Suggested topic: IR integration with research information management systems

Blurring the boundaries between an institutional repository and a research information registry: where's the join?

Introduction

Key motivations for provision of an institutional repository (IR) for research outputs within a higher education institution (HEI) are storage, retention, dissemination and preservation of digital research materials. Increasingly IRs are being considered as tools for research management as part of pan-institutional systems. This might include statutory reporting such as that required for the forthcoming UK REF (Research Excellence Framework). Such functionality generally requires integration with other management systems within the HEI. It is common to find that each research management system has been selected to serve a specific need within an organisational department, any broader aim being out of scope. As a result, data is held in many silos, is duplicated and can even be 'locked in' to those systems. This results in problems with data sharing, as well as lacks of efficiency and consistency. Some institutions are addressing this problem by considering CRISs (Current Research Information Systems) or business intelligence systems. The need for easy deposit in the institutional repository at the University of Oxford has prompted the development of a registry and tools to support research information management. Many of the motivations behind the repository are common with those for research information management. Not only do the two areas of focus have many common aims, but there is considerable overlap of design, data, services, and stakeholder requirements. This overlap means that the boundaries between the repository and the resulting tools being implemented for publicly available research activity data are blurred. By considering these two areas together with other related digital repository services, new opportunities and efficiencies can be revealed to the benefit of all stakeholders.

The motivations and structure of the IR

In 2006 the then Oxford University Library Services (now Bodleian Libraries) took steps to set up an institutional repository for the preservation of the University's research outputs. The repository needs at Oxford resulted in an architecture comprising a suite of federated repositories that could hold not only research output content, but various digital collections, already existing and not yet created. The aim was to build an underlying repository system that would serve all the digital collections and provide robust preservation functionality, resulting in the Oxford DAMS (Digital Asset Management System). The DAMS is a layered and modular structure that allows for separation of storage and digital object management, and for services and applications to be built on top as required. The system had to be able to cope not only with repository content, but vast collections such as the Google books content, heritage collections such as John Johnson ephemera and private 'dark' collections such as those being stored as part of the BEAM work. The mantra during the building of the system has been 'keep it simple' the aim being to create a robust yet easily maintained system that offers flexibility for use and further extension. It has been designed so that it can be regenerated in the case of failure, and so that services and parts can be removed and replaced without placing the whole DAMS in jeopardy.

The motivations behind the initial setup of the institutional repository named ORA (Oxford University Research Archive), were common to many similar repositories: the preservation, storage of and access to Oxford's research output. The system offers additional opportunities for the retention, reporting and management of research output, and to make Oxford research more visible and accessible. External factors that ORA helps address include fulfilling the requirements of and reporting to funding agencies and others.

By standing back from the IR as a whole and viewing it as a sum of many parts it can be seen as a machine where each component has a focused critical function. Each component has been selected to meet a specific requirement, with the criteria that it should be open source, in widespread use (in order to prove its suitability and to have built up a robust support community) and scaleable. In this way the search and indexing is currently provided by Apache SOLR, the RDF storage and query database is 4Store[1], and PairTree[2] is used for digital object storage.

[1] http://4store.org/about

The repository has been constructed using semantic web technologies including use of RDF triples and linked data. It uses the BagIt[3] concept and resource graphs. Vocabularies and ontologies are used where appropriate (eg SIOC[4] and LCSH as linked data). Each data element has a URI which is resolvable. The underlying precept is that the system can operate or interact with other systems and that data are re-usable. The system is metadata agnostic. For the most part, items in ORA comprise bibliographic text based materials such as research publications: articles, conference papers and so on. The core metadata is MODS, but items also have DC and MARCXML metadata, in addition to RDF. Other metadata schemas can be used as required.

In common with findings at other institutions, many researchers at Oxford do not have the time, nor always see the need to deposit their work in ORA. To make deposit quicker and simpler it was felt beneficial to obtain data that described research publications from appropriate existing internal sources. This would enable some metadata fields to be automatically filled. However, this was to prove more difficult than had been anticipated. Despite the problems, it was decided to pursue this idea further and a registry to store entity data for ORA was planned. Data were to be stored as separate entities (eg person details including different forms of the name, organisational units, the funding agency funding the research etc). In doing this, the potential for wider use of the registry and the data it contained began to emerge.

Overlap with research information management

It became clear that the entity data being gathered for the IR could be viewed in much broader terms as data describing research being undertaken at Oxford. This we called 'research activity data.' Funding was awarded by the JISC to create the registry as part of the BRII (Building the Research Information Infrastructure) project which ran Oct 2008 – March 2010. Data were harvested from existing sources, stored and processed in the registry and then made available for re-use within ORA. The potential uses for this data could be envisaged far beyond that of populating the repository.

ORA aimed from the outset to gather the richest metadata possible to provide for reporting and searching using different criteria. Many of these data are held in other existing locations. Some were already held in ORA, could be extracted from existing metadata and then re-used to save re-keying (eg author name). The entities forming the separate elements of metadata describing publications could be extracted and re-combined, enhanced and additional descriptors added. The types of entities being gathered include:

- People (names)
- Organisational unit (eg faculty, department, college)
- Project (title and other details)
- Funding agency (ie the agency/ies that funded the research)

They could also be used within ontologies and related to other entities. Such data can be used to answer questions such as 'who in Oxford is doing research on a specific topic?' That might sound a simple query, but in reality, a comprehensive answer was not always easy to find. In this way the repository, or more correctly, the ability to query the registry within the IR, becomes a fundamental component in the university's research management toolkit. Some metadata is repository and therefore bibliographically focused (such as publisher name) but much of it is highly relevant to the needs of research managers and administrators. When investigating the drivers and motivations of the institution for gathering research activity data during the extensive BRII stakeholder analysis[5], it was discovered that there is much in common with those underpinning the IR. They include the need for accurate data, including information about funders, and to facilitate research reporting (both internal and external) and dissemination.

In the same way, services being developed for the repository are of critical importance to the storage, management and use of research activity data. Some services are designed to support the trustworthiness and longevity of the data.

---

[2] http://www.cdlib.org/inside/diglib/pairtree/pairtreespec.html

[3] http://oxfordrepo.blogspot.com/2009/02/pushing-bagit-manifest-concept-little.html

[4] http://sioc-project.org/

[5] Available at http://brii.ouls.ox.ac.uk/

- Preservation: the ability to rebuild the system without loss of damage to data and the continuing ability to access and read data. This facility is provided by the DAMS
- Provenance: an indication of both the source of the data (ie from where it was harvested) and its known validity. For the metadata qualifying terms we have currently: validFrom; validUntil; validAt
- Relevant standards are adopted where appropriate including metadata schema and controlled vocabularies and data exchange standards such as OAI-ORE

The types of research activity data being harvested and aggregated for use in the repository are relevant for research information management and business intelligence. Data are harvested from sources such as departmental websites and databases. This is achieved with minimal impact on the data source. Because data are mirrored in the registry, data ownership is retained by the original source. The original data are dispersed across many diverse and disparate data stores. Drawing them into a single location adds value even before any more processing takes place.

Common motivations between those involved in promoting the repository and those involved in research management include the widest possible dissemination of Oxford research, and easy discovery and access to information about that research.

Data aggregation revealing opportunities

Aggregating previously scattered research activity data presents new opportunities for identifying connections between data entities. Ontologies and taxonomies are employed to define and categorise the data, so that connections between researchers, grants, projects and publications can be forged. This is not possible when information is held only in process-led silos. The information can then be used for trend spotting, business information, increasing efficiency and for knowledge transfer. Collaborations and the location of those collaborators can be ascertained using information drawn from publication (ie co-authors), project (eg co-investigators) and other data. Rather than creating a crude list of people and projects they have been associated with, this is done using a mixture of heuristics and co-referencing techniques being developed at the University of Southampton[6].

The services built as part of the registry and IR are extensible and adaptable for other collections within the Oxford federated repositories. For example, the provenance capability is being adapted and employed for digital objects in the Mellon funded 'Cultures of Knowledge' and the 'Medieval Libraries of Great Britain' projects which are describing digital objects of 18th century manuscripts and medieval catalogues. They will also be used as part of the library's repository for research output data, DataBank, which is being developed to support the JISC funded ADMIRAL project.

One of the key functions of the data entity registry is that the data it contains can be easily re-used. To this end it has been provided as linked data with resolvable URIs and is machine and human readable. It is designed so that data extraction is simple using RSS/Atom feeds and other simple solutions. Also, APIs are provided for easy and customised re-use. It is this that opens up the possibilities for others to build web services using the data. One example of additional use of the data, is that of the Medical Sciences Division (MSD) graduate opportunities website. This website uses data harvested into the registry from MSD departments which is then drawn back to be re-combined and re-used in new ways.

Where blurred edges obscure clear boundaries

Whilst the repository structure and metadata are being used and re-purposed for services and collections within library services, there are few problems of ownership and responsibility. However, once the remit goes beyond library mainstream services and starts to impact on central research information management services, there can be difficulties in assigning responsibilities. Not only that, libraries that are hard-pressed to fund the collections and services they already provide are unwilling to shoulder the cost of what is in effect a central administration service. The problem lies in the need for the service being targeted at administrators and management but the data management and semantic web expertise lying within libraries. It could be argued that many

[6] Glaser et al http://eprints.ecs.soton.ac.uk/17587/

library services fall into this category, but the difference lies in that it is not publication or literary collections that are being dealt with.

These questions can only be resolved by negotiation and by each stakeholder group having a clear understanding of the benefits of the system. The cost of not providing the service should also be taken into account. Where uses overlap to such a large extent, the aims of the system should be clarified from different perspectives. The people involved approach the services with different, often conflicting, perspectives. Policy and support aspects of the research information infrastructure registry are currently under discussion at Oxford to try to ascertain where responsibilities lie. This problem is likely to become further complicated if additional research management information is added to the registry. This might include financial and other sensitive information that is only visible to restricted groups.

As departmental systems (ie those which are or could be harvested and their data included the registry) are replaced and developed, more seamless data sharing could be employed, enhancing the overlap between services. By sharing and re-using data all parties can benefit: administrators by increasing efficiencies using pre-existing data and reducing re-keying, by maintaining a single canonical source of the data, and by easy discovery of information; management by interrogating the data and provision of reports for management information purposes; and by researchers themselves by not having to provide the same information many times.

Conclusions

The repository and the research information registry are one and the same in many respects, and in terms of data gathering, both feed and are fed by each other. The structure and content of the entity registry results in data sharing and system overlap rather than integration of services. When considering such data sharing and use of services for multiple purposes it is difficult to define where one service stops and the next starts. Integration should be assigned to the processes and policies of the services rather than at a technical level.