The process of building of a national trusted digital repository: A user centric approach for requirements gathering and policy development

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

Sharon Webb
Digital Repository of Ireland (DRI)
National University of Ireland, Maynooth
Co. Kildare – Ireland
Sharon.webb@nuim.ie

ABSTRACT
In this paper we describe a process of consultation and data gathering with key stakeholders conduced by the Digital Repository of Ireland (DRI) in 2011/2013. This paper will examine the contributions of the interview process to policy development and requirements gathering with a particular focus on access, reuse and community engagement.

Keywords
Case Studies and Best Practices: Processes, Metadata, Systems, Infrastructure, Community, Policy, Requirements.

1. INTRODUCTION
The Digital Repository of Ireland 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. It 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. A key task is to link together and preserve the rich data held by Irish institutions, provide a central internet access point and interactive multimedia tools. Enabling access and reuse to research data is a central challenge. This article outlines how the process of qualitative interviews conducted by DRI allowed us to develop a complex understanding of the barriers which might limit the ability of data to be shared. An unexpected outcome of this process was that it facilitated community engagement. This assisted in developing the relations of trust which are so important for overcoming barriers to access and data sharing.

2. METHODOLOGY
In this section we outline the contribution of qualitative interviews to requirements gathering and policy development. We then briefly describe the interview process. Chituc (2012) argues that "research on requirements engineering in the context of LTDP is scarce More effort should be allocated to pursue research on requirements engineering targeting information systems ensuring long term preservation of digital data" [1]. From the project's inception DRI emphasised the need to carry out a thorough evaluation of the needs and requirements of its target audience. It sought to underpin the development of this national infrastructure by understanding the activities, task, goals and behaviours of its' users rather than building a solution to an unspecified and unknown problem. To achieve this and to fully understand the problem domain we utilised traditional software engineering techniques and incorporated the use of qualitative interviews in these activities.

Users of the Digital Repository of Ireland can be defined in two ways; firstly users are content holders (cultural institutions, social science archives and libraries) who may either be sharing their digital content directly with DRI or will be sharing their metadata. A second set of users are the researchers and general public who will be making use of the digital content. The boundaries between the two are not clear cut as in some cases, the content holders are also researchers who both archive and use content (for example, the Irish Qualitative Data Archive, the All Ireland Research Observatory, and An Foras Feasa). Additionally the project design includes a number of researcher-led demonstrator projects who are tasked to test the repository and to illustrate the power of the archive. The first round of interviews, which this paper is based on, addressed representatives from the first group of users, the content holders, and the demonstrator projects. Within the interview, the interviewees were asked to describe their needs as both content-holders, and as end-users.

Requirements engineering extracts, derives and specifies system behaviours, operations and functions and ensures the system is built upon and reflects authentic user requirements. DRI considers the problems related to data preservation, manipulation and dissemination associated with the humanities and social science data. This context shapes the DRI solution by ensuring the various user and stakeholder data requirements (e.g. access control) are met. As such one of the most important phases in requirements engineering and development is requirements elicitation, or information gathering. In order to inform the development of the system we need to listen to the target community of users and specify their requirements formally. A user-centric approach, that is listening and learning from the target end-user, is an essential feature of any requirements methodology.

Understanding the problem domain is an essential activity within software engineering and is part of the software life-cycle known as requirements engineering (RE). RE is a subject area in its own right but may also be described as a sub-discipline of software engineering. The RE process informs the development of the system that will be and emphasizes the need for project goals and objectives that are informed by the target audience, or indeed the community of

users. The aim of the RE process is to explicitly state the required features and characteristics of the system from the users' point of view. It is composed of a number of phases, of which elicitation, analysis, specification, verification and evaluation are of significant importance. Our stakeholder interviews were part of the requirements elicitation or information gathering phase. While DRI's mission and vision statement have clear goals and objectives, from which we extracted core business, as well as functional requirements, these interviews highlighted a number of challenges to specifying a clear, generic set of user requirements for DRI.

Policy development is similarly central to the process of becoming a Trustworthy Digital Repository (TDR). The RLG-OCLC Report on 'Trusted digital repositories: Attributes and responsibilities' [2] explicitly identifies policy development as a central function of TDRs. In order to meet these policy obligations, DRI has adopted an eight step policy development cycle;

1. Issue identification
2. Policy analysis
3. Policy instrument development
4. Consultation (which permeates the entire process)
5. Coordination
6. Decision
7. Implementation
8. Evaluation

Part of the policy analysis process required DRI to review national and international practice. Conducting a review, through qualitative interviews with key participants allowed us to not only review policies in existence but to map emerging challenges and areas of concern. This allowed us to develop a richer understanding of policy issues, than if we had limited our review to the collation of published outputs and documentation.

DRI conducted 40 requirement interviews with key stakeholders from December 2011 to August 2012[1]. The representatives were drawn from the following spheres; digital repositories, university libraries, cultural institutions, social researchers, media organisations, public libraries and government content holders. The interviews were semi-structured. Our aim was to establish how users/stakeholders currently support their digital resources/objects and how they develop and maintain their data archives/repositories. The key approach is to use open ended questions (e.g. can you tell me about, can you describe, etc.), following the flow of the interviewee, and only directing, if the issues that need to be discussed do not emerge naturally in course of the conversation.

A topic guide (see appendix) was prepared which addressed the resource/archive in terms of its current data life-cycle.

Pre-ingest Stage: The activities surrounding the data before it is prepared for archiving.

Ingest Stage: Preparation and deposit of data into archive.

Preservation Stage: Fulfilling archive's responsibility to preserve data.

Dissemination Stage: Fulfilling an archive's responsibility to enable reuse of data.

Future development within a federated repository.

Issues addressed included software or computer systems in use, whether it was astatic or living archive, whether there were multilingual data, metadata and database formats, future proofing, data security and user tools. Policy issues relating to ownership, copyright, IP issues, and data sensitivity were also addressed. Where permission was granted, the interviews were recorded and the majority were transcribed. It is intended to archive the interviews so that they will form part of the DRI collection.

3. TRUSTED DIGITAL REPOSITORY - CHALLENGES TO POLICY

Two key discipline-specific policy themes emerged in the course of the interview discussions on facilitating open access. For the humanities and cultural heritage organisations copyright was a key concern, particularly in the face of shifting national and European legal frameworks. Social science organisations required policy frameworks which address data protection needs and the obligation to meet ethical research standards.

Copyright issues were of concern to many. While, there was an eagerness to enable sharing and re-use of digital data, some collections had copyright or ethical restrictions that limited these possibilities. Libraries were affected by the impact of copyright legalisation which placed access restrictions on books, journals and collections they held. Some institutions exercised copyright to generate revenue. Others exercised their copyright in order to limit unwanted re-use of their data. For example, one institution cited the re-use of a photograph in their collection by a commercial entity, in a way that exposed the individuals in the photograph to ridicule. This type of misuse could be prevented by denying the right to re-use. However, this also required that the institution was both aware of the re-use and in a position to defend its copyright – circumstances that would not always be true.

Most social scientific data (and some donations to libraries and archives) had re-use restrictions placed on them which limited who would be able to access the data and required that the anonymity of the original interviewees be maintained. These limitations lessen over time; in 100 years all data can be shared. Our interviewees also expressed concern about long-term preservation of digital content which had time embargoes restricting access, in some cases as long as 30 to 100 years. The time and resources needed to ensure sustainable access to these objects, in order for them to become publicly available in the far future, had not been fully explored by any of our interviewees.

The review found a marked interest in increasing access to digital data, including the use by many institutions of social media to engage with the general public. However there were important tensions. Within the social sciences, where data are collected on the lives of contemporary individuals, a balance needs to be maintained between the rights of the

[1] Ethical approval was granted by the Ethical Committee at National University of Ireland, Maynooth and consent forms (which included consent for future archiving) were used in all interviews.

public to access publicly funded data and the rights of research participants to have their confidentiality protected. Copyright brought with it an additional set of tensions that both restricted the sharing of data and also protected the interests of individuals and institutions. While the copyright concerns attached to digital and physical objects are in many ways similar, digital data carries with it additional opportunities and challenges to make collections and objects widely available by sharing them on the internet, but there was a clear sense that once an object was released, it would then be extremely difficult, if not impossible to police how that object might be used. Given that we are living in an increasingly digital world, there is a need in Ireland for a national digital policy which capitalizes on the possibility attached to digital data and provides guidance on how to facilitate sharing and re-use of digital data. Additionally, internationally a significant trend towards sharing of publicly generated data is evident, and as new copyright and ethical frameworks are developed, barriers against sharing may be reduced. Since the completion of the initial phase of the research, DRI has contributed to the publication of a 'National Principles for Open Access Policy Statement" [3] which in terms of research data states:

Research data should be deposited whenever this is feasible, and linked to associated publications where this is appropriate:

- European and national data protection rules must be taken into account in relation to research data, as well as concerns regarding trade secrets, confidentiality or national security.
- At a minimum, metadata describing research data and its location and access rights should be deposited.

This is an important first step in developing a national infrastructure which facilitates open access and re-use of research data. As such the DRI has adopted an open metadata policy and will make available its metadata under appropriate broad-use licences. It has selected a number of metadata standards which it will recommend for use with textual and visual data and is reviewing metadata standards for other data types. Our decision to support a range of metadata standards requirements is drawn from a recognition, drawn from the interview process, that the various domains served by the DRI have differing experiences in terms of metadata use. Depositors will be advised to use the metadata standard appropriate to their discipline. Our choice of standards reflects common practice in Ireland and internationally[2] Many users are involved with Europeana, therefore an additional policy became evident - the need for interoperability with Europeana. As such EDM will be supported by DRI.

4. BUILDING INFRASTRUCTURE - CHALLENGES TO REQUIREMENTS

A number of issues emerged in the course of the interviews which impacted DRI's requirements specifications; the requirement to retain a local stakeholder identity, clear identification of copyright, variable access controls and the development of user tools, for example time-lines and mapping interfaces. An unexpected outcome was the realisation that as an emerging field many stakeholders did not have a clear understanding of the requirements. The stakeholder consultation was as much a process of discussion as it was of gathering information.

The purpose of the stakeholder interviews was to establish, and learn from, the current activities of the community (relative to the data life-cycle stages mentioned previously) and from this to extract core user, as well as interface, storage and system requirements. Their aim was (and is) to ensure that the system is based on, and supports, *authentic* user requirements. However, through the interview process it became apparent that while we could identify some generic features, there were conflicts and tensions between particular requirements surrounding access, re-use and storage. This is related to the fact that DRI's designated community is quite diverse, both in scope and scale. More worryingly, but perhaps unsurprisingly, many were unclear or unsure of how DRI fitted with their current activities. This created further challenges in extracting requirements (a common problem related to requirements engineering - the customer or user not knowing what they want).

DRI's number one, core, business requirement is that it must be a trusted digital repository (TDR):

> REQ-1 A Trusted Digital Repository.
> The system shall be a trusted digital repository.
> 1.1 It shall supply provide 'reliable, long-term access to managed digital resources to its designated community, now and in the future'. (RLG-OCLC Report). (REQ-34)
> 1.2 It shall conform to the Data Seal of Approval guidelines or equivalent. (Defined by policy).
> 1.3 It shall be an access repository for the humanities and social sciences (HSS).
> 1.4 It shall have disaster recovery process in place. (REQ-57)

This requirement is mandated by the project description and is supported by policy guidelines and decisions. From this high-level, business requirement we specified numerous functional requirements to support the creation of a TDR. These include, but are not limited to, data integrity checks, disaster recovery mechanisms, export functionality and audit trail/reporting. Alongside this it must be an access repository for the humanities and social science data it stores, harvests and aggregates. Our access requirements, that is, how a user or actor can retrieve or view data, state that access to digital objects must be managed through authentication and authorization mechanisms. While we advocate open data and open access it was evident from our interviews that some stakeholders, beyond those with concerns over sensitive data or legislatively imposed embargoes, wanted to maintain some control over data access by particular users. In terms of requirements this solidified the need to implement role based access to content. Conversations about access also raised important questions and concerns over brand identity. Individual institutions expressed anxieties about becoming detached from their own collections within a system such as DRI. This revealed to us an essential user and interface requirement, namely, that of displaying the identity of hosting or contributing institutions or depositors to users

[2] They are Dublin Core, Modified Dublin Core, MARCXML, EAD, MODS and METS

when content was accessed or searched. Alongside this, our stakeholder interviews revealed important issues surrounding data re-use. A key concern was how best to support data aggregation and curation across different collections from different sources without creating copyright and licensing conflicts. Our requirements ensure that copyright statements are displayed to all end users and the systems maps copyright to all digital objects. Access rules foreground all our requirements and specify what a user can and cannot do within the system and in terms of data use and reuse.

5. CONCLUSION

Although the interview process was designed to gather requirements and map and develop policy, it quickly became evident that this was a process of joint discussion between DRI and the stakeholders; interviewees introduced us to the specificity of the issues facing them and we were able to alert interviewees to issues not previously considered. Some features identified are not traditionally seen as part of the remit of a TDR; these tended to be at the level of end-user needs rather than preservation needs (eg. smart phone/tablet use, end-user tools (visualisations, time/maps, user curated collections, crowdsourcing, etc.)). However, the importance of this review process is not necessarily in terms of innovation in terms of data management planning but in creating user-buy in and developing a closer connection between DRI and its user community. In times of decreasing resources and financial pressures (which was a common concern among the community), which creates competition for scarce resources, an approach which develops for community rather than with a community is unlikely to be successful. An "if you built it, they will come" approach is not feasible.

The interviews also highlighted that DRI is unlikely to succeed if seen only as a technical infrastructure. It is a socio-technical system in which the additional roles of training, skill sharing, and national policy development are also central to its mission. Digital archiving was a relatively new field to many; the interviews allowed for mutual learning and fulfilled an unexpected community engagement function. The 'bottom up' approach ensures that DRI will develop in response to stakeholder needs. Policy development continues as an iterative process as both a National Stakeholder Advisory Group and International Stakeholder Advisory Group have been established. Additional stakeholders continue to be interviewed on a rolling basis.

While the interview process has fruitfully contributed to policy development and requirements specification it also alerted us to the necessity for DRI to engage in training and development in order to ensure continued stakeholder engagement with the infrastructure. Building an infrastructure should not be considered a series of linear steps but rather a process of discussion and engagement.

6. REFERENCES

[1] Chituc, C.M. (2012) Requirements Engineering Research and Long Term Digital Preservation Open Research Challenges Workshop, Toronto.

[2] http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf (assessed 24th July 2013).

[3] National Principles for Open Access Policy Statement. http://www.dri.ie/sites/default/files/files/National%20Principles%20on%20Open%20Access%20Policy (assessed 24th July 2013).

7. APPENDIX: TOPIC GUIDE

Stage in Archive Life-cycle	Key Topics	Questions
Pre-ingest	Digital objects/resources? Quantity; data formats (txt, doc), processes of digitisation (crowdsourcing?) Computer or software systems in use. User-interfaces (bespoke, particular product?) Static or living archive? Bi-lingual data?	Can you tell me about your resource/archive/repository? Can you describe your data/content? Is all your data digitised? Can you describe the digitising process? Can you describe the current system you use for your data collection? How do you envisage your resource developing in the future?
	Data Quality Assessment/ Quality Control Process (in terms of data formats and data content)	How do you assess data/content quality?
Ingest	Nature of data (specific concerns, sensitive? rarity, commercial issues). Access issues/policy.	In terms of archiving or storing your data, are there any particular concerns or considerations? How did you address them?
	Ownership/ copyright IP	Who owns the data? Are there copyright issues? Do you have licensing agreements? Are there any IP issues?
	Collection priorities.	How do you source the data? Do you have specific priorities?
	Catalogue Ontology/ Thesaurus	Have you developed a catalogue? If so, can you describe it?
	Metadata formats? Database formats? Linked Data? Open Data/	What metadata standards do you use? Would you know what the database system you are using is? (MySQL, Excel, XML etc.)?
Preservation	Future-proofing - data formats/longevity of data.	Can you describe your preservation process, if any?
	Data security (physical threats, virtual threats) /Redundancy	Where is the data physically stored? What security systems do you have in place if any?
Dissemination/Data Re-use	User Experience /expectations (Actors e.g. students, researchers etc.).	Can you describe who uses your data? How do you see users in the future?
	What tools etc. do users currently use? (bespoke or not)	Do you provide any tools to enable the user to interact with the data?