Metadata and identifiers for research data

Thomas Ensom
University of Essex

@RDEssex

University of Essex

UK • DATA ARCHIVE

I'm going to talk briefly about the work we've been doing at Essex to create a generic metadata profile to describe research data, and then even more briefly run through our plans for creating persistent identifiers.

Introduction

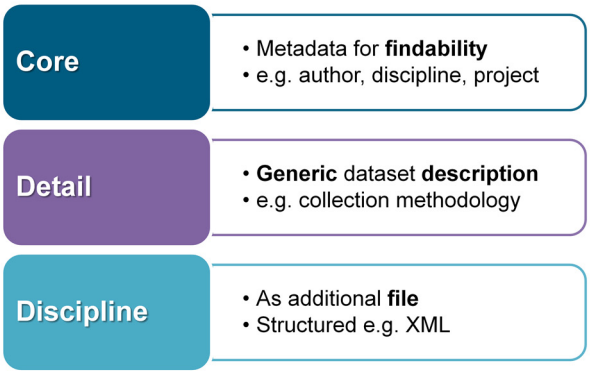
- EPrints repository package tailored to single article type deposits
- …not complex data collections
- Data can also be extremely diverse – varying file types, sizes, quantities, structure etc.

We chose to use EPrints as our technical solution for data storage. It's a popular institutional repository solution but is very much tailored to articles, not complex data collection which may contain many files with different purposes (including documentation). There is also a considerable diversity in the types of data likely to be deposited in an institution with a wide research base, such as Essex.

Our goal was to launch a data repository tailored to meet these challenges.

Introduction

- Define a set of generic metadata elements to allow the effective description of research data



Based on JISC IDMB project outputs, Southampton University

The basis of our metadata approach is an adapted version of the 3 layer metadata model developed by the Institutional Data Management Blueprint project.

Philosophy

- Minimising barriers for researchers to deposit
  …while satisfying requirements for re-use

Deposit should be as easy as possible

But should be more than just a ticked box!

If they can't be bothered the repository could end up full of rubbish..

- **How best to compromise?**

There is a kind of tug-of-war between two conflicting requirements – satisfying conditions for re-use while minimising barriers to deposit. 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 because depositors can't be bothered. Can we find a compromise?

Metadata

- Developed a set of metadata elements based on existing schema:
  - **DataCite:** Minimal mandatory metadata (discovery)
  - **INSPIRE:** Solid basic description of research data with specialist geographic elements, EU standard
  - **DDI 2.1:** Extra detail we think is valuable e.g. collection methodology, ethical issues
  - Inspiration from **DataShare**

We have developed an set of rich descriptive metadata based on several existing schema. DataCite schema is our basis, and also positions us well for using DataCite DOIs in the future. INSPIRE is an EU standard for geospatial data which has the potential to be required for some data generated by institutions in the UK. It also provides an excellent basic description of research data of any kind. Finally we leveraged DDI to capture further descriptive detail such as collection methodology, ethical/consent issues and data processing activities. DataShare provided inspiration on how to present our metadata and the controlled vocabularies and field validations that we might use. This combination allows for easy interpretation for reuse, standards compliance and interoperability.

Metadata challenges

- Lots of metadata schema
  - What is the best route to interoperability?
  - We need to re-use, and include formal definitions so meaning is retained
- Dealing with complex data collections
  - Recording metadata for large numbers of files
  - Making clear the relationships between files
  - Tracking provenance
- Collecting metadata before deposit

Research data management projects, particularly in universities, are faced with an overwhelming number of metadata options. There needs to be some general agreement on a baseline – perhaps our profile would be a good place to start! DataCite minimal is not enough.

Complex data collections present challenges in terms of metadata capture. For example, capturing metadata for individual files in large collections would be impractical for most researchers.

How also, should we link up data collection with what ends up in the repository? Perhaps tools like DataStage could provide the answer?

Identifiers

- We have a **DataCite** test account
  - Specify structure (suffix from eprint #)
  - Mandatory metadata captured
  - Plugin triggers push to DataCite API on indexing (using EPrints event functionality)
  - Receive back DOI which is stored in metadata
- But is DataCite the 'right' route? Also considering other options
  - **Archival Resource Key (ARK)** identifiers?
  - Or just use the **URI?**

The 'Right Way' is still not clear for us. Can the community agree? Will DataCite adjust it's cost model to fit with institutional budgets? A standard which requires buy-in should not be a standard!