Data repositories and storage: developments at Essex

Thomas Ensom
UK Data Archive, University of Essex

@RDEssex

University of Essex

UK • DATA ARCHIVE

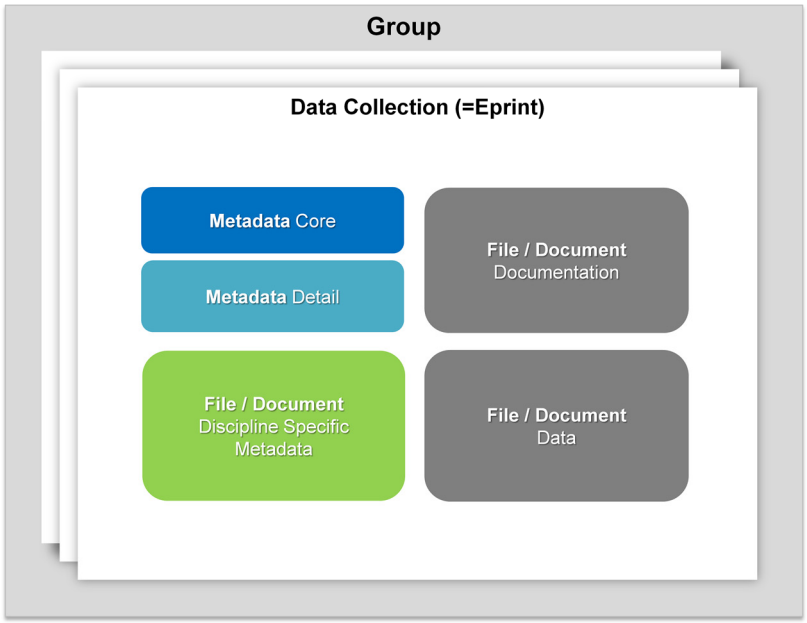
Background

- A JISC MRD project developing RDM infrastructure at the University of Essex
- EPrints is a widely used, open source institutional repository solution
- Geared toward article type deposits
- Development work at Essex to adapt for data

Research Data @Essex is a JISC-funded project aiming to develop a sustainable research data management and sharing infrastructure, built on best practise 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.

Groups, data collections & files



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 lets define our terms in an the EPrints context.

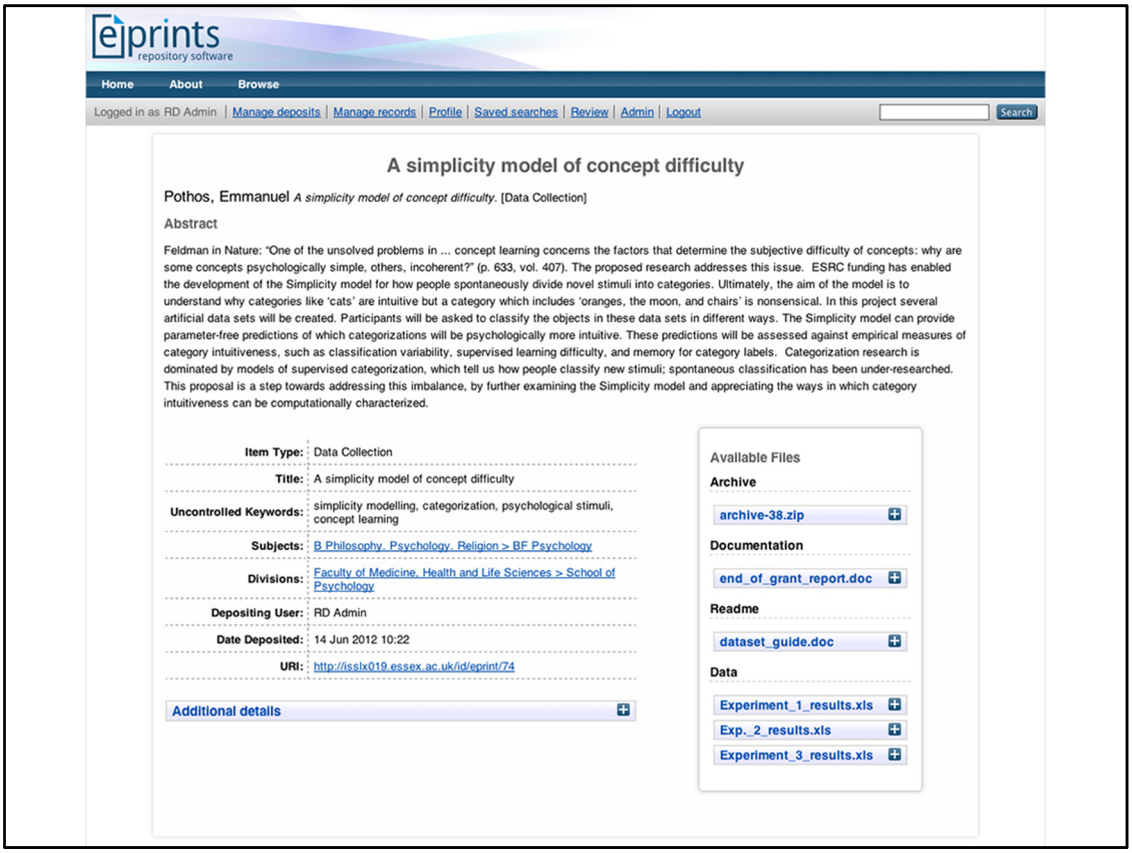
A data collection is our ‘eprint’, the key unit. This could be a anything from a set of audio interviews with transcripts, to a single spreadsheet. Within each collection there is a set of descriptive metadata, and a series of files. These files can be of the types: data, documentation and metadata. 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

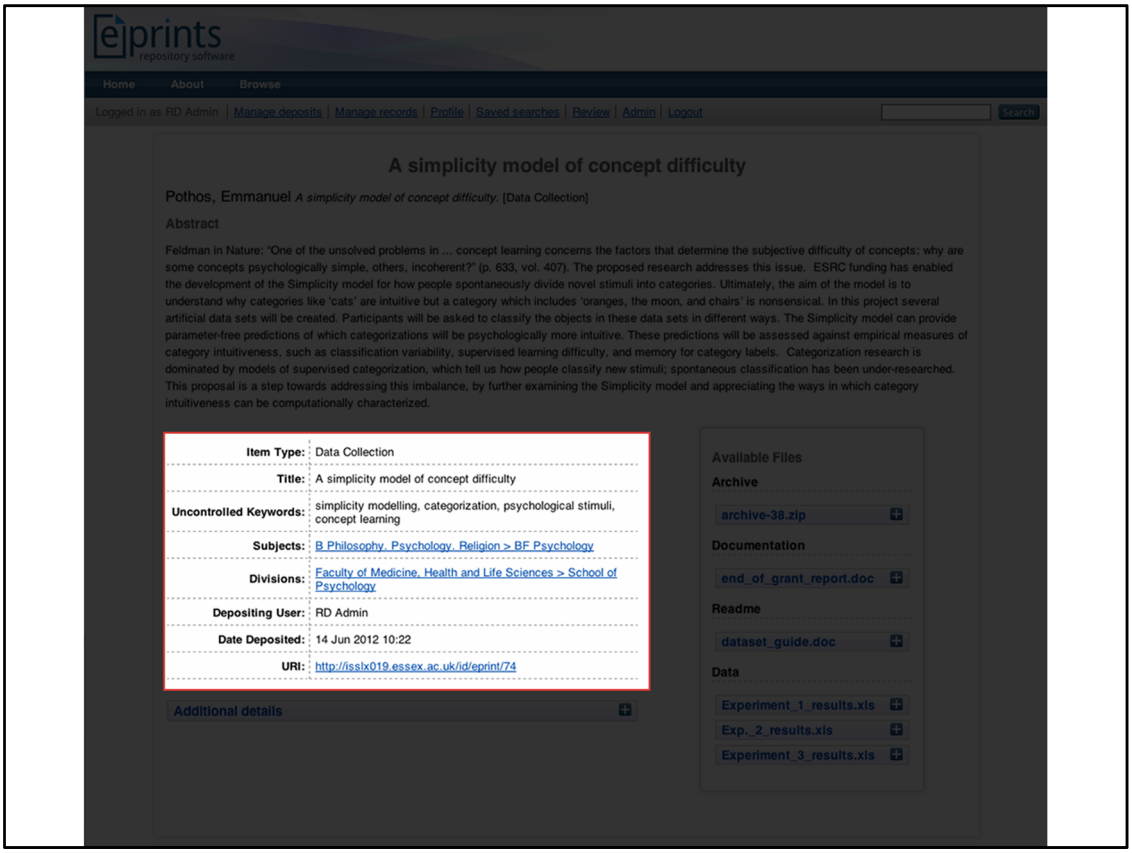
- Changes to rendering of a ‘data collection’
- A number of challenges presenting our metadata (presenting a 50% increase in same space!)
  - e.g. How to clearly present collection description and file listing
  - e.g. Separate a potentially large number of data, documentation and metadata files

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.



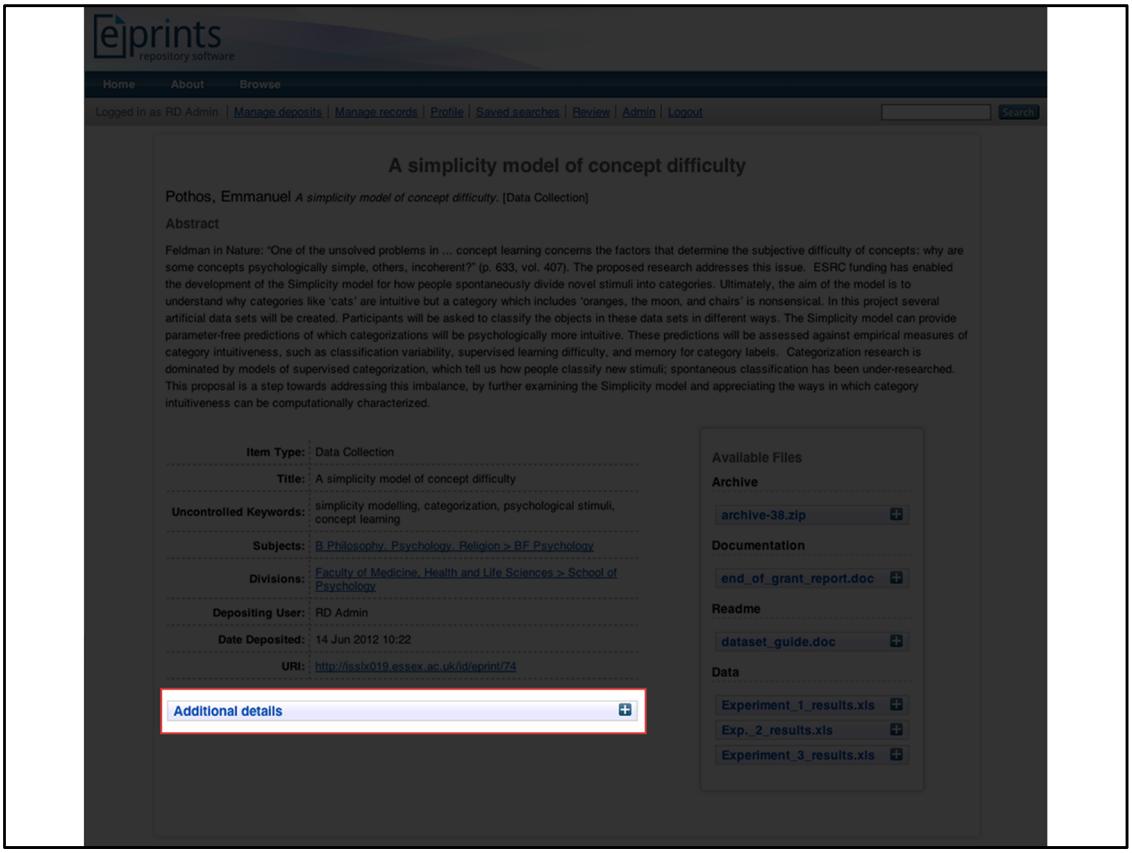
This is the current RD@Essex citation screen on our test server – it's a work in progress.

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.



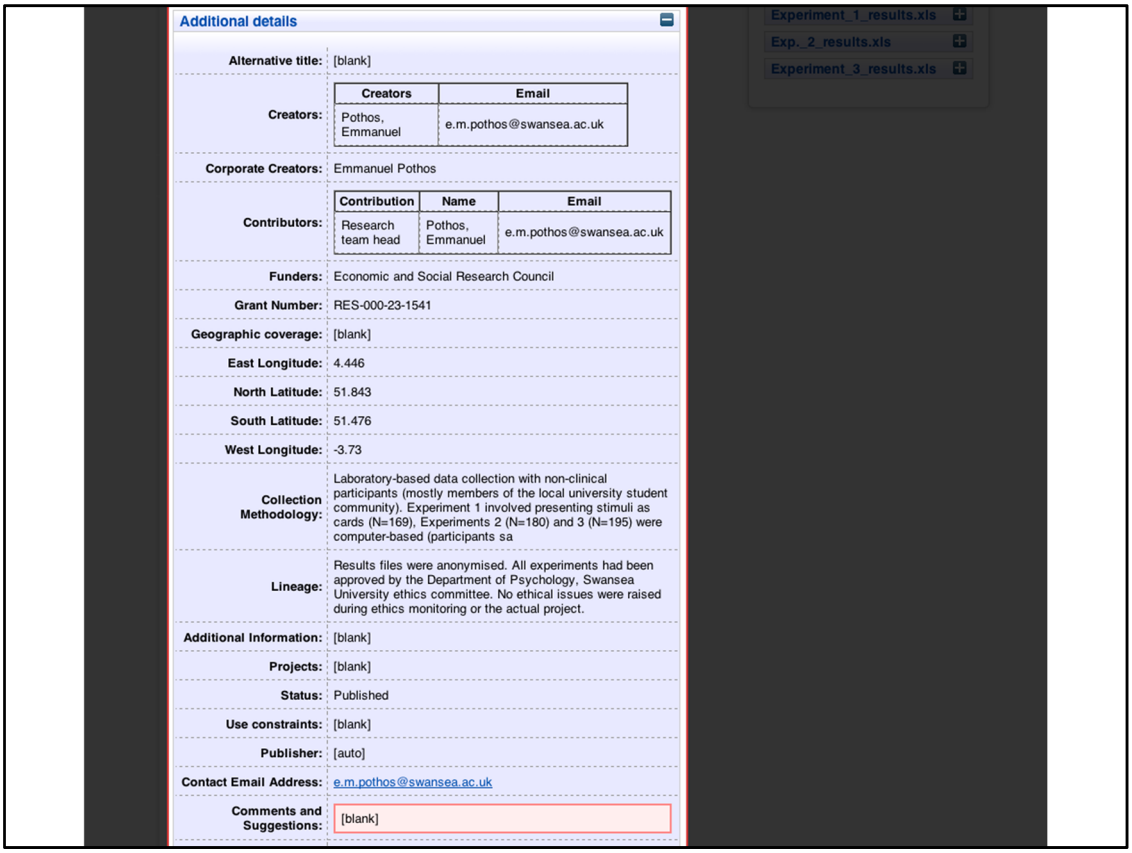
2. Core metadata

Remains mostly the same as with a base EPrints install
Visible here is the new Data Collection item type we're using

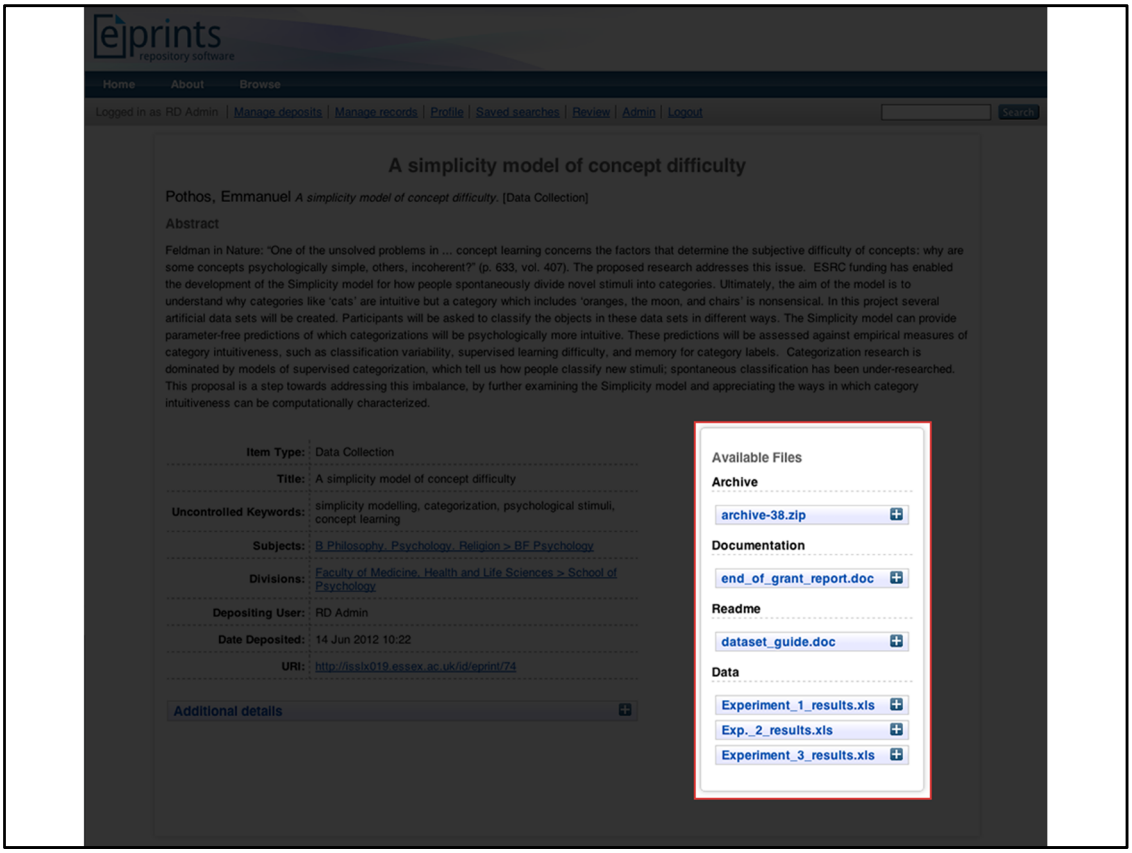


3. Metadata detail

Rendered as a collapsed box by default
Unrolled shows the complete metadata record



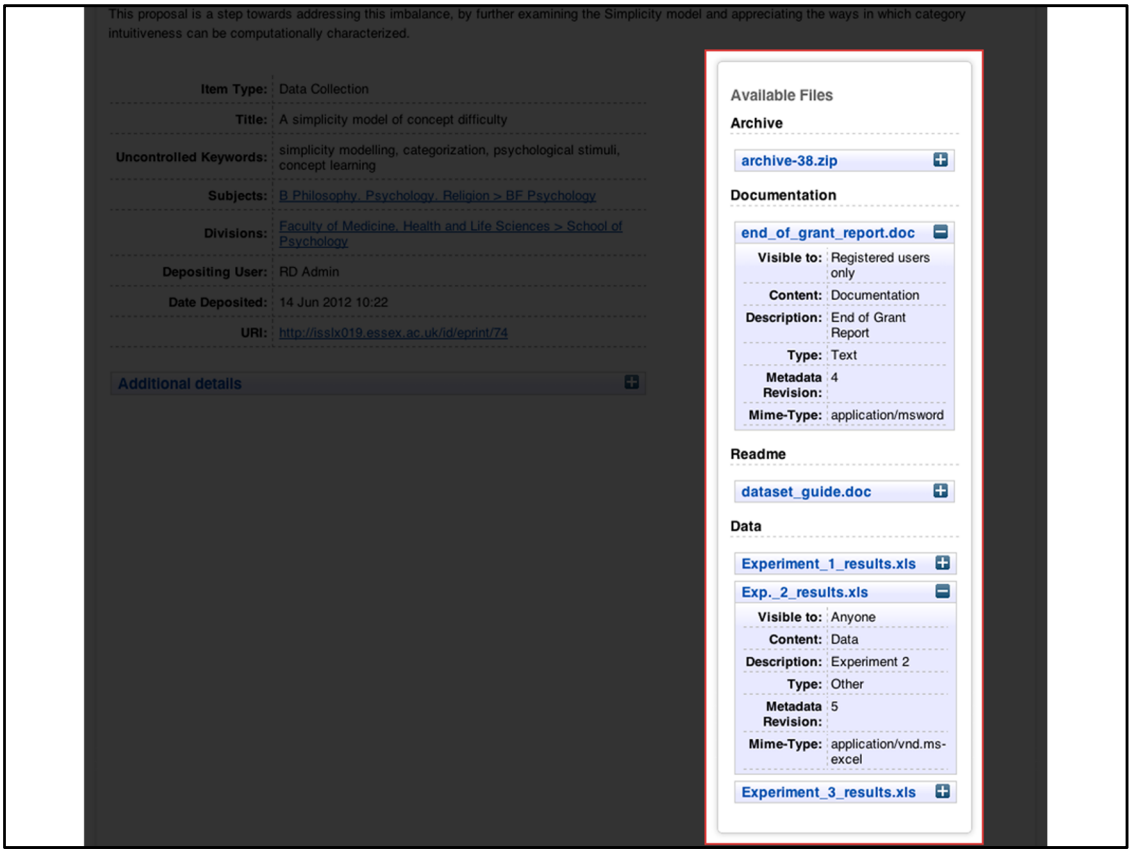
Metadata fully unrolled
Shows the extent of metadata we've added!



4. Documents associated with each Eprint=Data Collection

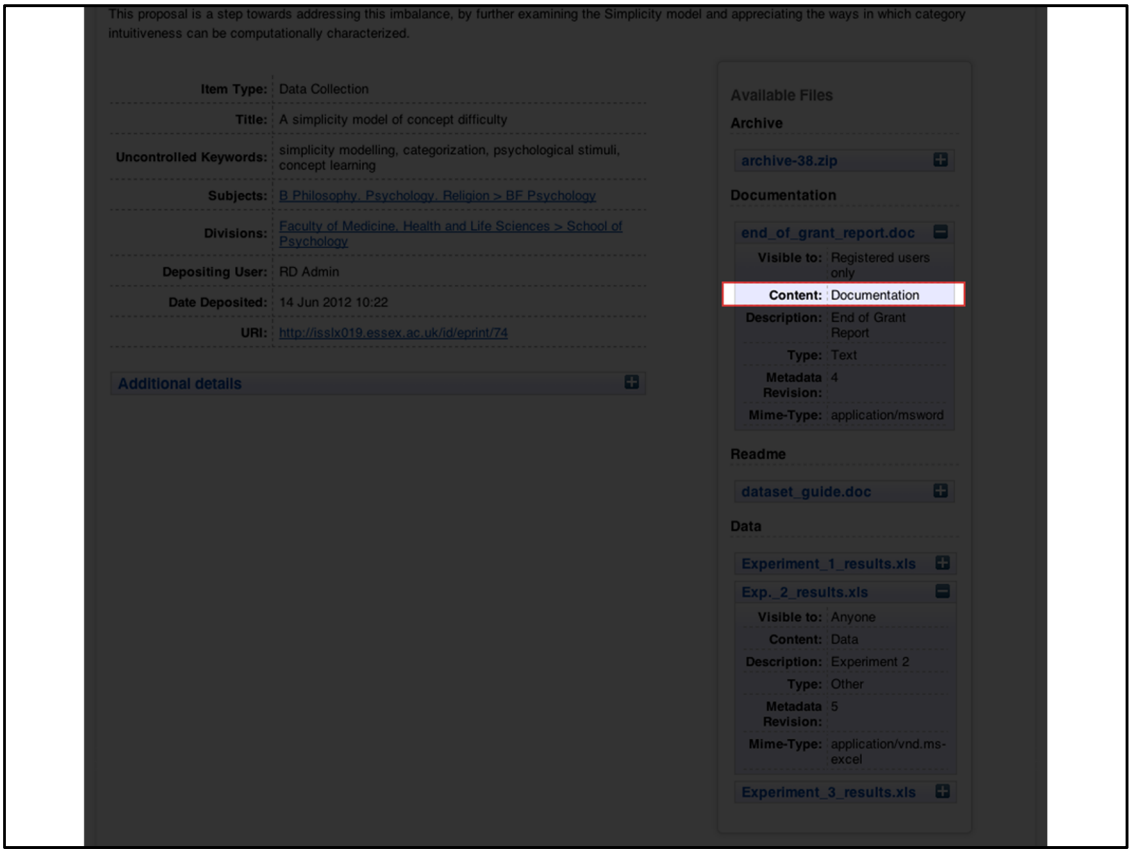
We wanted to sort uploaded files according to a type: Data, Documentation, Readme, Additional Metadata or Archive (i.e. the whole lot)

Quite a bit of debate as to the best way to sort the files – inspired by the ecrystals (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



This proposal is a step towards addressing this imbalance, by further examining the Simplicity model and appreciating the ways in which category intuitiveness can be computationally characterized.

Item Type:	Data Collection
Title:	A simplicity model of concept difficulty
Uncontrolled Keywords:	simplicity modelling, categorization, psychological stimuli, concept learning
Subjects:	B Philosophy. Psychology. Religion > BF Psychology
Divisions:	Faculty of Medicine, Health and Life Sciences > School of Psychology
Depositing User:	RD Admin
Date Deposited:	14 Jun 2012 10:22
URI:	http://isslx019.essex.ac.uk/id/eprint/74

Additional details

Available Files

Archive

archive-38.zip

Documentation

end_of_grant_report.doc

Visible to:	Registered users only
Content:	Documentation
Description:	End of Grant Report
Type:	Text
Metadata Revision:	4
Mime-Type:	application/msword

Readme

dataset_guide.doc

Data

Experiment_1_results.xls

Exp._2_results.xls

Visible to:	Anyone
Content:	Data
Description:	Experiment 2
Type:	Other
Metadata Revision:	5
Mime-Type:	application/vnd.ms-excel

Experiment_3_results.xls

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.