Data Archiving and Networked Services

Introduction to Data Management Planning

Matthias Hemmje (FTK)
Peter Doorn (DANS)

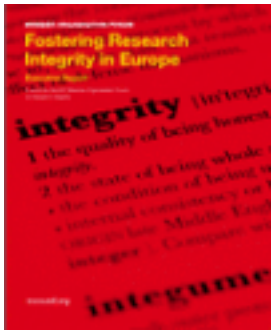
APARSEN-EGI-Community-Forum Training on Data Preservation
Helsinki, 22nd of May 2014

What is data management?

- Data management is a general term covering how you organize, structure, store, and care for the information used or generated during a research project
- It includes:
  - How you deal with information on a day-to-day basis over the lifetime of a project
  - What happens to data in the longer term – what you do with it after the project concludes
- Acronyms:
  - DMP = Data Management Plan
  - RDM = Research Data Management

First 3 slides based on “Research Data Management Training Materials”, University of Oxford,http://damaro.oucs.ox.ac.uk/training_materials.xml

Why spend time and effort on this?



- So you can work efficiently and effectively
  - Save time and reduce frustration
  - Highlight patterns or connections that might otherwise be missed
- Because your data is precious
- To enable data re-use and sharing
- To meet funders' and institutional requirements: good data management is becoming a standard research practice (research integrity, code of conduct)

Funders' requirements

- Funding bodies are taking an increasing interest in what happens to research data
- You may be required to make your data publicly available at the end of a project
  - Check the small print in your grant conditions
- Many funders require a data management plan as part of grant applications

Life cycle of research data

DANS

Driven by data

Data Management Policies in the United States

- The Office of Management and Budget sets forth standards for obtaining consistency and uniformity among federal agencies in the administration of grants to and agreements with institutions of higher education, hospitals, and other non-profit organizations.
- A number of US funding agencies have drawn up data management policies based on that circular.
- Since 2011, the National Science Foundation in the United States has made a data management plan mandatory when submitting grants proposals. Proposals must now include a supplementary document of no more than two pages labelled 'Data Management Plan' which should include the following information:

NSF requirements of DMPs

- **Products of the Research:** The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project.
- **Data Formats:** The standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate this should be documented along with any proposed solutions or remedies).
- Access to Data, Data Sharing Practices and Policies: Policies for access and sharing, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements.
- **Policies for Re-Use, Re-Distribution and Production of Derivatives**
- **Archiving of Data:** Plans for archiving data, samples, other research products, and for preservation of access to them

Celina Ramjoué, Head of Sector, Digital Science Unit, CONNECT.C3 presentation on "Open Access to Scientific Publications and Data"

European Commission

Data management in Horizon 2020

- **Data Management Plans (DMPs) mandatory for all projects participating in the pilot**
- **Other projects invited to submit a DMP if relevant for their planned research**
- DMP questions:
  - What data will be collected / generated?
  - What standards will be used / how will metadata be generated?
  - What data will be exploited? What data will be shared/made open?
  - How will data be curated and preserved?

Digital Agenda

DANS

Driven by data

H2020 Guidelines on RDM

Annex 1: Data Management Plan (DMP) template

- The purpose of the Data Management Plan (DMP) is to provide an analysis of the main elements of the data management policy that will be used by the applicants with regard to all the datasets that will be generated by the project.
- The DMP is not a fixed document, but evolves during the lifespan of the project.
- The DMP should address the points below on a dataset by dataset basis and should reflect the current status of reflection within the consortium about the data that will be produced.

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

DANS

Driven by data

- **Data set reference and name:** Identifier for the data set to be produced.
- **Data set description:** Description of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.
- **Standards and metadata:** Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.
- **Data sharing:** Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.). In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).
- **Archiving and preservation (including storage and backup):** Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.

Annex 2: Additional guidance for Data Management Plans

Scientific research data should be easily:

1. **Discoverable:** are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier)?
2. **Accessible:** are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses (e.g. licencing framework for research and education, embargo periods, commercial exploitation, etc.)?
3. **Assessable and intelligible:** are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review (e.g. are the minimal datasets handled together with scientific papers for the purpose of peer review, are data is provided in a way that judgments can be made about their reliability and the competence of those who created them)?

Annex 2: Additional guidance for DMPs (cont'd)

4. **Useable beyond the original purpose for which it was collected:** are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data (e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists)?
5. **Interoperable to specific quality standards:** are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc. (e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing recombinations with different datasets from different origins)?

Final remarks & some questions

- DMP will be mandatory
- Agreement on basic principles
- Variations in the degree of detail:
  - depending on scientific domain, size of project, funding instrument/organisation, institution…
- Procedures are under construction
  - data management section in proposal stage
  - data management plan as part of contract or in first phase of project?
  - DMP as a living document?
- Not just a paper (or digital) tiger…
  - who will check? Sanctions?
- To what degree will RDM be fundable?

Data Archiving and Networked Services

Thank you for your attention

www.dans.knaw.nl
www.narcis.nl
peter.doorn@dans.knaw.nl

www.ftk.de
mhemmje@ftk.de