California Digital Library (CDL). UC3 will provide project oversight and management and will hire two new staff for the project—a project architect and a developer. The proposal was accepted by the Library of Congress in early 2010 and UC3 has now begun the hiring process for the project, which is scheduled to run from July 2010 to July 2011.

Future plans: UDFR and beyond

In parallel to the technical work that will occur at the UC3 over the next year, the interim governance working group will establish the permanent governing body for the UDFR. This permanent group is needed to define registry policies and procedures, such as the editorial process to ensure registry information is accurate, how future enhancements will be defined and prioritized, and intellectual property policies related to the registry software and information. In addition, this group is needed to fund the UDFR's administration, maintenance, and future enhancements.

A key future enhancement is to transform the initial UDFR design into a network. Initially there will be a single registry instance hosted by UC3. However, the long-term goal of the UDFR project is to establish a network of registry instances operated by various institutions around the world, with automatic processes to copy the UDFR content among the registry instances. This will increase the safety of the registry information and reduce the dependency on any single institution.

The initial version of the UDFR will provide interoperability with existing applications used for digital preservation. It will supply format identification information to DROID, and it will provide export services that could be used to import format or environment information into local repository databases. Ultimately though, it is the intention that the UDFR will serve as a source of format information to many tools and services that will be developed by the preservation community over time for format identification, assessment, validation, characterization, transformation, delivery, and emulation. | FE |

doi: 10.3789/isqv22n2.2010.04

ANDREA GOETHALS <andrea_goethals@harvard.edu> is Digital Preservation and Repository Services Manager at Harvard University Library.

RELEVANT LINKS

DROID
sourceforge.net/projects/droid/

Global Digital Format Registry
www.gdfr.info/

NDIIPP
www.digitalpreservation.gov/library/

PRONOM
www.nationalarchives.gov.uk/PRONOM/

Unified Digital Format Registry
udfr.org

University of California Curation Center (UC3)
www.cdlib.org/services/uc3/