RDF Data Management Plans

Practical plans for applications

Andy Turner, EPCC
a.turner@epcc.ed.ac.uk

www.epcc.ed.ac.uk
www.archer.ac.uk

Outline

- Data Management Plans
  - What are they?
  - What will we cover?
- Data organisation
- Data transfer
- Other considerations
  - Repeatability/Reuse

Data Management Plans

Data Management Plans (DMPs)

- Cover all aspects of data lifecycle within a project
  - Policies – who can access data, copyright, security, etc.?
  - Processes – how will data be managed, curated, validated, etc.?
  - Documentation – data layout, metadata, etc.
  - Technology – transfer, storage, sharing, etc.
- Digital Curation Centre have lots of useful resources
  - DMP Checklist: http://www.dcc.ac.uk/resources/data-management-plans/checklist
  - DMPOnline: http://www.dcc.ac.uk/dmponline
    - Online tool to help produce research DMP
- Worth doing for every project!

DMP considerations for RDF/ARCHER

- Focus on *technical* considerations for working with data on ARCHER and RDF
- What do I have to consider if I am using ARCHER/RDF storage?
- What are the performance implications?
- Is my plan feasible from a technical standpoint?
- Will not really cover policy, process and documentation
  - This does not mean that they are not important!!!
  - These may be required by your funder…

Data Organisation

File Systems

- All data on ARCHER and RDF is currently stored in file systems of one sort or another
- The directory structure can be used to encode metadata
  - Still needs to be documented!
  - May need to use tools such as “tar” to preserve this implicit metadata when moving data around
- Often need to think about how to capture that the same dataset may have different versions
  - For source code a version control system (e.g. git) is often used but this is impractical for large files (particularly binary files)
- DMP should describe directory structure and how versioning will be captured

Data Types

- Think about what distinct data types you have, e.g.
  - Source code
  - Input data
  - Simulation data
  - Analysis scripts
  - Analysed/Processed data
- DMP should specify where data will reside
  - May be different places at different points in the workflow
  - Which parts need to be backed-up and how will this be done?
  - Which parts need to be kept long-term?
  - …of course, performance has an impact on how data is organised

Performance Considerations

- Particularly for parallel file systems
  - Not well suited to many small files – try to avoid this if possible
    - Performance will become limited by metadata server rather than exploiting the power parallel file system
  - Do not have many files in a single directory
    - Rule of Thumb: If you have more than 100 files in a single directory this can cause performance issues
- When writing the DMP you should consider any performance bottlenecks
  - Plan how you will organise data to try and avoid these issues
- DMP should specify size (roughly) of your files and how they are organised

Data Transfer

Internal Data Transfers

- Moving data between file systems on same system
- What is the best way to move data?
- Some observations:
  - Never use mv: chance of data loss in flight
  - After you have moved data check the integrity before removing original
  - Do not use compression – become limited by compute performance rather than file system performance
- DMP should specify how you will transfer data internally
  - Are your transfer requirements realistic?
  - How will you ensure the transferred data is valid?

External Data Transfers

- Moving data between different systems
- What data do I need to transfer and why?
  - Minimise the amount transferred
- What is the best supported way of moving the data?
- Is it feasible to use the method chosen to transfer the data?
  - Am I limited by bandwidth or numbers of files?
  - Is the required software available at both ends of the transfer?
- DMP should specify the amount and structure of data to be transferred and the method used

Other Considerations

Reproducibility/Reuse – Data Sharing

- All research should be as reproducible as possible
- This essentially boils down to sharing as much of your data publicly as possible
- Sharing your data also allows for reuse in different research
- When writing your DMP you should consider policies and technologies for making your data available
  - Probably also need to consider how you are going to import your data into a service that shares the data and maintains it long term
- Do not forget to include the source code for any software you used to produce or analyse the data!

Metadata

- As well as actually sharing data you need to ensure it is usable
- This is metadata
  - Describes the data
  - How to navigate the dataset
- Should be documented
  - Preferably in a way that is included with the dataset

Summary

RDF/ARCHER DMPs

- DMP's can cover a wide range of topics
  - Policy, processes, etc.
- Technical DMPs for using ARCHER/RDF
  - Think about how your data is organised in the file system
  - Think about how much data you are transferring and where to
  - Think about which data you really need to move off-site and store long-term
- Other considerations
  - Share your data wherever possible to enable reuse and reproducibility
  - Document your metadata