# The process of building a national trusted digital repository: solving the federation problem

Sharon Webb
Digital Repository of Ireland (DRI
National University of Ireland, Maynooth
Co. Kildare, Ireland
sharon.webb@nium.ie

Aileen O'Carroll
Digital Repository of Ireland (DRI
National University of Ireland, Maynooth
Co. Kildare, Ireland
aileen.ocarroll@nuim.ie

## ABSTRACT

The Digital Repository of Ireland (DRI) is building an interactive national trusted digital repository for contemporary and historical, social and cultural data held by Irish institutions. It will provide a central Internet access point and interactive multimedia tools, for use by the public, students and scholars and inform national policy for digital preservation and access. In 2011/2012 DRI conducted a requirements analysis of stakeholder needs [1]. This paper focuses on how aspects of this requirements analysis are translated into technical and policy solutions. We address how the project consortium, comprising six academic institutions, integrates with existing partner repositories and how the Digital Repository of Ireland tackles issues of repository federation in terms of storage, deposit and the legal frameworks associated with these activities.

## General Terms

Infrastructure, communities, preservation strategies and workflows, case studies and best practice

## Keywords

Requirements, policy, storage, deposit, user roles, use case, legal frameworks.

## 1. INTRODUCTION

The Hope consortium, tasked with building a federated repository of social history archives, detail a number of suggested benefits to adopting a federated model. They argue that users are less likely to turn to local catalogues to find content and that federation is more responsive to user needs. Clustering of content increases connections and links between content located in different collections, both at the national and international level, which enhances the contextual information about the digital object. Federation drives the adoption of open source solutions and shared standards which both increases the sustainability of the technical systems and the discoverability and quality of digital objects [2]. However, federation is not without its challenges. In this paper we outline how the Digital Repository of Ireland (DRI) has responded to the challenge of federation by discussing federation at the levels of storage, access management, and organisational structure.

The Digital Repository of Ireland (DRI) is building 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; and is seeking to inform national policy for digital preservation and access. The DRI research consortium comprises six academic partners: Royal Irish Academy, National University of Ireland Maynooth, Trinity College Dublin, Dublin Institute of Technology, National University of Ireland Galway, and the National College of Art and Design. DRI is a four-year exchequer funded project (funded by the Higher Education Authority PRTLI Cycle 5), and is collaborating with Irish cultural and social institutions such as the National Library of Ireland and the Irish national broadcaster RTÉ.

In parallel with a comprehensive requirements specification phase we have developed a lean repository prototype, and published a national report [3] with the findings from our nationwide programme of stakeholder interviews to determine the digital preservation and access practices in cultural institutions, libraries, higher education institutions and funding agencies. We are working to raise awareness of the need and benefits of digital preservation and open access, while respecting and acknowledging ownership, copyright, intellectual property rights, privacy and confidentiality.

In 2013 DRI carried out a mapping exercise, examining the range of institutions tasked with caring for digital content [4]. It is possible to classify three different architectural approaches to caring for digital content:

1. Single-site repositories, in which the technical and organisational function are located in one place (excluding off-site backup). The single-site approach is often adopted by national infrastructures.
2. In 2007–2009 a number of metadata aggregators were established. This approach brings together (aggregates) the metadata of a number of single-site repositories, thus increasing user awareness of content held in various repositories.
3. Since 2009 there has been a demonstrated shift towards the establishment of multi-site repositories, in which the technical infrastructure is federated across a number of repository sites. The Internet Archive and Dataverse were early adopters of such a multi-site approach.

The first challenge faced by DRI was how did we fit into this repository landscape and in particular how would we interpret our commitment to federation.

iPres 2014 conference proceedings will be made available under a Creative Commons license.
With the exception of any logos, emblems, trademarks or other nominated third-party images/text, this work is available for re-use under a Creative Commons Attribution 3.0 unported license. Authorship of this work must be attributed. View a copy of this licence.

## 2. FEDERATION

Federation may refer to an organization or group within which smaller divisions have some degree of internal autonomy. It can occur at a number of layers in the software and hardware infrastructure as well as at an organisational level. The OAIS reference model describes Federated Archives as "a group of Archives that has agreed to provide access to their holdings via one or more common finding aids"[5]. In this context they define a Global Community as "an extended Consumer community, in the context of Federated Archives, that accesses the holdings of several Archives via one or more common Finding Aids" [5]. Different types of federation are evident among those caring for digital content [4]. For example, Europeana is an example of a system in which DIPs (dissemination information package) containing the finding aids from each OAIS are ingested into the Common Catalog [5], it federates at the metadata layers and requires members to adhere to its standards. In contrast, the Institute for Qualitative Social Science (IQSS) model of federation is that it delegates the access controls to it's users, the systems are primarily located at IQSS's data centres [6], The IQSS provides the tools and infrastructure for contributors of data and meta-data and lets the 'user' decide on its own level of autonomy and trust. Of the three types of federation outlined in the OAIS reference model, DRI is closest to a Global Site structure, that is

> Global access is accomplished by the export of a standard-format Associated Description to a global site. The global site independently manages a set of descriptors from many Archives and has finding aids to locate which Archive owns a collection of interest. The Consumer is given a combined view of the holdings of multiple sites, which is maintained centrally. To view details of the documents, the user must access the site that contains the actual document. This is made easier when sites and clients support a standard set of protocols. [5]

In seeking to future-proof the DRI infrastructure, and in line with emerging trends, we have adopted a federated architectural approach for the DRI. In addition to the benefits outlined by the Hope project above, this also enables us to partner with existing and future digital archives, which we view as essential for a richer user experience, and to truly achieve our national mandate.

## 3. STORAGE

Federating at a storage level brings with it obvious advantages. DRI is building a trusted digital repository; it is a requirement of this trusted system to have high level availability (that is, with limited, controlled, downtime) and redundancy (duplicate copies of data available). Therefore, we are federating at redundancy and backup level. This approach fulfills a number of important business requirements, namely that the system is robust and reliable. Federated storage means that each federated member holds a copy of the repository, so if one goes down there are additional copies of the data and metadata available. This set up ensures that users have sustained access to content. This is a necessary feature from the user's perspective, as a reliable service garners trust - it also helps to build a user base that has confidence in the service provided. However, this is federation in a shallow sense and is not the focus of this paper. Here we focus on at other levels of federation - the first of which is delegating responsibility to federated partners in terms of deposit and access.

## 4. DEPOSIT AND ACCESS MANAGEMENT

The access management of an infrastructure, repository or application server can often be centralised or distributed.[1] Access management depends on the level of federation of the system and the policies governing access. Access management can occur in a central manner where it is centrally controlled or it may be delegated back to the community. For example in the IQSS and Europeana infrastructures it would be up to the contributors to decide what can and cannot be accessed. The control is delegated (federated) to members of these organisations. This is the approach, informed by our requirements and policy interviews, that DRI is taking.

As discussed in our 2013 paper, "The process of building a national trusted digital repository: a user centric approach for requirements gathering and policy development" [1], our requirements analysis informed us that it was necessary to build ingest functionality to support single as well as bulk ingest. This activity gives stakeholders high levels of autonomy and control over the ingest (or deposit) process. Although DRI is federated at an organisational level, one approach could have been to allocate central resources to manage the ingest process on behalf of DRI partners. Instead we chose to build an automated process that distributed responsibility to the stakeholders. The driver of this model is to ensure effectiveness in the context of resource limitations. However, an additional benefit of this model is that it builds the DRI federation at an organisational level, since in order to deposit, depositors must also act as partners. This involves legal agreements, as well as training and skill sharing within and among the community of DRI partners.

Our online work-flow facilitates data ingestion to the repository remotely (via ingest tools) by authorised third parties, namely partners of the DRI project. For this requirement we have developed a process to authenticate individuals who wish to deposit data on behalf of their institution/archive/library, etc. and have identified a hierarchy of those "users" that may work on such ingestion processes. In order to create and populate collections in DRI, representatives from an institution (library, archive, museum, etc.) need to apply to DRI to become an Organisational Manager. Once signed up the Organisational Manager can assign different roles to staff (see below for legal frameworks).[2] Additional roles include Manager User and Edit User.

The Organisational Manager is a user who has full access rights to particular collections and who has signed the Organisational Manager Agreement (see below), as such they act on behalf of their particular "organization" (university, archive, research center, library). They may or may not be the depositor of content but they have permission within the system to create collections and grant Manager and Edit roles to preferred users. In most cases this will be a librarian and/or a professional archivist. An Organisational Manager can:

---

[1] Access management should however not be confused with authentication and identity management of users of a given system, these issues are not dealt with in this paper

[2] The Repository Administrator will grant Organisational Manager privileges following instruction by the DRI Director.

1. Create a new collection in which to deposit digital objects.
2. Assign Manager User (see definition below) roles to a registered user in DRI.

A challenge that we faced was that many large institutions, such as a university, often themselves had federated structures. Therefore, it is envisaged that there will be more than one Organisational Manager associated with these types of federated institutions. The role of the Organisational Manager is illustrated in the following use cases.

> **Use Case 1**: An Organisational Manager, the Head Librarian, wants four collections from the library (1798 Pamphlets, 20th Century Fanzines, 15th Manuscripts and Irish Soldier's Wills) ingested into the repository (DRI). The Head Librarian wants to assign the management of these collections to four members of staff who are individually knowledgeable of one area each. The Head Librarian assigns four members of staff as a Manager User, one for each collection/project.

> **Use Case 2**: The head of the Department of Sociology wants to use the repository (DRI) as their main repository for research data generated by their PhD students. The head of the department asks their administration staff to register to DRI and apply to become an Organisational Manager on behalf of the department. The Organisational Manager (i.e. the admin. staff) is the point of contact for all PhD students who want to deposit their research data into DRI. The Organisational Manager will create a new collection for each student and assign him or her a Manager User role.

The role of a Manager User therefore reflects the need to allocate or grant responsibility for the day-to-day management or maintenance of a collection. An Organisational Manager automatically inherits the functionality or capabilities of the Manager User and can chose to delegate or not. A Manager User is a user who has manage permission on a particular collection or collections. Although strictly speaking, this is a permission-based role, it can be thought of as a distinct user type. These user permissions should, however, be interpreted as applying only with respect to the specific collection or collections on which the user has manage permissions.

A Manager User is an authorised user who can ingest content into collections, which an Organisational Manager has assigned to them. A Manager User can manage a number of collections. They have permission to:

1. Set the metadata standard for the collection.
2. Edit the collection title.
3. Provide a description of the collection.
4. Upload funding and partner logos related to the collection.
5. Assign and remove Manager User roles.[3]
6. Assign and remove Edit User roles.
7. Set and edit access permissions.
8. Review a collection.
9. Publish a collection.
10. Review collection activity.
11. Create folders

Importantly, a Manager User must "review" a collection (e.g. access permissions, metadata, etc.,) before a collection is "published" and visible on the DRI repository. This step is both a quality review for the Manager User and a chance to ensure that access permissions are correct in cases where a Manager User is relying on an Edit User to upload content. The Manager User automatically has the same permissions as an Edit User (see definition below).

The role of the Manager User is illustrated in the following use case:

> **Use Case 3**: A librarian is assigned as a Manager User and given access to the "1798 Pamphlet" collection. They write a description of the collection to give contextual information to the project and upload their institutional logo. There are 10,000 digital objects in the collection, each of which consists of the digital asset (the image) and a metadata file (Dublin core in XML). The library has two interns to help ingest the collection into DRI, the Manager User assigns these interns the Edit User role.

Finally, an Edit User is an authorised user who can ingest content into collections they have access to. An Edit User has limited functionality/permissions but must also adhere to the deposit terms and conditions (see legal framework below).

They have permission to

1. Ingest digital objects (asset and metadata) into the repository. They can use the single ingest web form or the bulk ingest tool (currently a command line tool).
2. Edit object metadata
3. Delete unpublished objects
4. Set a collection from draft to "for review" by a manager user.

The role of the Edit User is illustrated in the following use case:

> **Use Case 4:** The library's summer intern is allocated the Edit User role by a Manager User to help ingest objects into a collection. The collection is publically accessible and contains no objects that are restricted or sensitive in nature. The Edit User uses the single ingest web form to upload objects into the repository and creates the metadata upon ingest.

DRI have developed the above user hierarchy to facilitate the various institutional constraints. It supports the distribution of work and effort when users deposit data into the DRI repository. Each user type described above can ingest into a collection for which they have access and ingest permissions. As such at any given point an Organisational Manager, a Manager User or an Edit User may be a depositor of a collection.[4] Therefore, it is important that each of these users confirm that they agree to the terms and conditions of the deposit agreement.

[3] This functionality allows the Manager User to delegate responsibilities to staff, however, we are currently reviewing whether the remove Manager User functionality should remain with the Manager User or rest solely with the Organisational Manager.

[4] A Depositor is an authorised user who can ingest objects into a collection. A Depositor may be a Organisational Manager, a Manager User or an Edit User. An Edit User cannot set access permissions to a collection or digital objects.

This user hierarchy supports the automated system that DRI have developed to ingest content from DRI partners. This automated system introduces a number of issues in terms of, "trust" - DRI partners trust DRI to hold, make available and preserve their content, while DRI must trust that depositors will adhere to the deposit agreements and in particular set the access controls on their content. Trust is introduced and based here on social and political relationships, which are then codified in a technical solution and a legal framework addressed in the next section.

## 5. LEGAL FRAMEWORKS

As noted above, at an organisational level, DRI is a consortium of six academic partners. Partners, in the main, not only contribute to the building of the repository at technical, policy and business levels, but also populate the repository with digital objects through demonstrator projects [7]. These demonstrator projects serve to test the repository as well as populate it with content. DRI is following the ISO 163163 (the ISO standard pertaining to Trusted Digital Repositories (TDR)) in the development of its policy framework. This standard mandates that deposit of data must take place within a specific legal framework of agreements between the repository and those who deposit -

> 3.5.1 The repository shall have and maintain appropriate contracts or deposit agreements for digital materials that it manages, preserves, and/or to which it provides access [5].
>
> The repository shall have contracts or deposit agreements which specify and transfer all necessary preservation rights, and those rights transferred shall be documented [5].

DRI faced two related challenges in developing the legal frameworks attached to deposit, access and re-use of data. Firstly, how to manage deposit licences in a federated structure and secondly, to what extent the system could be automated if paper trails or signed documentation was required.

The demonstrator projects allowed us to test the legal frameworks developed. Traditionally repositories take data from the depositor, ensure that a deposit agreement is signed and from there manage preparation and ingest of the data to the repository. There are two actors involved in this process; the depositor and the repository. Yet, as we have seen, DRI has an organisational structure that is distributed - that is, deposit in the main will not be managed by DRI personnel but instead by the depositing organisation. In many cases the depositor will not also be the owner of the data (e.g. an institution, such as a library, may be depositing data to DRI that is owned by a third party). However, the depositor will have permission from the original owner to re-use the content.

DRI is managing the distributed nature of deposit through an interconnecting network of legal agreements. Current DRI partners have, via the existing legal frameworks, the ability to assign staff to Organisational Manager roles. However, it is envisaged that DRI will expand to include new members, depositing new data. An *Organisational Manager Agreement* is an agreement between DRI and a DRI member organisation. The Organisational Agreement is attached to the Organisational Manager role and delegates responsibility for managing ingest to this user type. In contrast, the *Deposit Terms and Conditions* are attached to the collection being deposited within the archive. Either the Organisational Manager or, more likely someone they nominate, deposit the digital objects and thus have the responsibility of agreeing with the *Deposit Terms and Conditions* (discussed below).

In developing these agreements, and being mindful of the ISO 16363 standard for Trusted Digital Preservation, we encountered a number of issues that needed to be resolved. Firstly, what indeed constituted a "legitimate" deposit agreement? We noted that ISO 6363 required that "contracts and formal deposit agreements should be legitimate; that is, they need to be countersigned and current"[5] and that in most of the archives and repositories we surveyed, deposit agreements were indeed paper documents counter-signed by both parties. Instead we were proposing the use of a 'click-wrap' agreement, that is

> an agreement, formed entirely in an online environment such as the Internet, which sets forth the rights and obligations between parties. The term "click-wrap" is derived from the fact that such online agreements often require clicking with a mouse on an on-screen icon or button to signal a party's acceptance of the contract [19].

After legal consultation we were reassured that a 'click-wrap' license was as valid and legitimate as more traditional legal agreements, indeed 'legitimate' had no particular meaning in Irish contract law.

The second challenge we faced was, did we need the *Deposit Terms & Conditions* to explicitly state the access conditions, contact details and licenses attached to the deposited digital objects (as is traditionally the case) or could we transfer these responsibilities to the depositor? The Organisational Agreement outlines both organisational responsibilities and DRI responsibilities. Many of the issues covered by the Organisational Agreement are familiar to those utilised by single-site archives. From the organisational perspective there is a requirement that the digital objects deposited meet the repository documented standards (including but not exclusively those pertaining to licensing, metadata and formats), and that the repository is granted the right to make available the digital object and process them according to established data protection practices. In return, the repository undertakes to preserve the digital objects and maintain their long-term usability in accordance with the repository's preservation strategy. In addition, the agreement allows the Organisational Manager to authorise users to act as depositors, adding or modifying data within the system. The ISO 16363 requires appropriate contracts or deposit agreements. They suggest:

> An agreement should include, at a minimum, property rights, access rights, conditions for withdrawal, level of security, level of finding aids, SIP definitions, time, volume, and content of transfers [5].

DRI departs from traditional practice in that the *Organisational Agreement* states that the Organisational Manager will ensure that the appropriate access permissions are set per collection and/or object basis as applicable, that the appropriate re-use licence is set per collection and/or object basis as applicable e.g. CC-BY, etc., that any embargo dates (e.g. if the collection publication date should be delayed) are set on a collection, etc. The role of the *Deposit Terms and Condition* in this distributed system is not to record the conditions under which the repository may distribute data, rather it places responsibility on the depositor to apply these conditions themselves when depositing data. The ISO 16363 framework allows for responsibility to be placed on depositors, for example

> Agreements may place responsibilities on depositors, such as ensuring that Submission Information Packages (SIPs) conform to some pre-agreed standards, and may allow repositories to refuse SIPs that do not meet these standard [5].

In a repository which is federated at an organisational level, the depositor is delegated a much greater level of responsibility. This responsibility is captured in the various legal documents and agreements that DRI partners must agree and adhere to in order to participate in our federated system and organisation.

## 6. SECTIONS

The HOPE Project outlined many of the advantages of federation. In the most obvious way, federating technically at the storage layer facilitates robust and reliable back-up - this is reflected in DRI's approach to storage. This paper highlights other domains at which federation can occur. In particular DRI have developed workflows that provide a degree of internal autonomy to DRI partners - they are responsible for managing deposit of and access to their data. They have autonomous control of their data for all actions with the exception of hard delete (this is currently in discussion). Trust is embedded in contractual agreements and in the provision of appropriate training and skill development. To this end we have developed metadata user guidelines and fact sheets on formats, copyright, metadata and hosted a number of workshops and seminars. A key advantage of delegated responsibility is it drives sharing and interoperability. The delegation of control is only possible when accompanied by shared standards and protocols, however, these are not developed *by* the repository *for* depositors, rather they are created *by* the federation, *for* the federation. Our 2013 article on the process of requirement gathering and policy development concluded that "Building an infrastructure should not be considered a series of linear steps but rather a process of discussion and engagement." Most of the partner organisations have pre-existing repositories whose autonomy they wish to retain, yet they also need support for the task of long-term digital preservation and are cognisant of the benefits of building links between the collections they hold and collections in other partner institutions. The technical, organisational and legal infrastructure developed by DRI is responsive to the needs of our community - however it has the additional benefit of strengthening and supporting that community through the federated structures that encourage the development of shared infrastructure, policy and advocacy.

## 7. ACKNOWLEDGMENTS

We would like to thank Associate Professor Eoin O'Dell at Trinity College Dublin for his past and continued assistance on the various legal issues discussed here.

## 8. REFERENCES

[1] O'Carroll, A. and Webb, S. 2013. The process of building a national trusted digital repository: a user centric approach for requirements gathering and policy development. In *Proceedings of the 10th International Conference on Preservation of Digital Objects* (Biblioteca Nacional de Portugal, Lisboa). DOI - http://purl.pt/24107/1/

[2] HOPE (2012) Best Practices for Trusted Digital Content Repositories Best Practices for Trusted Digital Content Repositories http://www.peoplesheritage.eu/pdf/D2-4-Grant250549-HOPE-BestPracticesTrustedDigitalContentRepositories2-0.pdf Accessed 24th March 2014

[3] O'Carroll, A. and Webb, S. 2012. Digital O'Carroll, A. and Webb, S. (2012), Digital archiving in Ireland: national survey of the humanities and social sciences. National University of Ireland Maynooth. DOI: 10.3318/DRI.2012.1 available at http://dri.ie/digital-archiving-in-ireland-2012.pdf

[4] O'Carroll, A., Collins, S.,Gallagher, D.,Tang, J., & Webb, S. 2013. Caring for Digital Content, Mapping International Approaches. Maynooth: NUI Maynooth; Dublin: Trinity College Dublin; Dublin: Royal Irish Academy. DOI: 10.3318/DRI.2013.1 available at http://dri.ie/caring-for-digital-content-2013.pdf

[5] CCDS (2012) *MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS) Magenta Book, The Consultative Committee on Data Systems*, p1-11 http://public.ccsds.org/publications/archive/650x0m2.pdf Accessed 20th March 2014

[6] Institute for Qualitative Social Science, http://www.iq.harvard.edu/ (accessed 25 March 2014).

[7] National College of Art and Design: *Kilkenny Design Workshops*, NUI Galway: *A Visual-Audio Demonstration of Irish Language and Cultural Heritage*, NUI Maynooth: 1916 *Letters*, NUI Maynooth: *Irish Lifetimes;* TCD: *Harry Clarke Studios Archive*

[8] Buono, F.M. & Friedman, J.A (1999) "Maximizing the Enforceability of Click-Wrap Agreements", *Journal of Technology, Law and Policy*, 4:3 http://jtlp.org/vol4/issue3/friedman.html Accessed 19th March 2014