Tom Ensom
UK Data Archive, University of Essex

Research Data Essex repository development and ingest

Introduction

EPrints[1] is a repository software solution, and was designed with the primary goal of lowering the barrier to deploying an institutional publication repository. This gearing towards article-type deposits is apparent in its design. At the University of Essex, as part of the JISC-funded Research Data @Essex project[2], we set out to pilot EPrints as a comprehensive research data repository solution. We aimed to do this in accordance with the needs of the University of Essex as a research led institution, while utilising the UK Data Archive's research data management expertise. Ultimately, the pilot has resulted in the release of two key reusable outputs: a generic metadata profile[3] for describing research data, and a 'plugin' for EPrints called ReCollect[4], which allows anyone with an EPrints install to implement our customisations.

In this report each of these outputs is described in detail, particularly focusing on the rationale behind particular decisions taken in the development process. We also describe testing and community engagement activities and how they have impacted the outputs, as well commenting on outstanding issues that have become apparent.

The University of Essex research base is broad, and many types of data are produced. This could include terabyte-scale outputs of large scale proteomics, to at the other extreme, small-scale business management analyses in a single Excel spreadsheet. It was

an essential requirement that any system we developed be generic enough to store, describe and present data from any of the disciplines represented at Essex. This approach also ensures relevance to UK higher education institutions more widely. We were also keen to ensure that the deposit process did not become overly burdensome for users, in order to minimise uptake barriers. Taken together, the above considerations involved a necessary compromise between ease of deposit and the need for sufficient information to fully enable re-use.

Throughout the project we worked closely with a selection of four pilot departments at the University of Essex, this consultation process allowing us to meet as closely as possible the needs of real research groups. In in-depth demonstrations they were able to offer feedback and suggestions with their own unique perspective, and directly influence development.

We have also engaged with the EPrints and JISC communities, resulting in an approach that joins up with other institutions and reflects recent developments in research data management.

Definitions

This document assumes a baseline level of knowledge about repository technology and information management, but a few specialist or confusion terms are defined below.

Eprint

A structural component of an EPrints repository. An *Eprint* might be a journal article, a set of images or a complex data collection. Note typographic separation in meaning by up and lower case ‘p’.

Document

A structural component of an EPrints repository. At the level below Eprints (see above) are Documents. A Document would typically be a single file, but could be multiple dependent files that form a discrete unit (e.g. a GIS database or a complete web page).

Data collection

A data collection may consist of many data and documentation files. Taken together, the files form the discrete basis of a research project.

Ingest

The process of deposit and processing - by either depositor or repository staff - before a data collection is stored and the file system and (potentially) made available in the data catalogue.

Plugin

In an EPrints context, a plugin is an extension of functionality that can be optionally ‘plugged in’ to a repository installation through the EPrints Bazaar system.

ReCollect metadata profile

The default EPrints set of metadata fields is well suited to describing research publications. However, it does not provide sufficient detail for the description of research *data,* which requires extensive metadata if it is to be made re-usable. For example, there may be crucial methodological information required as to the apparatus and settings used to collect measurements in order for them to be meaningful.

We set out to define a set of elements for providing a description of generic research data; that is research data from any conceivable discipline. To do so we wanted to employ existing schema and standards appropriate to this kind of content, rather than formulate something from scratch and further contributing to the bloated and confusing world of metadata schemas. The basis of the extension of the default EPrints metadata was a three layer metadata model of:

- core (citation, discovery);
- detail (descriptive, contextual); and
- discipline specific (as additional file)

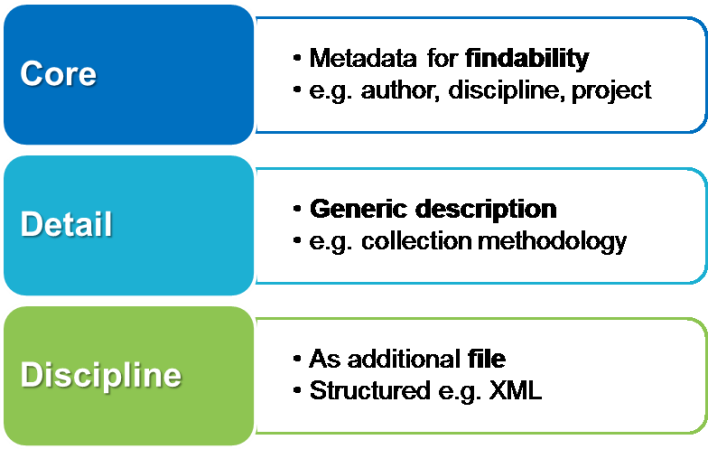

Figure 1. Metadata model, based on JISC IDMB project[5], University of Southampton

This model is built on work from the JISC IDMB project[5], a previous JISC MRD project based at the University of Southampton. It also drew on the UK Data Archive's (and international sister data archives) own approach to archiving social science data collections, developed over the past 40 years. The top two of layers of this model were constructed using elements from several existing schema which are detailed in Table 1 along with the reasons for their use.

We also mapped to the DataShare metadata profile[9], used in the University of Edinburgh's digital repository of the same name. This provided a useful comparison with another well-developed institutional approach and a tentative exploration of interoperability potential. In formulating this, the DataShare team themselves had also mapped various schema to suit broad collection-level description. Detailed documentation of the DataShare metadata schema in use provided useful guidelines for controlled vocabularies and field validations.

Table 1. Metadata schema used in construction of ReCollect metadata profile, with notes on compliance and rationale behind use.

SCHEMA	AREA	COMPLIANCE	REASON FOR USE
DataCite[6]	Data citation	Yes	Minimal mandatory. Essential if DataCite DOIs are to be minted. Emerging as a *de facto* standard among data repositories
INSPIRE[7]	Geospatial	Yes	EU standard. Intended to cover many different types of data with geospatial content – as a result, enables a fairly generic description
DDI 2.1[8]	Social science	No	Descriptive/contextual metadata beyond scope of above schema e.g. collection methodology, ethics/consent statement

We used the HESA JACS3[10] subject classification scheme over Library of Congress[11] headings that can be loaded into EPrints 'by default'. We felt this resulted in a much improved mapping to the disciplines represented at Essex, and indeed UK higher education as a whole. Whether a standard emerges among institutional data repositories will depend on community agreement. Initial investigations indicate that the recently published RCUK scheme is another strong candidate.

Modifying EPrints and developing the ReCollect plugin

The metadata profile outlined above was implemented in our pilot install of EPrints, through addition of new fields and modification of existing fields. This necessitated a number of changes to the data catalogue, primarily so that it could adequately display the more than 50% increase in number of metadata fields. It would be expected that most deposits would consist of many files, and on occasion hundreds, so adaptations were required to adequately present these. A comparison of the default and ReCollect versions of the citation page (EPrints terminology for the outward facing record for a data collection, presented through the data catalogue) can be seen in Figure 2.

Field labels and help text embedded in the workflow were tweaked considerably in order to make working with the expanded profile easier for depositors.

The ReCollect plugin is now available for free from the EPrints Bazaar[4]. This can be accessed through the Bazaar website, or directly through the EPrints admin interface. The latter gives the option to install (or uninstall) the plugin with a single click. This creates an extremely low barrier to deploying EPrints as a data repository. The plugin includes all the highlights mentioned in this section of the report, including the metadata profile. Technical documentation beyond the scope of this report will be provided through a dedicated page on the EPrints wiki[12] (maintained by the University of Southampton), complementing the detailed commenting of the code itself.

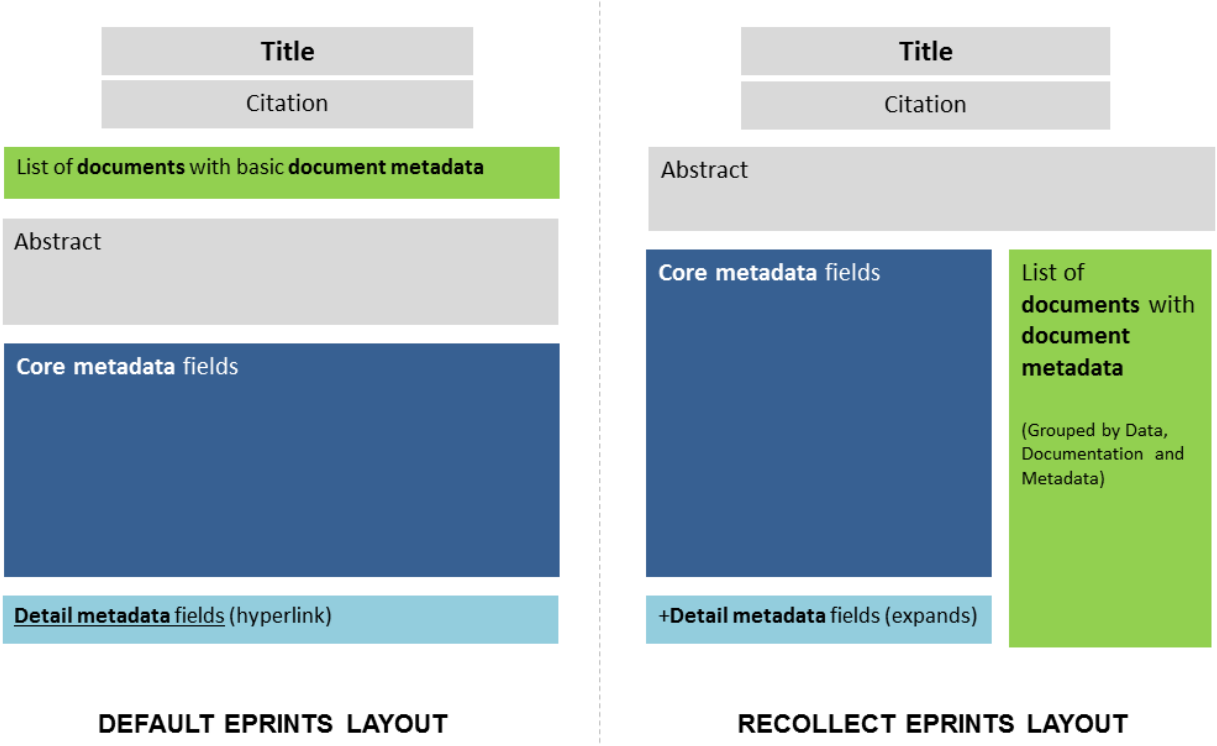

Figure 2. Comparison of the default EPrints citation page layout (EPrints terminology for a public record in the data catalogue) and the modified ReCollect layout: the ReCollect design is more space-efficient to allow for the presentation of multiple files and expanded metadata, while keeping to a single page.

Sample data ingest and user testing

Sample data ingest

The primary purpose of this stage in development was to test the pilot repository's suitability to real-world research data (and in doing so, the ReCollect plugin), while also engaging with peers in the research data management and repository communities. We gathered representative sample datasets from four departments at the University of Essex (Biological Sciences, Language & Linguistics, Business and Computer Science and Electronic Engineering), covering sub-disciplines within each. The researcher volunteered sample data collections underwent ingest procedures including the preparation and upload of files, and the addition of metadata at the document and collection level metadata.

In addition, we trialled ingest of a selection of datasets from the ESRC Data Store, a self-archiving service for ESRC-funded research operated by the UK Data Archive. These were useful for testing as exemplar material that has already been described with extensive metadata. Furthermore, given that the interaction between data centres and institutional repositories is likely to be of increasing importance, we thought this would be a worthwhile exercise in exploring interoperability. We found the two well aligned, despite the discipline-specific nature of ESRC Data Store.

Table 2. Sample data collections and the research groups and departments from which they originated

DEPARTMENT	RESEARCH AREA	SAMPLE DATA
Biological Sciences	Proteomics	mass spectrometry data from tumour tissue samples
	Bio-imaging	high resolution image data collected to examine cellular structure
Essex Business School	Management	tabular data on football managers, using publicly influenced performance metrics to examine managerial performance and succession
Language & Linguistics	Second language acquisition	audio and transcripts of classroom second language learners
	Sociolinguistics	audio and transcripts of interviews with multiple generations of Indian English speakers
Computing and Electronic Systems	Artificial intelligence	crowd sourced AI scripts with results of a competition between these AIs

Two brief descriptions of sample data and ingest follow, to give some flavour of the considerations and processes involved in developing the pilot repository.

Case study 1: Bio-imaging

The University of Essex has a heavily-used bio-imaging facility in the Department of Biological Sciences. Users generate large volumes of image data which must be stored and managed by dedicated staff. A key challenge is the aforementioned volume of data, which is particularly significant when there is a need for multiple versions to accommodate proprietary and open formats. This is the second key challenge of this data case study – how to present multiple representations of a file for download.

The sample dataset provided by the facility manager was a collection of image files underlying a published journal article. Each image file had two representations: raw imaging equipment output (proprietary) and TIFF format image (open). The two formats differ in the way they store image layers and the metadata associated with each of those layers. As a result, both formats would need to be download options.

The ReCollect citation page layout was well suited to presenting the two files types, due to the split column handling of multiple files. In this case, the recommended upload method was to provide two zip archives containing all the images in the two respective formats. Through the document metadata, in addition to the file naming, the difference between the two zip archives could be descriptively recorded. Metadata about the collection methodology and processed methods could be adequately captured during the deposit workflow.

This was a particularly good chance to look at how data may be linked to publication. We concluded that the crucial element in linking in an identifier – ideal a persistent identifier such as a DOI – for both items. EPrints automatically creates URIs for each data collection which would be suitable for the reverse publication-to-data linking.

Case study 2: Sociolinguistics

Linguists typically rely on large data collections that form a continuously used and expanded basis for their research. 'Corpus' culture of this kind results in large volumes of data that require careful curation and management. As a result of this necessity, linguists are perhaps more versed in research data management processes than many other disciplines, although may not intuitively refer to these activities in those terms. Although we assumed data had been anonymised for the purposes of these tests, there is clearly the potential for ethical issues pertaining to data collected from human participants. This will be a challenge for institutional repository administrators to control for, in terms of what is deposited, as the resources for in depth checking are unlikely to be available. Clear guidance and policy statements can help mitigate these risks, as well as ensuring guidance on ethical issues surrounding data is well integrated into the data management planning process.

A typical sociolinguistics dataset is composed of interview audio files, transcripts of the recordings, and transcript annotation files. The sample given for ingest testing was a corpus of interviews with speakers of Indian English. Each interview had multiple representations and interlinked files, including: audio interviews; typed transcripts; XML annotations. In addition, there was considerable amount of documentation such as consent forms and interview guides.

The ReCollect layout worked well here, as we had a clear divide between data and documentation. The many sets of files were presented in zip bundles according to format, in order to fit comfortably within the screen space. The files could be linked by participant ID codes. Descriptive metadata was sufficient to record the necessary information.

User testing

We have had input and feedback from the research data management community as well as active researchers throughout the development process. There were four stages to this testing process:

- Live demonstrations to the pilot department researchers, showcasing their sample data in the repository pilot, and inviting comment and criticism

- 19 pilot guest accounts for JISC repository and research data management practitioners (digital librarian, repository administrators etc.), and the invitation to test upload and browsing
- One-on-one testing with UK Data Archive staff, running through a simple testing protocol covering resource interpretation and deposit
- Open testing to researchers and postgraduate students across the University of Essex, encouraging upload of test data and feedback

Extensive qualitative feedback was the result of these various phases of testing. This feedback informed not only bug fixing, but an incremental process of refinement that is reflected in the ReCollect plugin.

Conclusions and future work

We have developed a comprehensive solution for institutional data storage, built on the back of the widely used open source software EPrints. Our approach is one that we hope will be adopted by institutions new to the research data management challenge, be that on a theoretical or technical level. Regarding the latter, the ReCollect plugin we have made available provides an ideal low-barrier solution for setup and standards compliance in an institutional context. Use of the EPrints platform can be particularly advantageous in getting institutional buy-in if EPrints is already deployed as a publications repository.

There are several gaps in the system (and indeed repository system more widely) that emerged during our work, and we set them out here briefly for the community to note. For practical reasons, individually tagging many files with their respective metadata is not practical. We would like to see new tools/extension enabling more efficient tagging of large numbers of files with metadata, or at least allowing for file level inheritance of metadata from a zip archive. In the meantime, we will not be recommending that depositors upload large numbers of files and individually tag them through the user interface, but rather upload and tag zip files containing logical groupings of files.

There are issues uploading files of roughly 2GB or more in certain browsers. We have found that it is possible to upload these files using certain browser setups, though even this is not consistently successful. This is seems to be to do with the limitation of browser protocols, so may be hard to address through the repository software itself. One option may be to encourage that such large files be hosted elsewhere, and linked to through the EPrints metadata record. There is also the possibility of using other protocols to provide alternate upload methods. Forthcoming improvements to the SWORD 2 protocol's provision for data transfer are likely to be important here. Another occasionally touted solution is the idea of harnessing BitTorrent-like peer to peer sharing.

Finally, we would have liked to have further explored searching and browsing repository contents, and how we might be able to control the searching of plain text and variables within various data and documentation formats. Due to uncertainties over the place of institutional repositories in the discovery of resources, this was not view as a priority for this project; a centralised cross-repository search facility may end up being the primary entry route to these kinds of data collections.

References

1. EPrints (2012) http://www.eprints.org/software/
2. UK Data Archive (2013). Research Data @Essex project page http://www.data-archive.ac.uk/create-manage/projects/rd-essex
3. Ensom, T. & Wolton, A. (2012). *RDE Metadata Profile for EPrints* http://www.data-archive.ac.uk/media/375386/rde_eprints_metadataprofile.pdf
4. ReCollect Bazaar page http://bazaar.eprints.org/280/
5. Brown, M., Parchment, O. & White, W. (2011) *Institutional data management blueprint. University of Southampton*. http://eprints.soton.ac.uk/196241/
6. DataCite Metadata Working Group (2011). *DataCite Metadata Schema for the Publication and Citation of Research Data: Version 2.2.* http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf
7. Association for Geographic Information (2012) *UK GEMINI v2.1*. http://www.agi.org.uk/storage/standards/uk-gemini/GEMINI2.2.pdf
8. DDI Alliance (2011). *DDI Codebook 2.1.* http://www.ddialliance.org/Specification/DDI-Codebook/2.1
9. Gibbs, H. (2009). *DataShare Metadata Schema for ePrints Soton (ePrints 3.1)* http://www.disc-uk.org/docs/ePrints_Soton_Metadata.pdf
10. Higher Education Statistics Agency (2011). *Joint Academic Coding System (JACS) Version 3.0* http://www.hesa.ac.uk/content/view/1776/649/
11. Library of Congress (2013). *Library of Congress Classification Outline* http://id.loc.gov/authorities/subjects.html
12. Ensom, T & Wolton, A. (2013). *EPrints Wiki: ReCollect* http://wiki.eprints.org/w/ReCollect.