An evaluation of the building of a new tool
to test the scalability of a trusted digital repository

Sinéad Redmond and Damien Gallagher

The **Digital Repository of Ireland (DRI)**[1] is an interactive national trusted digital repository for contemporary and historical, social and cultural data held by Irish institutions; providing a central internet access point and interactive multimedia tools, for use by the public, students and scholars. DRI is a four-year exchequer funded project, comprising six Irish academic partners, and is supported by the National Library of Ireland, the National Archives of Ireland (NAI) and the Irish national broadcaster RTÉ.

One of the aims of the DRI project is to construct a *trusted* digital humanities repository. A key requirement for a repository to qualify as trusted is that the users must be able to rely on it to remain functioning as expected, even under very high data loads.[2] [3] Thus, the challenge of testing the scalability of prototype solutions with large-scale datasets arose. As the DRI project is at an early stage of development, assigned datasets have yet to be repurposed for ingestion into DRI. Five test projects had been identified, named 'demo projects', which would come on board during the DRI project timeline, but these projects are coming on board in staggered stages. Furthermore, the metadata attached to these demo projects were either not created yet or not in suitable formats for use in the DRI. One project used a custom database schema that would not meet international metadata standards.

In this situation, the most obvious method of providing large-scale datasets with attached metadata to test the scalability of the prototype repositories was to simply create the metadata of the demo projects. However, there were drawbacks associated with this; primarily, the work (and therefore the cost) associated with the accurate creation of this metadata is very substantial. Who would be responsible for this work was also in

---

[1] Digital Repository of Ireland Home Page, http://www.dri.ie

[2] An RLG-OCLC Report, Attributes of a Trusted Digital Repository: Meeting the needs of research resources (August, 2001) p.16, 17 http://www.oclc.org/research/activities/past/rlg/trustedrep/attributes01.pdf accessed 31/10/2012

[3] DARIAH Policy Paper, Policy on Compliance as a Trusted Digital Repository (January, 2010) p. 7 http://dariah.eu/index.php?option=com_docman&task=doc_download&gid=472&Itemid=200 accessed 31/12/2012

question? If it were required by a specific time to test the prototypes' scalability and reliability, then perhaps it should have been the software engineers who require it responsible for its creation?

This, though, was not a viable solution; the software engineers are not archival specialists, are not familiar with the datasets of the demo projects, and were not in a position to be held accountable for the accuracy of the metadata. It would also have added hugely to the workload of the software engineers to mandate that they create this metadata, which was not a reasonable option either, given that there was already a sizeable existing, agreed and stipulated workload in existence for them. If, alternatively, the demo projects were required to generate the metadata, it would be equally unreasonable of the project to demand that the demo projects invest substantial resources much earlier than they had initially agreed to budget for.

Thus, a solution of building a tool called Hydrate, that would generate large sets of dummy data and metadata was proposed. These large-scale datasets would be ingestible into the digital repository prototypes for scalability testing of these prototypes. The proposed tool would be written in Rails, in order to sit within the Hydra framework already agreed upon for use in the DRI project. Java was also examined as a potential implementation language, but discarded in favour of Rails in order to ensure greater integration with the DRI project as a whole, as well as a higher likelihood of future use in other Hydra-based projects. This solution would also allow for the requirements already identified for the DRI to be tracked and tested properly in metadata generation.

Hydrate feature list:

- Hydrate creates a collection of data with associated metadata for ingest into the DRI repository.
- Hydrate allows the user to define what data it should output.
- Hydrate allows the user to select the generation of different data types; e.g. audio files, image files, text files, or video files, or a combination of these.
- Hydrate allows the user to select a metadata standard to generate associated metadata for each audio/image file generated.
- Hydrate allows the user to select of the quantity of data to be generated.
- Hydrate allows the specification of individual record size to be generated.

- Hydrate allows the specification of record quantity to be generated.

Hydrate was initially developed to generate Dublin Core or EAD metadata, depending on which of these options the user selects. From the requirements gathered from the DRI's national survey of humanities and social science institutions[4] it is clear that these metadata standards have become widely adopted in Ireland and choosing these two metadata formats allows Hydrate to be immediately beneficial to archives that intend to follow national best practice. However, Hydrate is expansible to allow for the inclusion of different metadata formats in future, should this option be required for the DRI at a later date. It is also a key point that Hydrate allows the user to specify different kinds of breakdown of data size to be output, as a repository of large size files and a repository of smaller size files, with the same amount of memory used by each, will face different challenges in file storage and file recovery.

We expect that with the continued increase in interest from academic and other institutions in creating reliable and accessible digital humanities repositories, that some of these many projects, if not most of them, will encounter a similar challenge in their development to that encountered by the DRI project and described here. Testing the scalability and reliability of trusted digital repositories is a task that will present to them, as it did to us, problems to be overcome. We believe that the use of the Hydrate tool by these projects could provide a potential and extremely workable solution for this problem, allowing as it does for the datasets generated to be specifically chosen by the user to suit each situation, and also allowing for the addition of other metadata standards as well as the existing choices of Dublin Core or EAD. Moreover, we expect that with the further development of Hydra as a reliable digital object repository management system, more and more digital humanities projects will be adopting it as a solution, thus making the facility with which Hydrate integrates with Hydra a particularly attractive option.

Currently, it is envisaged that Hydrate will be made available to the digital repository community as a whole under a Creative Commons rights license. This is in order to facilitate other, similar repository projects worldwide; particularly those developed in Hydra, which Hydrate was designed to complement as an added-value service.

---

[4] O'Carroll, A. and Webb, S. (2012), Digital archiving in Ireland: national survey of the humanities and social sciences. National University of Ireland Maynooth.