Welcome to C&I issue 187, the theme of this issue is metadata work in Institutional Repositories (IR). Feedback from the CIG conference that took place last year in Swansea indicated that people were interested in learning more about IR, and as it is an area that more and more cataloguers are becoming involved with at their institutions we felt that it would be timely to address the topic in C&I. We have five papers representing a range of metadata work, including a follow up article by Helen Williams to one published in 2012 which addressed the transformation of a bibliographic services team (at LSE) from copy cataloguers to metadata creators. A reflection from five years on details further transformations. Lizz Jennings looks at the development of a Research Data Archive at the University of Bath, and explains how metadata is at the heart of the solution. Louise Harrington and Tracey Andrews detail how ORCA (Online Research @ Cardiff) is the primary tool for the REF (Research Excellence Framework) at Cardiff University, and is embedded in researcher practice. Whilst Jason Partridge discusses the Oxford University Research Archive (ORA), in place since 2007, and the wide-ranging skill set that team members have developed.

Susan Miles and Katrina Clifford take us through the challenges and solutions of representing ‘practice-based’ items (such as live performances, sound recordings and exhibitions) in an institutional repository at Kingston University. Overall these articles provide a fascinating insight into repository metadata work which we hope will provide food for thought for our readers who are either engaged in similar work themselves, or for whom these tasks lie outside their own realm of occupation.

Karen F. Pierce: PierceKF@cardiff.ac.uk
Deborah Lee: Deborah.Lee@courtauld.ac.uk
Back in 2012 I wrote an article for C&I entitled ‘The past is a foreign country: transforming a bibliographic services team from copy cataloguers to metadata creators’. In it I summarized how our traditional Bibliographic Services team cataloguers became what one kind blogger described as ‘some kind of indispensable repository metadata ninja unit’, while still maintaining their usual cataloguing work. The article:

- outlined our background to the decision, and the preparation and practicalities involved in redefining our workflows;
- reflected on the successes and learning points of managing change and updating skillsets; and
- considered how we could remain at the heart of the Library’s mission to support the University.

While 2012 is not entirely a foreign country to our 2017 workflows, there have been changes in the last 5 years, as you would expect.

Firstly, a couple of structural changes mean that instead of a combined Bibliographic Services team, who shared acquisitions and cataloguing work, we now have separate teams for Acquisitions and Metadata Services. The Metadata Services team currently consists of the Manager, 2 Assistant Librarians (one of whom is part-time), 2 Senior Library Assistants, and 3 Library Assistants. From this group of 8, half are involved in working on our LSE Research Online (LSERO) repository. One of our Senior Library Assistants oversees and manages repository work in Metadata Services, and our 3 Library Assistants combine work on LSE Research Online with cataloguing work. This is in contrast to 6 Library Assistants working on LSERO in Bibliographic Services days, but compensated for by our team no longer carrying out acquisitions work.

Secondly we have seen the workload itself changing. Before we took on repository work our number crunching suggested that the team could spend 21 hours a week creating repository metadata without seeing a dip in our current print cataloguing output (these hours were instead freed up from historic gift cataloguing). At the end of our first year we found our average weekly time spent on LSERO was close to this at 24 hours, but with peaks of 42 hours a week, and lows of 16 hours a week, flexibility had been the key to our success.

In 2013 we used Eprints to manage our REF submission and so the year was dominated by adding and checking, and double checking, metadata to support this, which we managed with dedicated support from a couple of members of our (at that stage, still larger) team who developed significant REF expertise. By 2014 we had the skill of juggling our cataloguing and LSERO workflows down to the fine art you’d hope for from a ‘ninja unit’, and our average weekly time on LSERO was still pretty consistent with an average of 23 hours a week being spent on LSERO metadata creation, shared across 3 Library Assistants, with the help of the Senior Library Assistant for repositories.

---

But as 2015 rolled into 2016 we began to feel as if we were regularly ‘up against it’ in Metadata Services, with all our time going on cataloguing or LSERO data creation, rather than any of the project-type work we were used to having on the go at the same time. The inbox system we use to manage incoming LSERO work seemed to be constantly overflowing and we were regularly asking our team to delay cataloguing for a few days so that we could have all hands on deck to clear repository backlogs. Scrutinising our statistics backed up my suspicion that while incoming cataloguing remained reasonably constant, we were seeing changes in the LSERO data. As Table 1 shows, while we were seeing a reduction in the number of citation-only entries being submitted to the repository, we were seeing a significant increase in the number of full-text entries being received. The metadata for these takes longer to create, both by virtue of dealing with the full text and due to the additional metadata requirements of recent years, such as funder and project details to comply with the RCUK policy on Open Access, date of acceptance, preferred Creative Commons attribution, etc., which meant that overall we were seeing a significant increase in the hours required to keep on top of this work.

Table 1: Hours per week for the whole team

<table>
<thead>
<tr>
<th>Year</th>
<th>Average hours spent creating full text records each week</th>
<th>Average hours spent creating metadata only records each week</th>
<th>Average total time spent creating LSERO metadata each week</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>13 hours</td>
<td>10 hours</td>
<td>23 hours</td>
</tr>
<tr>
<td>2015</td>
<td>21 hours</td>
<td>8 hours</td>
<td>29 hours</td>
</tr>
<tr>
<td>2016</td>
<td>31 hours</td>
<td>6 hours</td>
<td>37 hours</td>
</tr>
</tbody>
</table>

Our current expectation is that Library Assistants will spend approximately one third of their time on cataloguing, a third on LSERO metadata and final third on counter duties and team or Library wide activities. Counter duties and other Library meetings are fixed timetable slots, so those continued to happen, but LSERO and cataloguing time competed with each other and required a lot of flexibility and juggling in the team to respond to competing priorities, particularly on the weeks where the hours required were above the averages in Table 1. Furthermore, we were finding that although we didn’t have acquisitions work to do anymore, 3 bodies instead of 6 meant it was harder to juggle workloads and urgent requests. If 1 or 2 people are off in a team of 6, the work is still manageable, but when that’s out of 3 your team can be depleted very quickly and very unexpectedly, making the best laid plans hard to execute. While the LSERO work doesn’t pile up visibly in quite the same way as books on shelves, or boxes on the floor, we are aware of our role as service providers, both to the Library’s Research Support Services team who have overall responsibility for the LSERO service, and to our academics who submit their publications to the repository, so backlogs sit uncomfortably with us, and we were keen to find a more effective way of managing the process.

We knew that possible solutions could not involve reducing the incoming work, or increasing the people to do it, which doesn’t leave a lot of ‘wriggle room’, so instead we tried changing our service level agreements from a 3 working day turnaround to a 6 working day turnaround for most content, but with the promise of dealing with urgent requests in 24 hours. We also have incoming blog content to be archived each month, and have increased the turnaround time for this to one month, though we aim to have protected time each month to keep up with this as while the metadata is simpler, the volume of content is higher. Obviously this doesn’t reduce the amount of work that needs doing, but the longer turnaround time frame has given us a greater degree of flexibility, which in turn has reduced the need to regularly adjust workflows and get the whole team focused on LSERO at the expense of cataloguing. This makes for a less pressured feel to the workflows as well as making them easier to manage.
In 2016 we successfully bid for some Higher Education Innovation Funding (HEIF) to increase the visibility of research outputs by extending the archiving of PDF versions of LSE blog posts in LSERO. The metadata applied to this content when PDFs are archived in LSERO enhances their discovery on the web, and facilitates exploiting them as a research resource. It also supports the Library’s strategic objective to collect and preserve the intellectual output of the School, by ensuring that outputs such as blog posts, which fall outside more traditional publishing channels, are still preserved, and made discoverable.

The Library was already archiving selected blog outputs (chosen either by blog editors who send content for addition, or individuals who request inclusion of their post) but this was not a systematic or comprehensive approach. The project began with LSE’s top 10 public facing blogs, and the daunting prospect of systematically adding somewhere between 5000 and 11,000 records to LSERO, in a 3-and-a-half-month period, in order to retrospectively archive all content from those blogs. At the time of writing, towards the end of April, we are just ahead of schedule with 3700 records already added to LSERO and the hope of retrospectively archiving content from a select number of LSE’s other 62 blogs as well.

The HEIF blogs project is being supported by a team of 3 temps and will finish at the beginning of July so it has been important to consider how we can sustain this increased amount of blog archiving work alongside all our usual incoming work. A secondary part of the project has been working with the School’s IT team to automate the creation of PDFs, which has been successful, and to auto-populate as many of the metadata fields as possible, which we are still working on. If both of these are successful then we would anticipate a significant reduction in the staff time required to create archived versions of blog posts and to assign appropriate metadata, allowing us to embed comprehensive archiving of the top 10 public facing blogs into business as usual work.

This may sound optimistic, given the increased volume of work outlined earlier in the article, but the School, including a team in the Library, is currently implementing a Current Research Information System (CRIS), which we anticipate will significantly change existing repository workflows and reduce the amount of manual metadata entry required, thereby freeing up staff time to deal with other unique LSE content, such as the blogs. However, since the project is not yet complete it remains to be seen exactly what the impact of the CRIS will be, and whether our current repository workflows will look like a foreign country in a year’s time.
Background

Requirements for sharing research data have increased in recent years, partly in response to the open science agenda and partly as a means to make better use of data generated using public money. Research funders increased their expectations of researchers in relation to research data sharing, and the Engineering and Physical Sciences Research Council (EPSRC) was the first Research Council to put the onus on institutions to provide support for this.

In response to the EPSRC expectations, the University of Bath developed a Research Data Archive for researchers to archive and share their data. This was developed using the institutional repository software, EPrints. EPrints has a flexible approach to metadata, so developing a metadata schema for use in the Archive was a significant task.

Priorities

The University's Roadmap for EPSRC set out several requirements for metadata, such as:

- making data discoverable within the University and through external data discovery services;
- relating data to publications and other information stored in the research information system, Pure;
- using external metadata standards such as DataCite to represent the data and provide unique identifiers;
- providing scope to include documentation with the data; and
- automating data and metadata capture where possible.

In addition to the discovery requirements, we had to ensure that research data uploaded to the system had sufficient management metadata to provide for its future care.

Developing a metadata schema

We decided to align with the DataCite metadata schema as we intended to use their service to create Digital Object Identifiers (DOIs) for datasets, and any metadata would be included in their global search facility. We also worked with the Pure datasets working group to provide feedback on their approach for dataset metadata. This was also based on the DataCite schema, and it was important to align with Pure to provide connections between projects, publications and equipment related to datasets.

The initial phase of work was overseen by a working group, and sign off for required metadata was essential. Some of the required metadata was specified by DataCite:

- Identifier (the DOI itself)
- Title
- Creator(s)
- Publisher (University of Bath)
- Publication Year
Although some institutions felt this was sufficient, we had concerns that making data available with this limited set of descriptors would not further the aims of making the data discoverable or understandable. We therefore added the following:

- Abstract (a description of the dataset to help users decide whether to access the data)
- Department
- Research Funder
- Methodology (a detailed explanation of what the data are and how they were created)
- Rights Holder(s)
- Contact Person

In practice all datasets have the following fields set automatically, although they can be overridden:

- Version
- Language
- Resource type (introduced as a required field in version 4 of the DataCite metadata schema)

Implementing the metadata schema

For each of the metadata fields we used, we had to decide on whether we wanted to follow the DataCite schema, the Pure schema, or our own implementation. Generally, we preferred the DataCite schema where possible, as this was most compatible with other databases.

Where a format was not specified by DataCite, we considered the Pure implementation. In some cases, we felt the Pure implementation was not effective, such as with contact information. There was a conflict of use cases for this field. For researchers, this was effectively a chance to specify a corresponding author. For us, this needed to be an institutional rather than a personal contact. We were able to implement both use cases in the Archive by enabling each Creator to have an additional “contact person” field. We asked for them to nominate at least one person, and made the first Creator listed a default contact if they did not specify. This also helped us to record the primary contact for future reference, should there be a query with the dataset. For external users, there is a request access button on the record which appears if the data are not publicly downloadable.

However, Pure introduced useful metadata fields for capturing legal or ethical restrictions for internal use. This meant we were able to design EPrints to mirror the metadata included in Pure. The tick box fields mapped exactly, but we decided to have a single text box for the text information about ethics. This gave people the freedom to include further information beyond the main categories. For this field, we concatenated the text from the separate text boxes in Pure.

We also implemented fields which were unique to our repository. In these cases, we considered the information we were trying to capture and how structured that needed to be. In some cases, we decided to capture the metadata in a range of formats. For example, we expected researchers to include documentation. We provided metadata fields to support describing what the data were and how they were created. We also allowed researchers to link to external documentation or upload readme files. As a result, our guidance for using the fields has many caveats and options, which actually makes it harder for researchers to use the system. If we were to review this section, we would favour providing template readme documents and limiting the format to an uploaded file. This would make guidance clearer and enable some automatic validation.
Using the metadata

Import

Researchers start the process by registering their datasets in Pure. The Archive has a custom script that transforms Pure metadata into EPrints metadata. At first, this process was manual. We downloaded metadata from Pure as XML, then uploaded it into EPrints. However, I improved the import to enable direct fetching of XML from the Pure web service. This not only made the process faster, but also enabled us to capture metadata from related sections such as people and projects. Although we copied these metadata manually before this script was updated, the change has affected our researcher engagement. We now encourage researchers to link to projects when adding metadata to Pure as it results in more consistent metadata in the Archive.

I also developed an import script to capture metadata from the UK Data Service (UKDS). This is the standard location for our ESRC-funded researchers to deposit their data. Initially, we intended this to form the basis of an import from DataCite metadata. However, it soon became clear that the UKDS DataCite metadata were minimal, but richer metadata were kept in Data Documentation Initiative (DDI) format. I decided to write a specific DDI import script. This was more challenging than the Pure import, as we had not mapped our metadata to DDI fields, and had to make decisions about which fields most closely matched our schema. We were able to use the same technical process as for Pure to provide an import from the accession number, rather than needing to save the metadata and upload it.

Export

EPrints has a wide range of export formats as standard. At first the most important format for us was the DataCite schema. The DataCiteDoi plugin we use to create DOIs, included version 2.2 of the schema, but only transformed a minimal set of fields. We updated and extended the default schema to version 3. This enabled us to make use of new relations, such as linking to additional metadata, and to list rights in detail, rather than having to summarise for the whole dataset.

We decided to include as much metadata as possible in our DataCite export. The aim was to make our datasets findable in their searchable metadata store. We realised that not all researchers would want full details being shared in this way, so developed the export to be able to exclude groups of metadata from the export file. All datasets still had to export the DataCite required metadata, plus some additional fields which were not considered open to unwanted use, such as version and language. Adding this flexibility had unintended benefits in practice, as we can exclude fields which meet our Archive metadata quality requirements, but do not work with the DataCite system. For example, if a researcher has included an ampersand in a file name, this caused an error when creating the DOI, but to escape the character to make it compatible might have negative consequences for future linking.

More recently, I introduced an export to JSON Linked Data. This embeds structured JSON metadata that Google understands in abstract page headers. Google have produced a draft content type for datasets. We were interested in whether this worked for our diverse data types. The structured data testing tool was important for developing this export format, as it provided a level of validation, and made the metadata human-readable. Of the required fields, only ‘variableMeasured’ was problematic. Although the description claimed the content type applied to any kinds of data, this field is only applicable to datasets providing quantitative data. We hope that this will make our research datasets discoverable outside of the academic bubble.

Reviewing and updating the schema

There have been two main drivers for updates we have made to the metadata in our Archive: external schema changes, and experience of how the system is used in practice.
The update of the DataCite schema to version 4 prompted a metadata review. Some of the new fields mapped closely to our existing metadata, such as the fields to capture detailed funding information. However, the new schema introduced a more detailed geographical polygon field. Previously we had intended to include values for all relevant fields in the Archive. However, we had not seen extensive usage of any geographical coordinate fields in any of our datasets. We decided not to implement this field in the Archive as it would have been a significant piece of development, and we did not feel there was demand to represent geographical information to this level of detail.

We have added new administrative fields in response to our improved understanding of how the system is used. For example, we needed to know different information about metadata-only records. We added metadata to capture information about where the data were held. We can now run reports to identify data which needs to be ‘rescued’ should another repository go out of business. We have also extended metadata in response to requests for data. We expect researchers to include a data access statement in their papers to explain how to access the underlying data. We added a new field to capture these statements, as this is already checked as part of our review process. This provides quantitative data on compliance, which is challenging to monitor in isolation. It has also given us a body of examples which we can use to inform guidance.

Conclusion

Metadata has been at the heart of developments of our research data archiving solution. The need to interoperate with internal and external systems has driven the direction of our developments. However, this has also been balanced by an increased understanding of the management and operational metadata that enable us to curate and preserve the data we hold.

Supporting Information

The metadata schema referred to in this article is available from https://doi.org/10.15125/BATH-00374.

References

The Cardiff University repository team is currently part of a wider research support team that includes Open Access funding and advocacy, research analytics, and the recently-established Cardiff University Press. The team consist of librarians and library staff who come from a wide variety of backgrounds: while the Repository Manager does come from a cataloguing background, the other team members originally worked as library and IT assistants or as library service managers, while several of us have subject librarian experience. Certainly, at Cardiff, we have found that the institutional repository role has widened out from a branch of cataloguing and metadata to one that involves more advocacy, training and direct customer interaction. One thing all of us have in common though is that we enjoy the detailed work involved in creating a perfect metadata record, while a subject specialist background is also useful as you get to know the subject resources and the researchers well, so will have some idea where to find elusive information to create or edit that perfect record.

The history of ORCA 2003-2017: where we got to where we are today

Cardiff University’s institutional repository, ORCA (the acronym stands for Online Research @ Cardiff) started life as Cardiff EPrints in 2003, and in 2008 was upgraded and relaunched as ORCA, becoming a fully supported library service that contained, amongst other research outputs, all the University’s Research Assessment Exercise (RAE) data. ORCA becoming embedded into researcher practice was given impetus by the University’s Research Committee to approve the required deposit of new publications in ORCA, “subject to detailed consideration as to the means by which staff up-load data to the repository”.¹ Between 2009 and 2011, ORCA received JISC funding to create a tool that would streamline the deposit process, minimise the deposit effort and simultaneously ensure capture of all required information without duplication of effort; in short create an easy, user-friendly workflow that would encourage authors to use the repository.² Web feeds were also created by an IT colleague so that schools could use the metadata in ORCA to feed their publications profiles on the school web pages. All new deposits are reviewed by repository staff before submission to the live database; the publication data is checked and, where available, information such as DOI, URL, abstracts, volume, issue and page numbers, keywords and subject fields are added. Because minimal data is captured from the user, there is more emphasis on the role of the library professional in adding data and making the record as complete as possible. This is where a cataloguing or metadata professional is invaluable in institutional repository work.

¹ Harrington, Louise and Scott Hill. 2011. The I-WIRE Project final report. Available at: http://orca.cf.ac.uk/74670/
The new deposit tool, the ability of the web feeds to update web pages within 24 hours, and the requirement from the University Research Committee for researchers to deposit bibliographic details of their work into the repository, combined with the decision to use ORCA for the 2014 Research Excellence Framework (REF) meant that researchers or administrative staff from all the schools began to deposit into ORCA in rapidly increasing numbers. Using ORCA as the repository for all the research outputs that could potentially be submitted to REF2014 meant a great deal of work for the central repository team, especially as many of the outputs were not already on ORCA, which resulted in enormous backlogs of outputs to be deposited, edited and reviewed. Getting the research into the repository also became very time-critical due to internal ‘mock’ REF programmes: a lot of schools would have internal deadlines that we did not know about which lead to a lot of pressure on the small central team. As well as the repository team and cataloguers working on ORCA, there was invaluable support from subject librarians and library assistants, who did a lot of adding and editing of records, while the repository team would do the final checks and submissions. The decision to involve more library staff in the process of course meant widespread training and ongoing support – another skill that ORCA team members have had to learn and develop!

No sooner had we submitted our return for REF 2014, then we had to learn new processes for the next round of the Research Excellence Framework (the phrase ‘moving the goalposts’ springs to mind!). The next REF requires journal articles and published conference papers accepted after 1 April 2016 to be made freely available as open access documents, which is a major adjustment for both academics and library staff, requires us all to learn new rules about copyright and embargo lengths, and also means a change to our editing processes. Because outputs have to be deposited at acceptance, our role is no longer so much checking and adding publications details, it is instead ensuring that deposits are added to the depositing in a critical time-frame and that publisher copyright rules are checked and embargo lengths are calculated correctly. It also involves a lot of liaison with our colleagues in the library Open Access team regarding payment of gold open access fees and correct license statements. While metadata skills are still needed, particular for items such as books, book chapters or reports which are not affected by the REF open access policy, knowledge of copyright and publisher requirements and advocacy skills are increasingly needed for institutional repository work. Institutional repositories were of course initially created as open access repositories in an attempt to circumvent the ‘serials crisis’ of the 1990s and 2000s where academic journal articles were hidden behind paywalls and library subscription charges skyrocketed. Due to many factors, however (for example, being unaware of copyright issues, suspicion of, or lack of understanding of, open access, simply not keeping the correct version of the file, using subject repositories such as ArXiv, the Physics pre-print server), institutional repositories have tended to be metadata archives rather than full text databases; at Cardiff full text files account for between 10-20% of the records. While there are always individual authors who support open access, and some disciplines which are open to it, it has been a slow process to educate and engage academics in the positive aspects of open access to research, one that is still ongoing.

**Tips, hints and pitfalls**

Institutional repositories have become increasingly central to university administration and research processes, especially with the REF and funder requirements. Good (or poor) quality institutional repository data affect as all audiences in the University – from academic staff to administrators to PhD students who, as a result of a change in University regulations, have been required to make their theses available online through ORCA since 2011. Positive aspects of the role are that you do get a lot of feedback, the vast majority of it being pleasant – something that you don’t always tend to get as a cataloguer! You also do get a lot more contact with colleagues across the University, whether via email, phone or in person, and there are a lot more opportunities for training and advocacy, if this is an area which interests you; it is good for continuing professional development.
As cataloguing colleagues will understand, the quality of the data that is uploaded to ORCA depends increasingly on the quality of the data that we are given or can source; put simply, if rubbish data goes in, then in most cases, rubbish data comes out! Because of the requirement to make deposit as easy as possible for researchers in order to engage them, our system allows bulk imports from external databases that we have found to be poor sources for good publication data. By making things easy for depositors, we have made things more difficult for ourselves which can be frustrating. The other reason why data is suffering is due to REF requirements: because articles and published conference papers now have to deposit at acceptance, rather than publication, authors have very little information about the paper, and there is nothing online that we can access to improve the record quality. It is also fair to say that unless you come from a cataloguing background, you can miss the face-to-face role of the library site, as institutional repository and scholar communication teams tend to be office-based. However, we have found that many colleagues in the library sites wish to help with ORCA so we do get to visit libraries to help train and engage people.

Conclusion

A background in cataloguing is a huge asset when working on an institutional repository because of the detailed data quality work involved, and the emphasis on findability, but as we have shown, staff from any background can work in this area. While perhaps in its initial incarnation, institutional repository work was seen as mostly for metadata professionals, the move towards open access and the subsequent need for training and advocacy in that areas means that we are developing lots of different and transferable skills. There are lots of opportunities in this area with universities moving towards using Current Research Information Systems (CRIS) and increasing interest in open Research Data Management (RDM) which could provide further opportunities to transfer skills used in metadata work across to these systems and wider research support areas. Cardiff University’s institutional repository’s main aim is ultimately to promote and make our research (and where possible, our research outputs) available to the widest possible audience across the internet in the true spirit of open access. Our role in enabling this with good quality metadata gives us a lot of pride and job satisfaction, whatever our library background, in a job well done.

Bibliography


Harrington, L. and Hill, S. 2011. The I-WIRE Project final report. Available at: http://orca.cf.ac.uk/74670/

Introducing ORA

The Oxford University Research Archive (ORA)\(^1\) has been in place as Institutional Repository (IR) for the University of Oxford since 2007, with the purpose of providing a permanent and secure online archive for research materials produced by the members of the University. This includes, but is not limited to the following:

- Journal articles
- Conference papers
- Working papers
- Technical reports
- Patents
- Book chapters
- Books
- Posters

The additional ORA-Data service\(^2\) was launched in 2015 as part of the existing IR in response to external drivers, namely the expectations laid forth by the Engineering and Physical Research Council (EPSRC) with regard to research data management.\(^3\)

ORA and its Open Access content\(^4\) is available to anyone with an Internet connection, though deposit to the archive requires University authentication. The repository, and its content, is managed and maintained by staff within the Bodleian Libraries\(^5\) although various stakeholders around the University are involved with decision making regarding its function.

Metadata records and files are received from multiple sources. At present most deposits are received from the University’s Current Research Information System (CRIS) – Symplectic Elements (SE).\(^6\)

SE is ORA’s main source for records of traditional research publications (journal articles and conference papers) and grey literature such as working papers and reports. Theses and research data are received directly to ORA via in-house\(^7\) deposit forms. The source of record content for ORA currently influences the means by which deposits are reviewed and enhanced before a record (and file(s) where allowed) becomes available on the public website.

The ORA review team comprises an institutional repository librarian, a curator of digital research data, six metadata assistants, one research archive assistant and one senior research archive assistant, with the length of time in post varying from 1-5 years. The team has doubled in number over the last few years to account for the increase in number of deposits being received to the repository.

---

\(^1\) https://ora.ox.ac.uk
\(^2\) http://libguides.bodleian.ox.ac.uk/ora-data
\(^3\) https://www.epsrc.ac.uk/about/standards/researchdata/expectations/
\(^4\) Content available for download from the public pages. Other material may be stored in the archive under an embargo period awaiting release.
\(^5\) http://www.bodleian.ox.ac.uk/
\(^6\) Information from the company Symplectic Ltd. can be found at: http://symplectic.co.uk/products/elements/ and local to Oxford at: https://www.admin.ox.ac.uk/researchsupport/awards/symplectic/
\(^7\) Forms developed by ORA that form part of the application, not third-party application components (i.e. SE).
ORA Metadata Assistants are predominantly responsible for the review and enhancement of metadata associated with deposits made via SE and research data deposits, whereas the research archive assistant’s main responsibility is for theses deposits.

In-house deposit and review

Creation and submission of records in-house, requires the user to complete an online deposit form which is broken into multiple stages. Each stage includes the option to add additional information as they proceed (Figure 1). The forms were developed with the intention of capturing sufficient information to facilitate the review of the deposit and to fulfil any requirements with regards to metadata from internal or external mandates.

Figure 1. Snapshots of sections of the deposit forms for theses and data. Each deposit form has specific fields intended to facilitate and ease review of the records and also ensure compliance with external requirements for metadata.

---

8 EPSRC policy affecting research data has already been mentioned. For theses at Oxford it is a mandatory requirement as part of exam regulations to deposit an electronic version of the final examined work to ORA for all students who began their programme of study after 1st October 2007.
Once submitted the record moves to a separate review stage accessible only to ORA review staff preventing further edits to the submission by the depositor, although a function to refer the deposit back is available if needed.

ORA review staff are notified of a new deposit via email as the record reaches the review area. Deposits are ‘claimed’ by individual team members before beginning the review of a record to avoid conflicting reviewer edits. The review and enhancement of metadata for the records received via the in-house forms is made within the forms themselves. Correction and completion of information to the fields is done within each stage of the form.

Title fields for deposits are amended to conform to RDA (Resource Description and Access) standards and the abstract is enhanced with HTML (Hypertext Markup Language) and Unicode\textsuperscript{9} where necessary, so to render the abstract that displays on the record page as close to the papers content as possible. Where abstracts contain equations or molecular structures these are reproduced using HTML \texttt{<table>} tags (Figure 2) or where this may be too difficult to render, a JPG image file may be created and uploaded to the record page as a downloadable object. The image can then be referred to and displayed in the record abstract using the HTML \texttt{<img src=URL>} syntax (Figure 3).

**Figure 2. Rendered equation and HTML \texttt{<table>} mark-up used.** The table \texttt{<table>}, table row \texttt{<tr>} and table cell \texttt{<td>} properties are used to determine the positioning of sections of the equation to compile it to the correct form (See ORA record: https://ora.ox.ac.uk/objects/uuid:4b48ebc4-dd04-4741-8ab3-1ac8620f1cb0/).

\begin{equation}
|\nabla a|^2 (b,\cdot) \leq C \frac{||a(0,\cdot)||^C_\infty - a^C (d,\cdot)}{b}\text{, and }\frac{|\nabla a|^2}{\alpha^1_0 (d,\cdot)} - \frac{\partial_b a}{\alpha^2_0}, (d,\cdot) \leq \frac{C_1}{b}.
\end{equation}

\begin{table}
\begin{tabular}{|c|c|}
\hline
a & b \\
\hline
\end{tabular}
\end{table}

\textsuperscript{9}http://unicode.org/
Controls within the forms have some crossover with traditional cataloguing through the standards being used in the input of specific fields. FAST (Faceted Application of Subject Terminology) subject headings are provided as choices for subject information added to the forms, and Library of Congress (LoC) ISO (International Organization for Standardization) numbers are used for the language selected for the work, and date information.

Key aspects of traditional cataloguing and RDA standards such as Name Authority Records (NARSs) are not used. Author names and affiliations used to populate author information to the forms are obtained by querying the University’s Core User Directory (CUD). As information is added to the field the user is able to select an existing University of Oxford author, which automatically populates institutional affiliation information (Division and department).

Due to not all authors being based at the University the field also allows for a free text entry. This adds to difficulty in linking works by a single author – a difficulty that is hoped to be overcome with the increased uptake of ORCID\(^{10}\) identifiers throughout the institution, and their incorporation into the underlying record metadata.

**Metadata received from external sources – Symplectic Elements (SE)**

ORA has been in receipt of metadata records and full text files from SE since 2008. In 2014 the Higher Education Funding Council for England (HEFCE) announced a Policy for open access in the post-2014 Research Excellence Framework (REF).\(^{11}\) The Bodleian Libraries ran a pilot project in response to the policy to determine the most suitable deposit method for research publications affected by the policy, with the least burden on academic depositors.

---

\(^{10}\) [https://orcid.org/](https://orcid.org/)

\(^{11}\) All journal articles and conference published within a journal or conference proceedings with an ISSN accepted after 1st April 2016 are to be deposited to a repository within three months of acceptance. Updates to the policy have since seen this changed to within three months of publication up until 1st April 2018.
SE was selected as the deposit method for these items as the outcome of the pilot. The policy came into effect on 1 April 2016 and as a result the interaction between ORA and SE has increased significantly. Deposits have increased from an average of 80-100 received for review by ORA staff per month pre-April 2016 to an average of 1100 deposits per month post-April 2016.

Records in SE are created by a user completing a short deposit form and uploading a file (Figure 4) or by ‘claiming’ an existing record that has been obtained from an external datasource such as PubMed Central\(^\text{12}\) or Web of Science.\(^\text{13}\) The deposit form was designed to be short and concise in order to make the deposit of an item quick and easy for academics, researchers and administrative staff. This does however mean that the additional content be added at review by ORA staff.

Review staff are notified of the deposit via email upon transfer of the file and metadata to ORA. Deposits made via SE are dealt with in a different review area to those received through the in-house forms, and involve editing the XML (Extensible Markup Language) directly.

There are multiple datastreams (XML and other metadata content files) for each record created for a record received from SE. These include:

- information of the related item within SE,
- a MODS (Metadata Object Description Schema) datastream,
- a DC (Dublin Core) datastream, and
- a citation datastream.

Enhancements and edits are made to the MODS datastream as part of the review process. This is done by incorporating the content of the MODS datastream into a new datastream called xml_ORA. Edits are then made to the xml_ORA datastream and the datastream is then preserved or ‘locked’ preventing further updates made to the record from SE overwriting the work done by reviewers.


\(^\text{13}\) [http://apps.webofknowledge.com](http://apps.webofknowledge.com)
Full text files deposited along with the metadata are looked at as part of the review process. Information gathered from the file(s), and any information that may be sourced from a publication website is used to add additional information to the record (Figure 5).
An XML template is used to add the additional information into the xml_ORA datastream. The same RDA principles are used in reviewing the XML as within the in-house forms, but for fields such as date of publication the ISO format is applied manually copying from the XML template. The template covers common fields that are added or changed as part of the standard review process, such as publisher information, publication status and information regarding intellectual property rights, including copyright.

The current MODS schema has been customised for ORA specific review and is therefore many versions behind the latest available. Once edits to the metadata are completed updates to the MODS datastream are pushed from the xml_ora datastream. It is the MODS datastream that is used for the creation of the information displayed on the record page. DC metadata is automatically created using the MODS datastream. A record that has completed the review process is significantly more enriched with information than the original deposit, aiding the discoverability of the record and its usefulness as a resource (Figure 6).
Traditional cataloguing experience within the ORA review team is varied and is a result of training, education or previous employment. While undeniably helpful, this experience is not essential to undertake the review work. Experience with HTML is a benefit, but again not essential. Most of the in-house form review consists of data-entry in forms, and the underlying code provides the controlled vocabulary for the information entered. This is in contrast to the review of SE items where a template is used.

It is worth noting that review of deposits requires the review team to have extensive knowledge of policies affecting the dissemination of scholarly research, as well as being able to interpret publisher policies surrounding self-archiving. This is in addition to the knowledge of issues such as third-party copyright, the inclusion of sensitive content, and licencing.

A large amount of liaison with depositors is necessary during the review process, including communication about policies and requirements surrounding the deposit and access of a record, and any available full text.

---

14 Often referred to as Green Open Access. Making an author accepted manuscript or Preprint available in a repository either immediately, after publication or on expiry of an imposed embargo period.
Onward sharing of metadata and future developments

As noted there are external policies that ORA needs to comply with in regards to metadata. These include policy regarding what information is recorded within the metadata and the format that this takes. For the HEFCE REF policy, one of these is the RIOXX Metadata Application Profile (RIOXX). 15 RIOXX combines standard DC with additional RIOXX terms with the aim of aiding institutions in compliance with HEFCE metadata requirements.

ORA currently uses mappings from the MODS information to conform to RIOXX standards. An OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) endpoint has also been developed to make record information available to internal and external services, including the Bodleian Libraries’ own catalogue, SOLO. 16

The ORA service is about to embark on a two-year project to upgrade most of its software, infrastructure and application stack design. An updated and scalable data model is being developed, which will see further improvement to ORA’s compliance with endpoint standards (including OpenDOAR 17, OpenAIRE 18, UKRDDS 19). It is anticipated that the current review process will see many changes as work begins, including the switch to writing metadata in RDF (Resource Description Framework) Triples. It is also planned that the review process will be performed through the use of further in-house forms, moving away from editing the xml datastreams directly.

15 http://rioxx.net/
16 http://solo.bodleian.ox.ac.uk/
17 http://www.opendoar.org/
18 https://www.openaire.eu/
19 https://www.jisc.ac.uk/rd/projects/uk-research-data-discovery
Kingston University is a multidisciplinary University based in Kingston upon Thames in London. Amongst the variety of subjects on offer are degrees from the Kingston School of Art, at the forefront of art, design, and architecture education and research internationally, and the School of Performance and Screen Studies, where the emphasis is on subjects relevant to the world today taught by world-leading researchers and talented and experienced creative practitioners.

History

Kingston University’s institutional repository (http://eprints.kingston.ac.uk) was developed during 2005-2006 to support the University’s RAE2008 submission. The initial project was led by Jane Savidge, at that time Head of Learning and Research Support at Kingston University. Her considerable previous experience as the Faculty Librarian for Art, Design and Music at Kingston University, as well as at Central St Martin’s and as Chief Cataloguer at the National Art Library, strongly influenced the shape of the repository.

Hence, practice-based outputs were central to the initial configuration of our EPrints repository. A list of item types was constructed to best support these types of research outputs:

- Artefact/device
- Composition
- Design
- Exhibition
- Still/graphic work
- Moving image/broadcast
- Performance/live event
- Sound recording

New metadata fields, some with controlled selection lists, were added to capture elements of the physical aspects of the output as well as its place within a larger exhibition or performance. These evolved further in response to feedback from academics that tested early versions of the repository. Citations were configured based on Harvard referencing. The structure of the content in each field was based on AACR2, the cataloguing convention at the time (influencing how we used punctuation and capitalisation). The Research Repository was launched in the autumn of 2006, using EPrints 2, with the first record being made live at the end of October 2006.

Kultur and Kultivate

The early work undertaken meant that we were very interested to follow the progress and outcomes of the Kultur project 2007-2009, particularly the final reports on Metadata and Technical Developments, and we were able to participate in various Kultur II (2009-2010) meetings.

---

4. VADS (no date) Kultur II Group. Available at: https://vads.ac.uk/kultur2group/index.html (Accessed: 8 June 2017).
The Kultivate project (2010-2011), which arose from Kultur II, was tasked with finding strategies (digital and otherwise) to increase the deposit of practice-based arts research outputs into repositories. A larger number of institutions participated in this project and a range of common themes and challenges in the area of supporting practice-based arts research outputs emerged. A recent article gives a timely and interesting discussion of the ten-year anniversary of Kultur and its ongoing impact.⁵

A major upgrade and refresh of the underlying EPrints software of our Research Repository was undertaken during the summer of 2011 and we took the opportunity to enhance our existing support with some of the outcomes from the Kultur project. This was before EPrints software became ‘Kulturised’ as a matter of course.

We implemented the following items which were adapted from the Kultur project final report:

- Promoting the visual presentation of images and sound files over metadata values in the public view of the record
- Splitting the initial record creation screen into two pathways. The first screen asks the question “Which option best describes this item?” with two choices
  - Text-based or non-text or practice-based
  - The phrasing was taken from Higher Education Funding Council for England (HEFCE) language
  - Upon selecting ‘Practice-based’, our original list of practice-based output item types is then available
- Introducing some new metadata fields
- Adding a new workflow stage for Events to capture both multiple locations and events for objects and performances or exhibitions. This expanded our original configuration which captured a single event only.
- Adding Scorch files (music score software) as an uploadable file type

Technical issues remain a considerable challenge in ensuring that already deposited sound and video files remain usable over time. During EPrints software upgrades, we have experienced problems with ensuring that existing video and sound files play and present properly. These issues are often connected with changes in underlying server technology or software and their interaction with web-based media players. We therefore include testing the media files in our test scripts for each upgrade.

**Putting it all into practice**

Regardless of the level of planning, the proof of the metadata pudding is in the eating recording – so how did we put this into practice?

Much of how we approach the recording of items is in collaboration with the academics concerned. Often we’ve met with them to discuss the sort of work they do, to demonstrate the Research Repository and to highlight what it can offer in terms of promoting their work.

One of the first challenges is that the way in which two people view their work will differ. Some view the creation (composition/play/artwork) as being the main output, with the different places it was performed/exhibited being simply the way it is shared. Others, however, view the location as integral to the piece, and that each performance/exhibition exists as separate and distinct outputs. Often the academics are quite surprised that we can be flexible (up to a point) in how we record the items, but they appreciate that we understand the difference for them.

---

Once you have assessed whether you require one record or more, then the next challenge is to determine what to actually put in the record. When we implemented the Kultivate project improvements, we spoke to the Music Subject Librarian and investigated what other repositories were doing in order to make decisions on how to interpret the fields on offer. The aim is to be as consistent as possible, but the unique nature of practice-based items can make it difficult.

The title is often the most straightforward, although the usual convention of only capitalising the first word and any proper noun after that often has to be overlooked as even the use of capitals can be considered important. The first main difficulty is in recording the creator’s contribution. Within the repository software we can record multiple creators and assign a ‘role’ to each. These roles are based on the Dublin Core Relator Terms. However, we cannot record multiple roles for the same creator. It is not unusual for a practitioner to identify three (or more) distinct roles that they had in the creation and expression of the piece – for example, director, performer, producer, researcher. In these instances, we have had to come to the compromise of either not specifying any role in the creator field or recording what they feel to be most significant in the creator field. We can then include information on the other roles in a general note field.

With practice-based items the traditional notion of publisher and place of publication goes out of the window. Even date can be a tricky one to capture. Whilst we can now record multiple dates in the system, only one date will appear in citation. We can display multiple events associated with the record, but these may be different from when the academic feels the work to be ‘created’. Verifying this ‘creation’ date can be almost impossible as it’s only the performances that are recorded, not the date it was finished. The date matters greatly in terms of inclusion in the Research Excellence Framework (REF), so getting the date consistent between items is important.

When it comes to recording events, we had to modify the practice developed for conference papers. With conferences, we recorded the place as town and country only, as the host institution of the conference is often not significant, just the location the organisers happened to choose. However, the gallery, concert hall or theatre that the practice-based piece event took place in can carry more importance, not only from the prestige point of view, but also from the differing layout or acoustics. Therefore, we record festival/event name (if appropriate), gallery/hall, town and country.

Verifying dates and places can be problematic, especially for the smaller galleries – webpages get updated, posters get binned or programmes lost. We’ve now put together guidelines for the academics which details what forms of evidence we accept. This means they can retain what we require at the time of the event, saving time and effort in having to ask for them later. Additionally, the Wayback machine has often been invaluable in delving back through the mists of internet history.

---

7. Contact us if you would like a copy.
With the HEFCE REF Policy the place of full text in Repositories has been given greater importance. We had always welcomed full text for all items, but had never actively sought it, valuing the recording of metadata for all outputs over trying to explain the complex minefield that is full text permissions and restrictions. Over the last year, in addition to dealing with the text-based items covered by the policy, we have had more queries from those academics with practice-based outputs concerned about how the policy will affect them. Their reluctance regarding full text is frequently because they feel a photograph or video cannot fully express the work as it has to be experienced in context. Having said that, some researchers are keen to enrich their records with pictures, video or audio. This has meant we have to find out about and explain the elements of copyright and re-use (including Creative Commons licences), which may be unfamiliar to them.

**Engagement and workload**

Despite our best efforts, not every meeting results in increased engagement, and we’ve had to accept that it’s not personal! Many of the academics involved are on very fractional contracts and some don’t feel the same engagement with the University as full-time staff may do.

Of course, if you do get engagement, then you have to be prepared for extra work. It is quite time-intensive to have meetings, deal with the records and follow things up, however worthwhile it is. With the HEFCE REF policy and its requisite time pressures (alongside the consequential uptake in general Research Repository engagement), it’s easy sometimes to overlook these records in favour of ‘easier’ or more ‘important’ work. You have to take each item separately and allow yourself the time to fully understand the complexity of the output, and how to record it. As with all Repository work, if you need to contact the academic it’s important to stress that it’s not about us doubting the existence of an item, it’s about us ensuring it’s represented in the fullest, most comprehensive way.

Early on in the Research Repository history we tried to get academics submitting records for their own items and for the text-based outputs. This has been fairly successful, especially as their outputs in the Repository are listed on their profile pages. Providing clear instructions for books or articles is fairly straightforward; however, due to many of the factors already listed, practice-based instructions are much more difficult. We have drawn up guidance on the main fields, but are always keen to emphasise to the academics that we will work with them in recording their work.

**Conclusion**

Writing this article has reminded us of all the work that has been put in over the last 10 years. We’ve constantly refined and updated our understanding of the recording of practice-based items and changing the architecture of the Research Repository to accommodate this knowledge means we can more easily record outputs in more common formats, whilst also giving us flexibility to capture more complex outputs. REF2021 is on the horizon so it’s more important than ever to continue to maintain the dialogue with all involved. We also follow with interest what other institutions are doing to support practice-based researchers.

---

The main premise of this book is that information professionals are, due to their experience of knowledge organisation systems and subject domain knowledge, ideally placed to contribute to the semantic web and information retrieval, particularly through the development and use of ontologies. Furthermore, as stated on the back cover, the author’s aim is to provide “an accessible introduction and exploration of ontologies”.

Why ontologies? Because “those who successfully find ways of managing the information overload, and of making use of the increasing quantities of data available, will have the competitive advantage.” (p. 4)

If you are a student or new to linked data, the semantic web and the role of ontologies within it, this book will be a useful guide to these areas as they relate to information professionals. The author’s writing style is concise, clear and each argument is supported by well-chosen references. The arrangement of the chapters follows a logical progression whilst equally allowing the reader to dip straight in to areas of specific interest.

The first two chapters are introductory ones. The first chapter clarifies the position of ontologies in relation to other knowledge organisation systems and emphasises their value before offering a preferred definition: “An ontology is a formal representation of knowledge with rich semantic relationships between terms.” (p. 12) The second chapter focuses on the elements and standards which form the semantic web stack, introducing the concepts of RDF, RDFS, OWL, SPARQL and others. It also evidences the involvement of libraries and other cultural heritage institutions with the semantic web to date.

Chapters three and four outline some of the dominant ontologies information professionals are likely to encounter and emphasise the importance of ontology re-use. The third chapter begins with a closer look at RDF, SKOS, OWL2 and others used in ontology construction. It then categorises and provides introductory paragraphs for, amongst others, Dublin Core, the Bibliographic Ontology, FRBR, RDA, EDM, CIDOC-CRM, FOAF, DBpedia and Schema.org. The fourth chapter guides the reader through available tools for sourcing existing ontologies and methods of evaluating an ontology’s suitability for use.

Chapters five and six, on building and interrogating ontologies respectively, are where things get really interesting. The fifth chapter offers a useful overview, identifying three types of methodologies for ontology creation and a twelve-step practical approach. An ontology development example, ‘Bibliometric Metrics Ontology element set’, following these steps is also provided. The sixth chapter explores three reasons for ontology interrogation: suitability for re-use, information extraction and analysis of an ontology’s use. It then looks at currently available search technologies appropriate to each of the three stated reasons.

The reality is that “information professionals are far more likely to use an ontology or knowledge base than develop one” (p. 135) and that many of the tools available for exploring ontologies are currently “aimed at computer scientists or professional ontologists rather than the casual user.” (p. 154) However, in these two chapters the author re-asserts his view that it is important for information professionals to not only use but also build ontologies because many areas “still require the development of ontologies” (p. 97) and “it is only through the development of ontologies that we come to fully appreciate the associated tools and technologies.” (p. 97)
It is then up to the final chapter, chapter seven, to sum up and look to the future of ontologies in general and the role of the information professional in particular. The author envisages that “ontologies will undoubtedly be increasingly embedded within many of the traditional roles of the information professional, such as cataloguing, classifying and indexing, as well as newer roles such as supporting the publication of data and gathering new indicators of impact.” (p. 158)

If you take one thing away having read this book, I hope it is this. As library services evolve to meet the information needs of today’s researcher the skills of information professionals must also adapt and evolve. Taking the time to become familiar with firstly, the role of ontologies both within the semantic web and within libraries, and secondly, the practicalities of ontology re-use and creation, whilst not necessarily being easy, is a very worthwhile, natural and perhaps even essential step forward for today’s cataloguer and indexer.
Catalogue & Index is electronically published by the Cataloguing and Indexing Group of the Chartered Institute of Library and Information Professionals (CILIP) (Charity No. 313014)

Advertising rates: GBP 70.00 full-page; GBP 40.00 half-page. Prices quoted without VAT.

Submissions: In the first instance, please contact the Co-editors:
Karen Pierce: PierceKF@Cardiff.ac.uk
Deborah Lee: Deborah.Lee@courtauld.ac.uk

Book reviews: Please contact the Book Reviews Editor:
Anne Welsh: a.welsh@ucl.ac.uk

ISSN 2399-9667

CIG website: http://www.cilip.org.uk/cataloguing-and-indexing-group/catalogue-index

CIG blog: http://catandindexgroup.wordpress.com/

CIG twitter: @CILIPCIG

Tags from the CIG blog on specific areas of interest:

authority control  book reviews  Catalogue and Index  cataloguing  CIG activities  CIGS classification  committees and working groups  conferences  Dewey  digitised material  Dublin Core  events  folksonomies  linkblog  MARC metadata  news  RDA  Semantic Web  social software  standards  taxonomies  UDC