Data repositories and storage: developments at Essex

Thomas Ensom
UK Data Archive, University of Essex

@RDEssex
Research Data @Essex is a JISC-funded project aiming to develop a sustainable research data management and sharing infrastructure, built on best practice guidance from the research data management community and UK Data Archive expertise.

The University has an EPrints institutional repository, and an important part of the project is setting up a data instance building on the same implementation. I will be talking about our approach to adapting it to better suit collections of data.
Research data differs greatly from research publications, in level of complexity. An article is typically a single file, while a dataset or data collection could (conceivably) be hundreds of files with multiple relationships between them. So let's define our terms in an the EPrints context.

A data collection is our ‘eprint’, the key unit. This could be anything from a set of audio interviews with transcripts, to a single spreadsheet. Within each collection there is a set of descriptive metadata, and a series of files. These files can be of the types: data, documentation and metadata. Data collections can be grouped inside larger containers. For example, a series of datasets produced as part of an umbrella project. We are trying to decide whether these higher level groupings should be formal or user instigated.
Key changes

- Extended the default EPrints metadata profile to better suit research data
- Based on existing schema to enable interoperability
- Changes to rendering of a ‘data collection’
- A number of challenges presenting our metadata (presenting a 50% increase in same space!)
  - e.g. How to clearly present collection description and file listing
  - e.g. Separate a potentially large number of data, documentation and metadata files

We have developed a metadata profile built on the DataCite schema - we intend to mint DataCite DOIs further down the road). To improve descriptive richness, and meet relevant standards, we also examined several other schema and expanded our profile:
- INSPIRE – for geospatial data, but also providing a neat generic description of research data
- DDI - a metadata schema originally from the social science community, but now finding applications in biomedical research and beyond due to it’s descriptive power
- DataShare – work done at Edinburgh University for sharing research datasets, this was primarily a source of inspiration for controlled vocabularies and form validations.
This is the current RD@Essex citation screen on our test server – it’s a work in progress.

Looks a lot like base EPrints, but has key differences
We wanted an organised, tidy screen, but without sacrificing any detail.
We’ve added two extra components to do the work, hooking into default EPrints javascript to control the amount of metadata onscreen at any one time
We wanted to work with what has already been done so well by the EPrints team, but adding the necessary detail our system captures.
2. Core metadata

Remains mostly the same as with a base EPrints install

Visible here is the new Data Collection item type we’re using
3. Metadata detail

Rendered as a collapsed box by default
Unrolled shows the complete metadata record
Metadata fully unrolled
Shows the extent of metadata we’ve added!
4. Documents associated with each Eprint=Data Collection

We wanted to sort uploaded files according to a type: Data, Documentation, Readme, Additional Metadata or Archive (i.e. the whole lot)

Quite a bit of debate as to the best way to sort the files – inspired by the ecystals (a discipline specific EPrints repository) layout we tried initially to order by file extension, quickly realised this wasn’t going to work as different content types (data, documentation etc.) could have the same mime type e.g. .doc, .pdf .xls
5. Documents associated with each EPrint/ Data Collection extended

We wanted file level metadata to be viewable, but not immediately so again we’ve used collapsible boxes to keep the screen tidy.
The proposal is a step towards addressing the issue, by further examining the tempting view and exploring the ways in which intuitive concepts can be computationally characterized.

<table>
<thead>
<tr>
<th>Item Type:</th>
<th>Data Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title:</td>
<td>A simplicity model of concept difficulty</td>
</tr>
<tr>
<td>Uncontrolled Keywords:</td>
<td>simplicity modeling, categorization, psychological stimuli, concept learning</td>
</tr>
<tr>
<td>Subjects:</td>
<td>Philosophy, Psychology, Religion &gt; BF Psychology</td>
</tr>
<tr>
<td>Divisions:</td>
<td>Faculty of Medicine, Health and Life Sciences &gt; School of Psychology</td>
</tr>
<tr>
<td>Depositing User:</td>
<td>RD Admin</td>
</tr>
<tr>
<td>Date Deposited:</td>
<td>14 Jun 2012 10:22</td>
</tr>
<tr>
<td>URI:</td>
<td><a href="http://eprints.essex.ac.uk/id/eprint/734">http://eprints.essex.ac.uk/id/eprint/734</a></td>
</tr>
</tbody>
</table>

### Available Files

#### Archive

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>archive-38.zip</td>
<td></td>
</tr>
</tbody>
</table>

#### Documentation

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>end_of_grant_report.doc</td>
<td></td>
</tr>
</tbody>
</table>

#### Content: Documentation

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>dataset_guide.doc</td>
<td></td>
</tr>
</tbody>
</table>

#### Data

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment_1_results.xls</td>
<td></td>
</tr>
<tr>
<td>Exp_2_results.xls</td>
<td></td>
</tr>
</tbody>
</table>

#### Visible to:

- Anyone

#### Content:

- Data

#### Description: Experiment 2

- Other

#### Metadata

- Revision: 5

#### Mime Type:

- application/vnd.ms-excel

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment_3_results.xls</td>
<td></td>
</tr>
</tbody>
</table>
Challenges

- Dealing with complex collections
  - Very large file sizes
  - Multiple versions of the same file(s)
  - Inter-dependent files e.g. GIS database

- Standardising the ‘pre-repository stage’ e.g. collecting metadata, naming files

- Looking forward to SWORD2 for data

- Researchers do not necessarily think like repository designers! + vice versa

We’ve had problems uploading large files during testing – it tends to fall over. Could be the same with download? There are still practical problems uploading and adding metadata to very complex collections e.g. many files. How also, to ensure inter-dependent files such as those making up a GIS database, maintain their essential file/folder structure while still being adequately described in file level metadata. Current approach to both of these problems is to recommend upload of problem data in zip files.

Looking forward to new tools that will help manage the pre-repository stages of the data lifecycle, enabling collection of metadata at this stage which can then be passed to EPrints.

Final note – community needs to work with researchers in their institutions to get repositories accepted and integrated. We found researcher feedback invaluable and illuminating. Not necessarily easy to engage though.