Opening up research data at Essex: experiments with EPrints

Thomas Ensom & Alexis Wolton

University of Essex

@RDEssex

University of Essex

UK · DATA ARCHIVE

Today we're going to be reporting on our work developing a pilot EPrints data repository at the University of Essex

Background

JISC

- JISC-funded Managing Research Data programme project
- Developing MRD infrastructure and policy at the University of Essex
- Utilising UK Data Archive expertise in data management

- University of Essex has an EPrints IR, we have piloted a separate instance for data
- How best to adapt it 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. Today we will be talking about our approach to adapting it to better suit collections of data.

Accommodating diverse data

- We spoke to researchers from four pilot departments:
  - **Language and Linguistics**
  - **Biological Sciences**
  - **Computing and Electronic Systems**
  - **Business School**
- Interviews, inventory and sample data gathered for testing
- On-going contact and consultation with researchers throughout development

We have been working with four pilot departments, covering a broad range of disciplines Essex.

This has involved interviewing researchers and asking for sample data collections to trial ingest into the test-bed repository. We have continued to work with these researchers as the project has continued.

Design ethos

- Minimising barriers for researchers to deposit
- …while satisfying requirements for re-use (i.e. sufficient metadata and documentation)

- Yes, we want deposit to be as easy as possible
  - But we want the data to be more than just a ticked box
  - But filling repository with rubbish is pointless

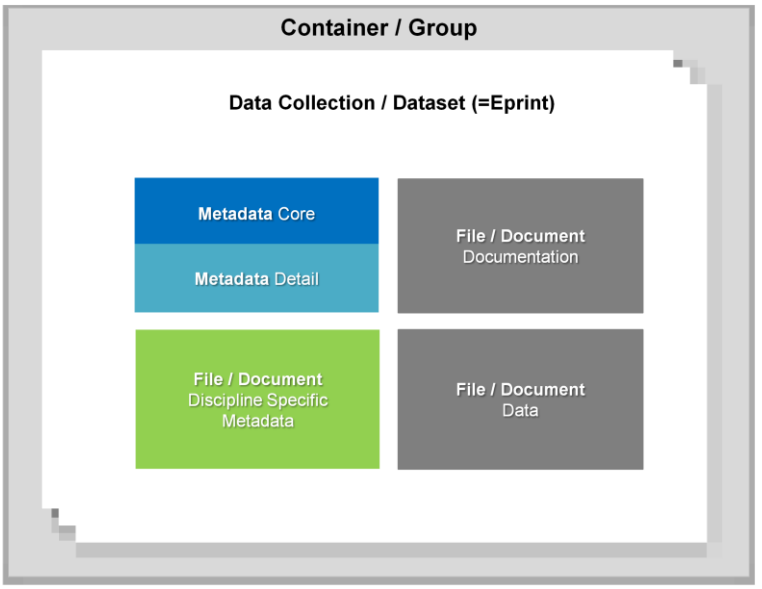
How should we compromise?

Our design maxim has been to minimise barriers while enabling re-use.

Yes, we do want the deposit process to be as straightforward as possible for the user, but we also want publishing data to be more than just a tick in a box – we need rich metadata and as much documentation as possible.

We realise though, that asking for too much you might end up with a load of junk. Can we find a compromise?

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.

Metadata

- Extended the default EPrints metadata profile to better suit research data
- Based on existing schema to enable interoperability

- Minimal mandatory elements based on DataCite metadata, to enable DOI minting further down the road
- Rich metadata based on
  - DataShare, for Edinburgh digital repository
  - INSPIRE, and EU standard for data with geospatial content
  - DDI (Data Documentation Initiative), from the social science community but now being by others e.g. to describe biomedical data

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, we also examined several other schema including:
DataShare – work done at Edinburgh University for sharing research datasets
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 due to it's depth and power

This will be published within the next few months

Implementation

- The next phase of the project looked at rendering a Data Collection in EPrints
- We knew how we wanted to describe our data, but we faced a number of challenges turning our metadata into a useable EPrints screen:
  - how to display a metadata schema that had been extended by nearly 50%
  - How to clearly present the 3 tiers of our data collection
  - how best to group together and display the different files that make up the collection.
- We looked around at what others were doing / had done
  - Ecrystals – orders files according to type
  - Kulture – uses 'Containers' that inherit metadata

We've talked about the requirements gathering exercise and the metadata profiles that were been generated as a consequence.

The next phase was to render a Data Collection in Eprints

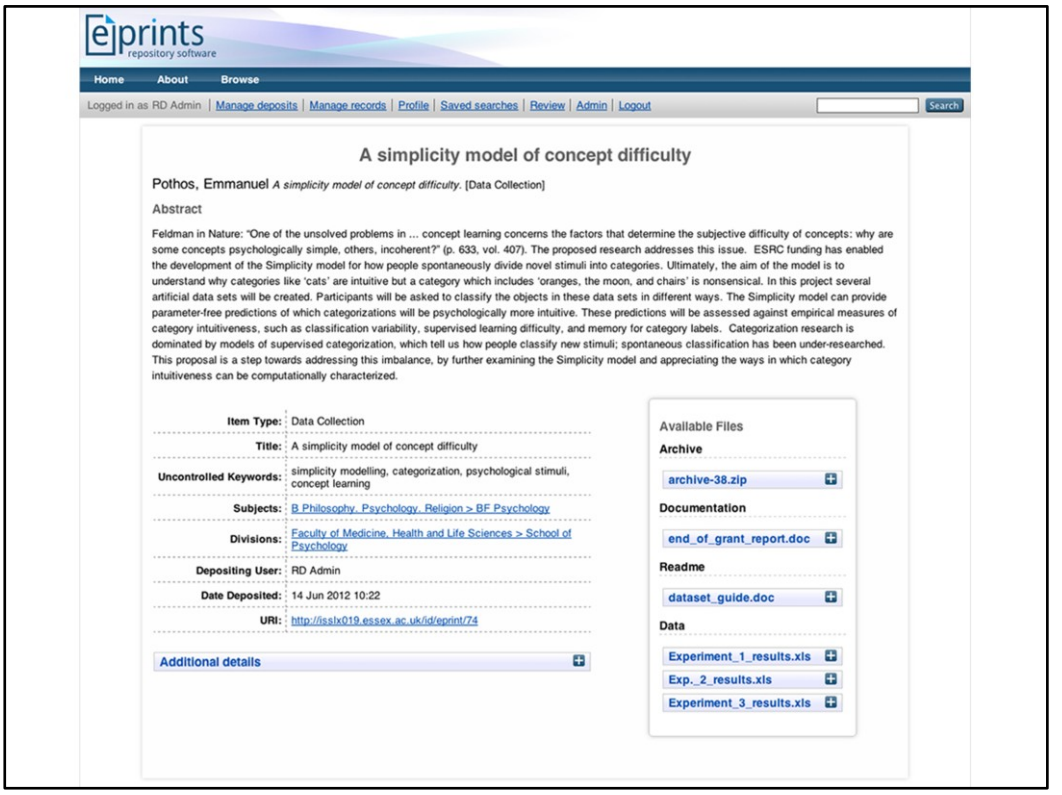
We were faced with a number of challenges

We knew how we wanted to describe our data, but we faced a number of challenges turning our metadata into a useable Eprints screen

- how to display a metadata schema that had been extended by nearly 50%
- How to clearly present the 3 tiers of our data collection
- how could we group together and display the different files that make up the collection.

We looked around at what others were doing / had done in the past.

We were particularly impressed by the way Ecrystals ordered files according to file type,
And were interested in the way Kulture was using 'Containers' that inherited metadata



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

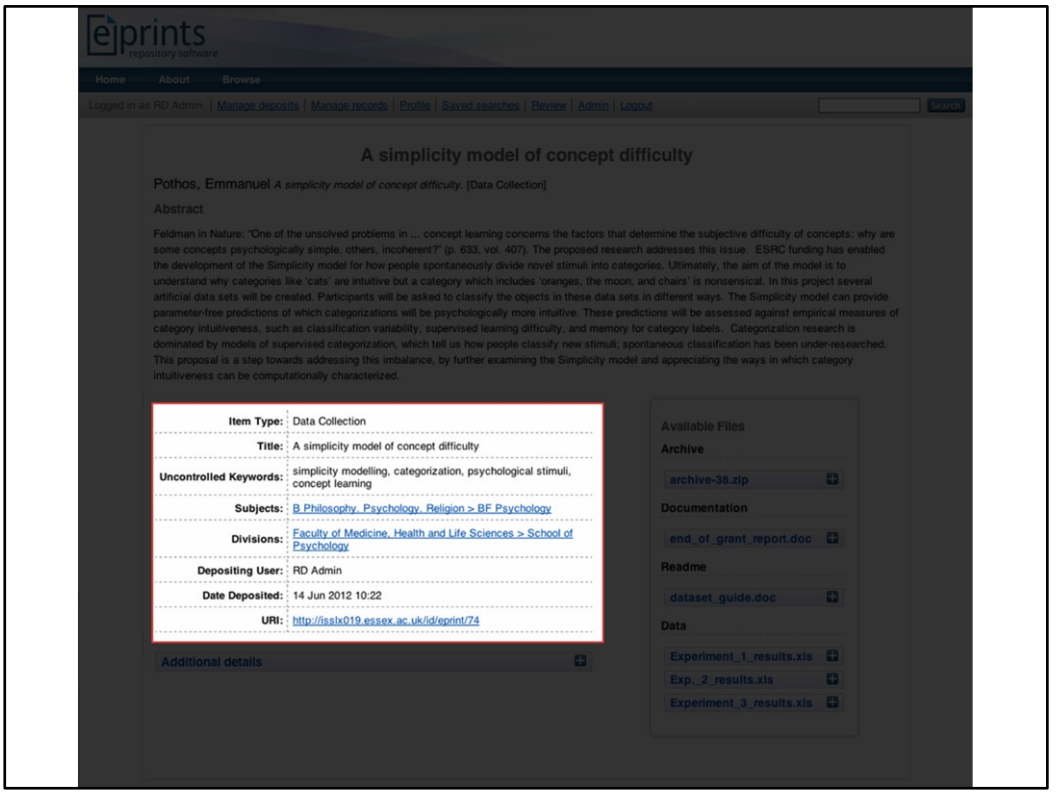
Looks a lot like base Eprints, but has key differences

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, building on top of the solid base that already exists, but adding Research Data specific elementsto it.

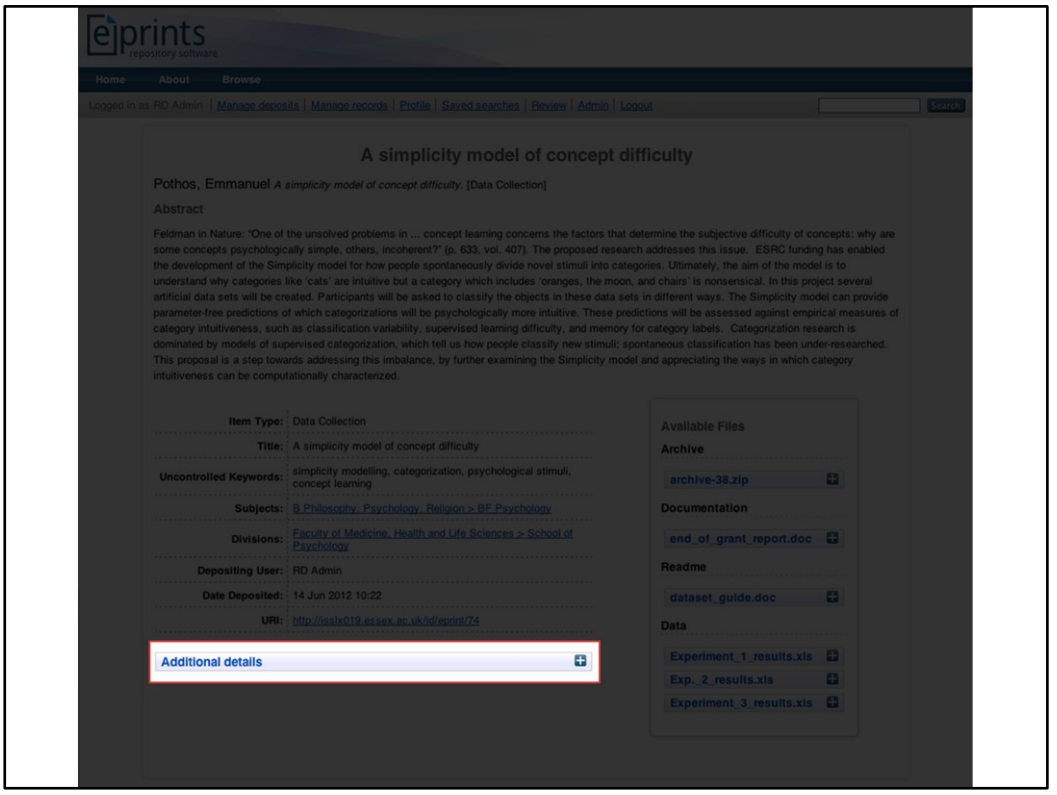
The different elements of the screen:



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:

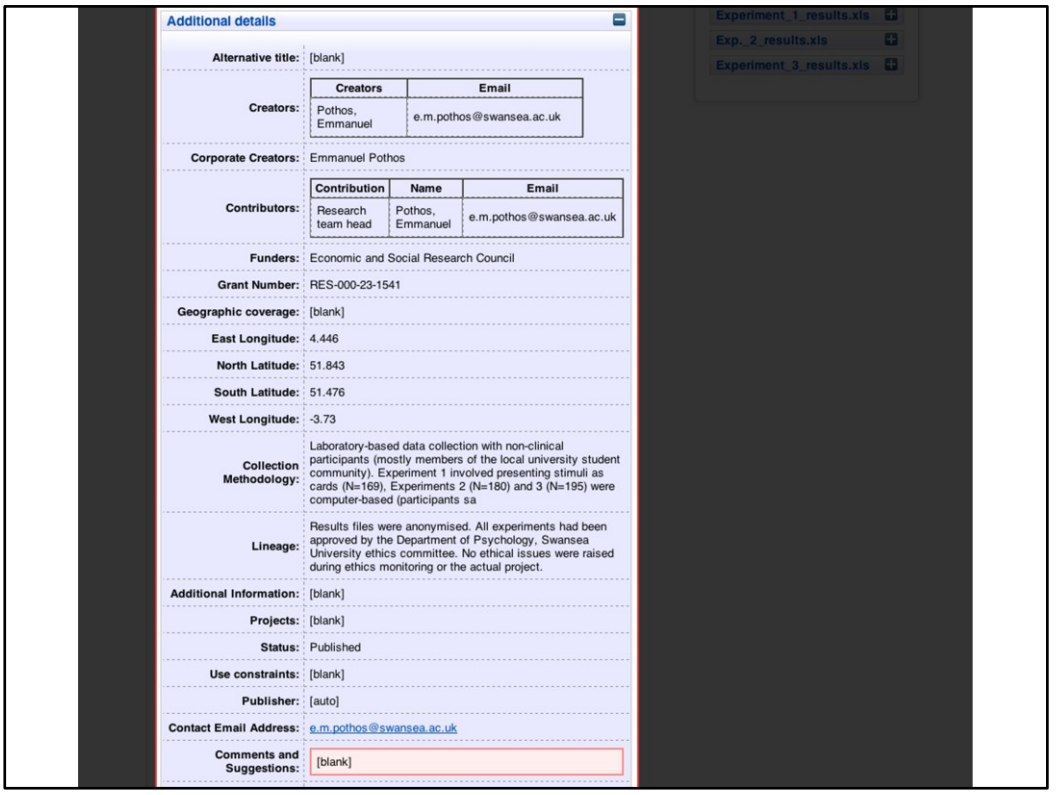
Data Collection



3. Metadata detail

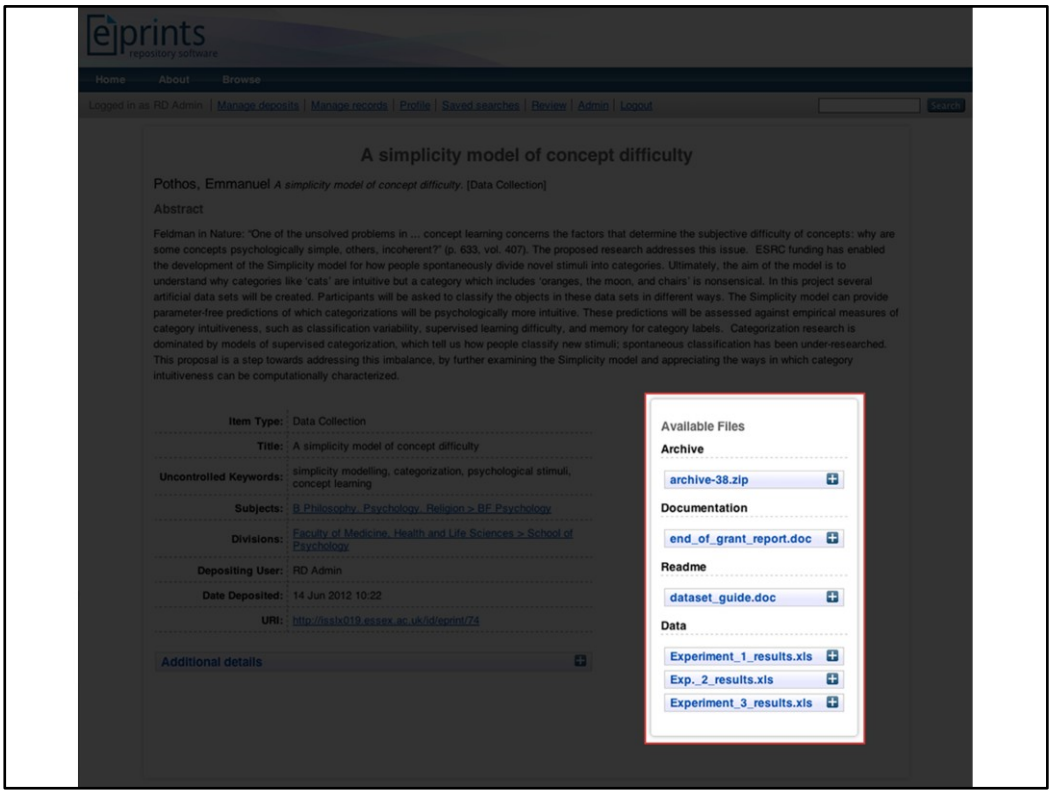
Rendered as a collapsed box by default

Unrolled forms the complete metadata record



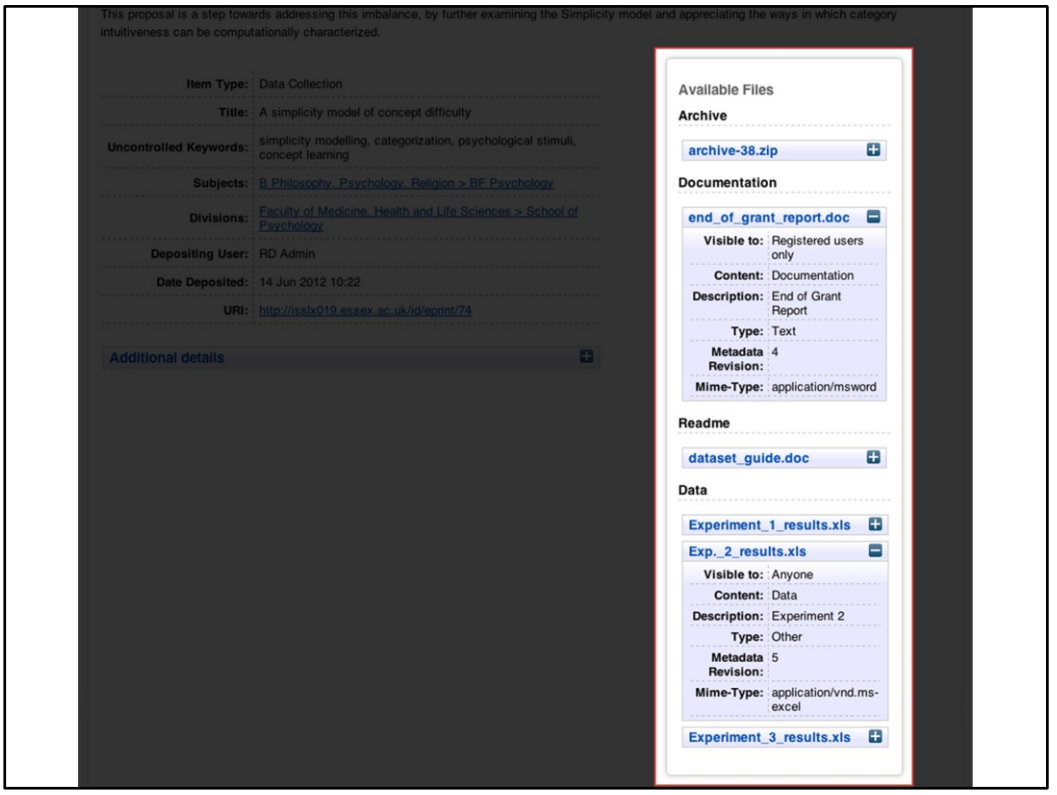
Metadata full unrolled

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: Archive, Documentation, Readme and Data

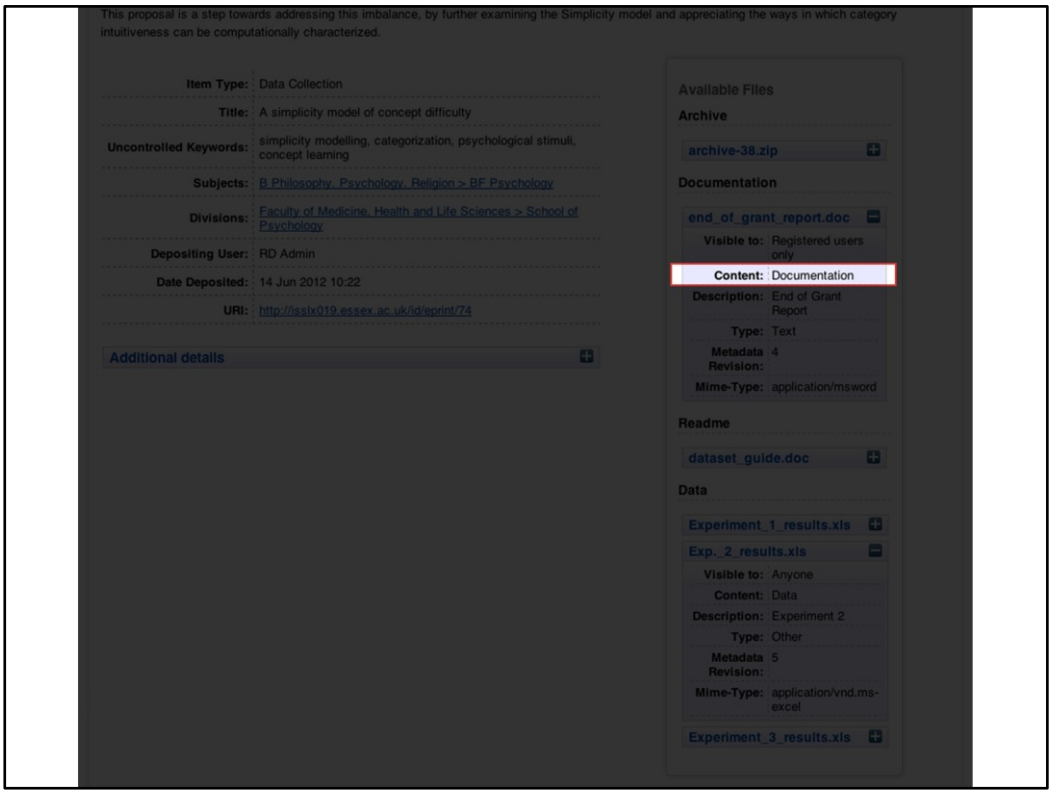


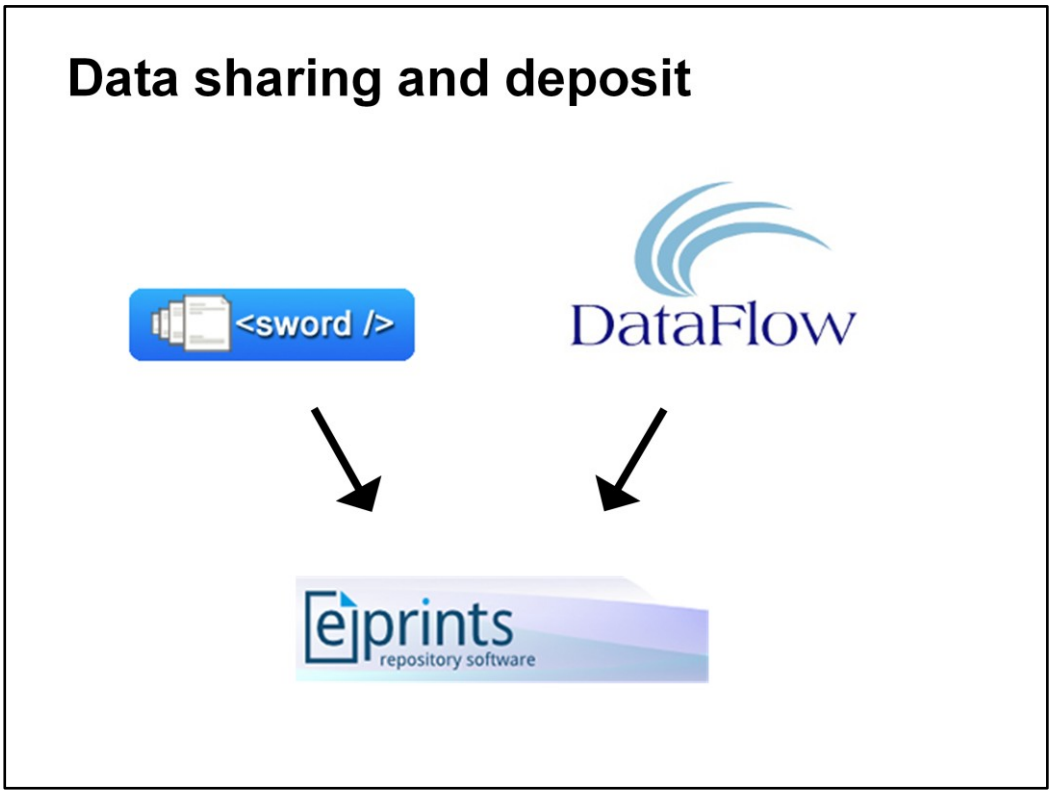
4 Documents associated with each Eprint/ Data Collection extended

We wanted file metadata to be viewable, but not immediately so again we've used collapsible boxes to keep the screen tidy

Quite a bit of debate as to the best way to sort the files – inspired by the ecrystals layout we tried initially to order by file extension, quickly realised this wasn't going to work as
Different content types could have multiple mime types ie .doc, .pdf .xls etc

We decided it's better to sort using a metafield - content


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



While metadata schema was being finalised, we were also asking:
What are the practicalities of technically managing University data?

We were interested in four key areas:

1. Pre-deposit - collaborative local storage environments?
2. Moving data from local storage and depositing data into EPrints
3. Displaying the data files and metadata to users
4. Persistent identification

Sword2:

Interested because Sword could potentially facilitate an easier deposit and therefore encourage researchers to add data.

Dataflow/Datastage

Interested because it's an interesting idea putting together a collaborative environment and Sword based deposit.

Do they help us at this stage of their development? Not really. We need to upload multiple files with very complex metadata. At the moment, it makes much more sense to continue to use the GUI.

We'll be keeping an eye on developments.

What next?

- Continue deposit & ingest testing with real data collections
- Implement DataCite DOIs
  - Using UK Data Archive methodology

- As alternative to discipline specific metadata, test the use of assignable fields for addition of non-standard metadata
  - Gives freedom
  - Recommendation from Southampton's IDMB project
- Should access be at the file or eprint level?
- What data licence options can/should we provide?

Work will continue on testing what we have so far, using real data collections.

We will implement a system for minting data DOIs, adapting a methodology developed at the UK Data Archive

An idea proposed by the IDMB project, Southampton, is to allow depositors to create their own fields using blank fields. This is something we'd like to explore.
OR an alternative approach – discipline specific metadata could be included with data and documentation as additional files e.g. XML

Considering access control options required to cover every scenario, including use of embargos and other item level restrictions.
Related is how to licence data – another chance to draw on UK Data Archive expertise

Challenges for the EPrints community

- Dealing with **complex collections**
  - Large file sizes
  - Large number of files
  - Multiple versions of the same files
  - Inter-dependent files e.g. GIS database
- Adding metadata to files during deposit – how to mass apply
- Looking forward to SWORD2 for data (see http://swordapp.org/2012/07/data-deposit-scenarios)
- Can we visualise data eprints? Many different file types/formats.

Questions?

researchdataessex.posterous.com

@RDEssex

University of Essex

UK · DATA ARCHIVE