# CASE STUDY

A Digital Curation Centre Case Study
August 2015

# Using EPrints to Build a Research Data Repository for UEL

**Stephen Grace (University of East London), Angus Whyte and Jonathan Rans (DCC)**

## Introduction

This case study describes the early incarnation of the University of East London's (UEL) research data repository, which at time of writing was based on the EPrints platform (v.3.3.12) and associated plug-ins. The study draws mainly on an interview with Stephen Grace, Research Services Librarian at the University, and related publications.

The case study is one of three accompanying the DCC guide *How to Evaluate Research Data Repository and Catalogue Platforms.*([1]) Each looks at these in different contexts and portrays issues affecting platform choice and implementation decisions. Another case study, co-produced with Jisc, offers a broader overview of the steps UEL took to meet the Engineering and Physical Sciences Research Council's (EPSRC) expectation that universities establish research data management (RDM) policies.

## Background and context

The university is a post-92 institution ([2]) , based on three campuses in London's Stratford and Docklands areas. UEL's research strategy aims to embed it as one of the UK's leading modern research universities. Library and Learning Services (LLS) support for that strategy has driven the data repository initiative, along with institutional commitment to meet EPSRC compliance requirements.

In January 2012, an LLS-led project was initiated to identify requirements for research data management training and infrastructure across disciplines, and at different stages of the research lifecycle. Two months later the project identified two main infrastructure requirements for the support service

- Data repository for sharing data held by the university.
- Data catalogue/ register to provide metadata records for all data including that held externally.

Key issues:

- EPSRC mandate for data publishing & preservation
- Outsourcing
- Development capacity
- Persistent Identification
- Domain & context metadata
- Information security policy control
- User licensing flexibility
- Version control
- Storage management

[1] *Forthcoming in 2015*

[2] *The term refers to 'new universities' in the UK, specifically those that gained university status and degree-awarding powers through legislation introduced in 1992 see Wikipedia entry New universities at: https://en.wikipedia.org/wiki/New_universities (retrieved 2/7/15)*

The university Research Committee provides general oversight for RDM. It includes representation from senior academics and services, including IT, and Research and Development Support, as well as LLS.

The choice of platform was heavily influenced by previous experience with setting up a publications repository. UEL's publications repository ROAR (roar.uel.ac.uk) runs on the EPrints platform. This replaced a DSpace installation that had been established in response to the IT policy of the time, which favoured the Oracle database underlying that platform. That policy is in remission, and UEL now sees EPrints as more aligned with the UK higher education sector's needs.

UEL also opted for a hosting service to run their EPrints repository, using the University of London Computer Centre's (ULCC) offering. This allowed them to incorporate customisations to the basic installation, based on their own specifications, while minimising the need for in-house development capacity.

The publications repository is defined by institutional policy as an Open Access (OA) repository. This presented no barrier to using it as a research data catalogue, and currently externally held metadata records are being added to the repository. However it was recognised that some research datasets would require more restricted access, ruling out the OA repository.

With the hosting arrangement newly established, and with ULCC also offering the Recollect plug-in for EPrints, this became the natural choice for a data repository. The repository requirements were provided to ULCC in the form of a spreadsheet, and screen mock-ups were used to refer to these during the repository design and implementation, as illustrated in Figure 1.

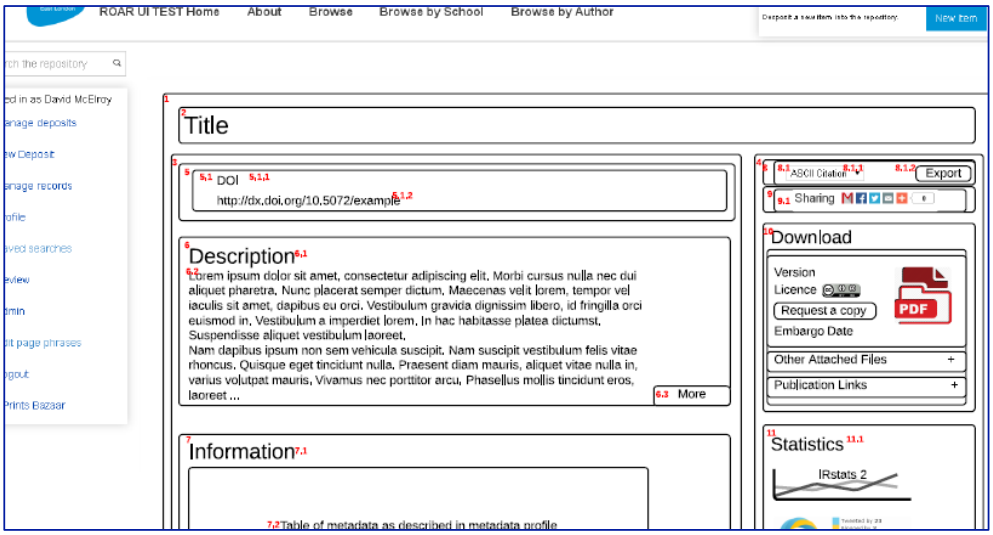


Figure 1. Screen mock-up cross-referencing numbered requirements statements.

Stephen had built a good working relationship with the developers at ULCC, having worked closely with them to identify the right set of specifications for the publications repository, so he was confident that he would benefit from expert guidance and advice. The working knowledge of EPrints across central services also minimised the impact introducing a new system was likely to have on day-to-day business.

## Key requirements and current challenges

The UEL data repository operates under the name 'Data.uel' and incorporates a number of EPrints plug-ins to ensure the installation meets the key requirements identified. These include:

- Repository linking - ability to link across repositories between data packages or 'items' and publications;
- Containers - allowing items to be grouped together and nested;
- Recollect - supporting emerging best practice for mandatory and optional metadata;
- Datacite - plug-in to enable minting of DOIs for data items;
- Access Control Layer (when available) - forthcoming EPrints functionality developed by University of Leeds to provide role-based access to data items;
- Arkivum A-Stor storage integration - to allow data items to be stored and retrieved from the bit-level archiving service provided by Arkivum.

## Linking, identifying and grouping data for discoverability

An important factor in selecting EPrints was the potential to effectively link the publication and data repositories, to connect publications with underlying data. The platform does not support this 'out of the box' however; the functionality was added through a plug-in written to UEL requirements by the ULCC developers over the course of the repository implementation project. The RepoLink plug-in has now been made available to all EPrints users via the EPrints bazaar (bazaar.eprints.org).

UEL subscribes to DataCite so that it can publish metadata for data deposited in the Data.uel repository. The repository deploys the DataCite plug-in to support DOI minting, and in line with DataCite expectations, the mandatory metadata is included on a landing page for each data item.

UEL also wanted to make relationships between data items discoverable, for example enabling these to be grouped within projects. An EPrints 'Containers' plug-in provides that functionality. This was developed through the Jisc supported project *Kultivate*, to address the levels of complexity in artistic research outputs and workflows.([3])

## Supporting metadata description

UEL wanted to take advantage of emerging best practice in 'core' metadata for research datasets, responding to the EPSRC expectation that institutions 'ensure that appropriately structured metadata describing the research data they hold is published and made freely accessible on the internet…'. The EPSRC acknowledges DataCite DOIs in its policy, but makes no specific recommendation to use a particular standard. For UEL this meant having room to manoeuvre, enabling the repository to gather best-practice descriptive terms beyond the five mandatory fields that Datacite stipulates.

The Recollect plug-in offers a more extensive metadata schema, developed by the Jisc-funded *RD@Essex* project and taken further by the EPrints user group. This schema has a larger number of mandatory fields than UEL wanted to mandate. UEL are considering only a few fields beyond the mandatory Datacite ones, with Recollect fields as an option for more enhanced description. Unlike the UEL publications repository, which has a process of mediated deposit, the data repository will be relying on self-archiving (including assisted deposit), so it will need to present as low a barrier to researchers' deposit as possible. The data repository is also discipline-agnostic, so UEL wanted to avoid specifying any domain-specific documentation requirements.

## Dealing with access restriction

Datasets that are subject to access restrictions will still typically be housed in the UEL's data repository, whether for a temporary, pre-publication embargo period or for restricted long-term access on a restricted basis. EPrints offers access restrictions suitable for time-limited embargoes, and access to a data package may be limited to authorised users of the repository. However UEL need to provide restricted access to specific data packages, limited to specific identified roles and users. This more fine-grained control may eventually be provided through an Access Control Layer (ACL) for EPrints, being developed by Leeds University.

Currently secure data archiving is provided through the Jisc framework agreement with third-party provider Arkivum, whose service conforms to ISO 27001 standards for data security. The ULCC-hosted data repository uses Arkivum for storage, and archived data can be retrieved through the repository interface, but UEL has additional Arkivum storage that is not currently integrated with the repository. This may be used to archive data securely and to offer mediated access.

Keeping track of data access requests is essential for meeting the EPSRC requirements on data retention. EPrints tracks online access to items. UEL also employs the same ticketing system used by their Information Services helpdesk to keep track of data access requests from individuals, and to maintain an audit trail for released datasets.

## Licensing and stewardship

The repository offers a broad range of licences that can be applied on deposit including all available Creative Commons licences and the option to apply user-defined licences to particular datasets. This is intended to accommodate the need to describe unusual access or use conditions, and it is hoped that this will encourage broader use of the service.

[3] *Gramdstadt, Marie-Therese (2012) Kultivating Kultar: Increasing Arts Research Deposit - Retrieved 6 July 2015 from: http://www.ariadne.ac.uk/issue68/gramstadt*

All datasets will have a responsible party assigned to them at the point of deposit, if not sooner. There is no restriction on who that might be, but the expectation is that the project's PI would be the most likely candidate. Naturally, there also needs to be someone with ultimate authority within the institution should the original data owner be unable to make necessary decisions. There is a nominated position within each School that holds this responsibility; commonly this falls to the Associate Dean.

### Version control

At the moment, there is no real versions control in EPrints running the Recollect plug-in. Changes are possible although this is only really appropriate for relatively small edits that have no impact on the structure or interpretation of the data, for example, minor spelling changes to table headings. Should a substantive change to a deposited dataset be required, UEL's approach is in line with DataCite expectations that a new version of the dataset metadata must be created, with its own newly minted DOI.

### Data storage management and preservation

In addition to the open access repository, UEL will be providing researchers with data archiving infrastructure via Arkivum, a third party provider of tape archiving through Jisc's data archiving framework agreement. UEL already had an Arkivum appliance in place, before developing their data repository, to enable automatic archiving of data from UEL's network.The archival infrastructure offers a possible solution for dealing with datasets that are too large to consider storing and delivering through the open access repository.

## Further development

Initially, IT services were not involved in the RDM project as this coincided with the University's support for the 2012 London Olympics. Now that pressure is past, LLS is working with IT services to take the infrastructure further. Their involvement is expected to ramp up UEL's offering for both active data storage and archival storage.

The on-site Arkivum service could provide a secure dark-storage solution for any data unsuitable for an open-access repository, or which does not need to be kept on expensive, spinning-disc storage. The main outstanding issue with using this service to store research datasets is that the ULCC-hosted repository will be using the same service from its site. Having two modes of interaction with the Arkivum service has somewhat complicated the management of this. UEL have not finalised the processes that they will use to manage the access and availability policies applying to data held on their on-site Arkivum service, but the aim is to integrate these with the workflow for repository deposit.

Currently the UEL repository's support for preservation is limited to bit-wise checks for data integrity, however a library-based working party is looking into what further levels of preservation should be provided. There is a wariness of making commitments to support format migration or emulation without further knowledge of the scale of the requirement, including the range of formats that will need to be supported.

## Lessons learned

- Leveraging the skills and experience gained from existing in-house repository platforms is likely to be a major factor in the selection of data repository software.
- Outsourcing to a repository hosted service may be a way to access expert guidance.
- Other factors aside, there may be policy reasons not to use a publications repository for datasets, as the constraints on offering open access to data are greater than for publications, making it difficult for a single repository to apply an open access policy consistently.
- Don't necessarily try to solve all issues around the repository before implementation; it is likely that some functionality will need fine-tuning as researchers begin to deposit data.
- Clearly identify which elements of support for the repository, if any, should be paid for through direct grant recharging.

**Please cite as:** Grace, S., Whyte, A. and Rans, J. (2015). 'Building UEL's Research Data Repository Capabilities with EPrints'. *DCC Publications*, Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/case-studies