MATES ED2MIT

Education and Training for Data Driven Maritime Industry

Tutorial A03

Case Study: Research Data Management

Maritime Alliance for fostering the
European Blue economy through a
Marine Technology Skilling Strategy

Yuri Demchenko MATES Project
University of Amsterdam

Co-funded by the
Erasmus+ Programme
of the European Union

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Outline

- Overview status and initiatives: RDM in Europe
  - Open Science, Open Data, Open Access, European Open Data Pilot
  - FAIR Data Principles and Data Stewardship
- Data lifecycle and data management factors
  - Data Management Plan
- Data Management basics
  - Creating documentation and metadata, metadata for discovery, PID
- Backing up your data
- Responsible Data Use (citation, copyright, data restrictions)
  - Handling sensitive data
  - Personal data management, GDPR
  - Ethical aspects
- Data Stewardship, Competences and demand

- Practice: Creating your Data Management Plan (DMP)
  - Industry Data Management and Governance

Co-funded by the
Erasmus+ Programme
of the European Union

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Research Data Management - Part 1

- Open Access, Open Data, Open Science
- PID, ORCID

Open Access to Scientific Publications

- EC initiative on Open Access scientific publications from publicly funded projects
  - Included into Declaration from the H2020 Rome meeting (2012)
  - Approx 3500 publicly funded ROs and 2000 privately funded ROs
  - Special funding scheme for reimbursing publications
  - Issues with China, India, Russia compliance to OA principles
    - Consultation at high governmental level
- OpenAIRE project exploring models for open access to publications - https://www.openaire.eu/
  - PID (Persistent ID for data), ORCHID (Open Researcher ID), Linked data
  - Started as EU funded project, now is a member funded service

EU policy on Open Research Data

- Research data can be defined as whatever is either produced in the research process or evidences research outputs such as articles
- The European Commission's Research Data definition is: "information, in particular facts or numbers, collected to be examined and considered and as a basis for reasoning, discussion, or calculation"
  - https://ec.europa.eu/research/openscience/index.cfm?pg=openaccess
  - https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm
  - https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
- Examples include: statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings, images
- Open data are deposited in institutional or specialist repositories and licensed appropriately so that prospective users know clearly any limitations on re-use.

Horizon 2020 Open Research Data (ORD) Pilot

- ORD pilot aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects, taking into account
  - the need to balance openness and protection of scientific information
  - commercialisation and IPR
  - privacy concerns
  - security
  - data management and preservation questions
- Applying principle '**as open as possible, as closed as necessary**'
- Complying with FAIR Data principles
- ORD applies primarily to the data needed to validate the results presented in scientific publications.
  - Other data can also be provided by the beneficiaries on a voluntary basis.

Horizon 2020 Data Management and Data Management Plan

- Data Management Plans (DMPs) are a key element of good data management.
  - A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project.
  - Help making research data Findable, Accessible, Interoperable and Reusable (FAIR)
- DMP should include information on:
  - the handling of research data during & after the end of the project
  - what data will be collected, processed and/or generated
  - which methodology & standards will be applied
  - whether data will be shared/made open access and
  - how data will be curated & preserved (including after the end of the project).
- The project **must submit a first version of DMP** (as a deliverable) within the **first 6 months** of the project.
  - DMP is updated if data are changed
  - DMP is mandatory for projects participating in ORD Pilot

[ref] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

Open and Toll Access (OA and TA)

- Open Access generally refers to the outputs of research, such as journal articles, as distinct from research data, which are produced as part of the research process
- Open Access is differentiated from the traditional method of access to research outputs, known as Toll Access
  - Toll Access can be by means of institutional or personal subscription to journals, or to aggregations of content, or by means of paying publishers for access to individual articles
  - Toll Access payment is reader-side

Open Access Definition

Budapest Open Access Initiative (BOAI) 2002, reaffirmed in 2012:

- By "open access" to … literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.
  - http://www.budapestopenaccessinitiative.org/boai-10-recommendations
  - Copyright constraints are applied to protect integrity of work

Peter Suber's Concise Definition:

- Open Access literature is "digital, online, free of charge, and free of most copyright and licensing restrictions" (Suber, P. Open access. MIT Press, 2012. Available at:
  - https://mitpress.mit.edu/sites/default/files/9780262517638_Open_Access_PDF_Version.pdf

Gratis and Libre OA

- Context: Intellectual property laws generally offer ***limited "fair dealing" or "fair use" exemptions***
- Gratis OA is free of charge to access but subject to the limits of fair dealing
  - it removes toll barriers but not permission barriers
- Libre OA is both free of charge and free of at least some legal and licensing restrictions
  - it removes toll barriers and at least some permission barriers
- The BOAI (Budapest Open Access Initiative) definition is Libre.

Green OA –1 and Green OA - 2

Green OA -1 is delivered through **self-archiving**: authors deposit manuscripts in institutional or disciplinary repositories;

- Relies on a recent but well established infrastructure of repositories
- Is easy and cheap: each article only incurs a very small portion of the overhead costs of setting up and running repositories
- Does not incur the overheads of peer-review;
- However, deposited articles may be, most often have been, peer-reviewed for publication in traditional Toll Access journals

Green OA – 2 is compatible with subscription journal publishing: scholars can **publish in TA** (Transactional Analysis) journals and, through **self-archiving**, still make their articles OA (author's final peer-reviewed manuscript, without the formatting or pagination of the published version)

- Is often subject to an **embargo period** imposed by publishers, generally of between 6 and 12 months
- Depends on authors' obtaining rights from publishers to deposit and make articles available
- Is hospitable to many other types of document, notably pre-prints, theses, and reports.

Gold OA-1 and Gold OA-2

Gold OA – 1: **Offers articles that are paid for by the authors or their institutions or funders**

- Articles may be either in completely OA journals or in hybrid journals, containing both OA and TA articles
- Articles are peer-reviewed for publication
- Incurs much the same costs for the editorial and peer review process as TA journal publishing
- Is always immediate, while Green OA is often subject to time embargoes imposed by subscription journal publishers.

Gold OA – 2: Provides access to the **published version of an article**, while Green OA generally provides access only to the author’s final peer-reviewed manuscript, without the formatting or pagination of the published version

- By its nature is confined to post-prints
- Generally obtains rights and permissions direct from the rights-holder (usually the author);
- Is delivered through journals: these may be completely OA or hybrid, where some articles are OA and others toll access;
- Both Green and Gold OA are gratis. Green OA generally is only gratis; Gold OA may be Libre.

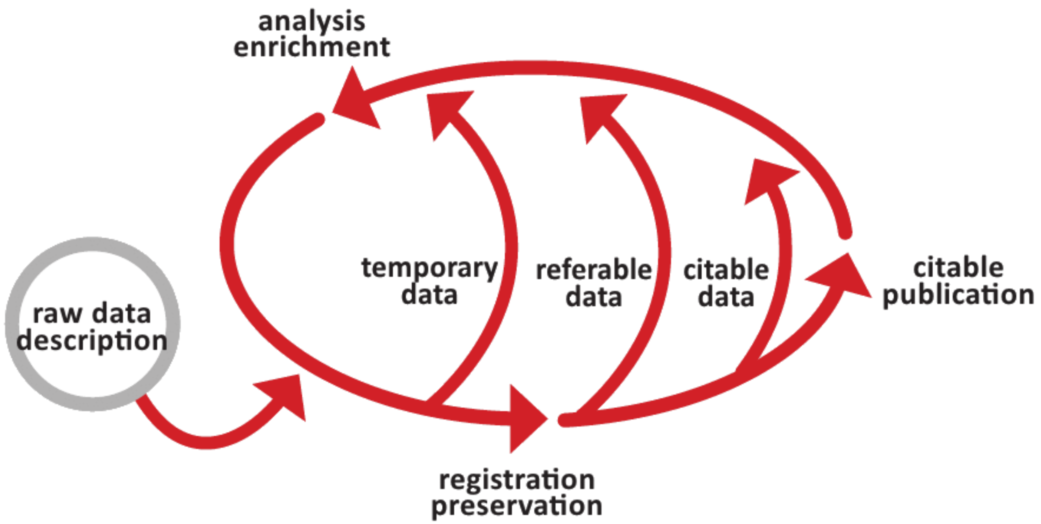
Self-archiving services

- Zenodo - https://zenodo.org/
  - Zenodo helps researchers receive credit by making the research results citable and through OpenAIRE integrates them into existing reporting lines to funding agencies like the European Commission.
  - Citation information is also passed to DataCite and onto the scholarly aggregators.
  - Collects rich metadata on the archived publications
  - Publications recognised by EC as project related publication – mandatory for some programmes and projects

- Arxiv (Cornell Univ service) - https://arxiv.org/
  - arXiv is a free distribution service and an open-access archive for 1,777,731 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics.

Persistent Identifier (PID)

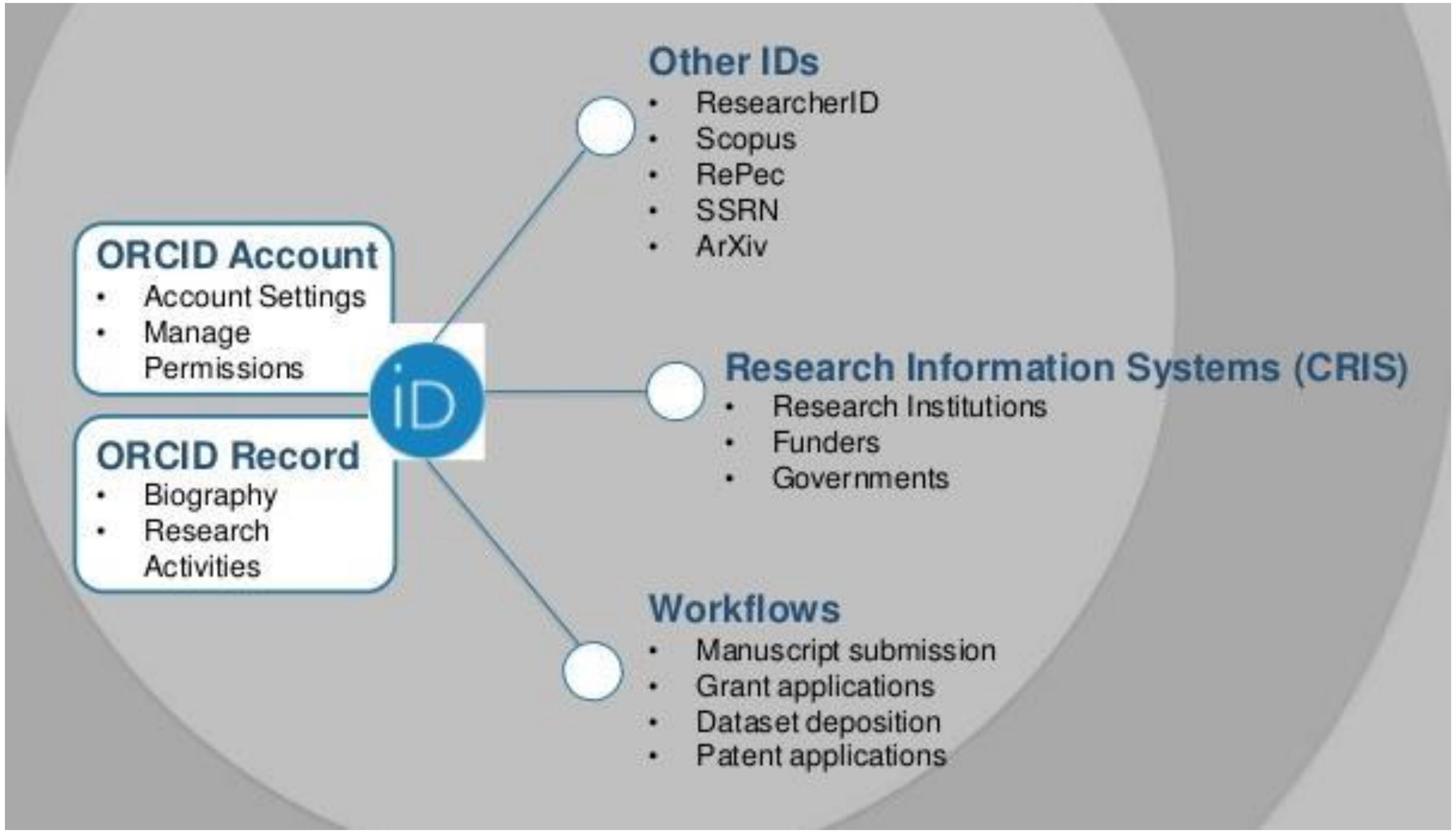
- PID – Persistent Identifier for Digital Objects
  - Managed by European PID Consortium (EPIC) http://www.pidconsortium.eu/
  - Superset of DOI - Digital Object Identifier (http://www.doi.org/)
  - Handle System by CNRI (Corporation for National Research Initiatives) for resolving DOI (http://www.handle.net/)
- PID provides a mechanism to link data during the whole research data transformation cycle
  - EPIC RESTful Web Service API published May 2013



ORCID - Connecting research and researchers

- Research in the digital realm is becoming increasingly linked up
  - Leverage this to increase your profile
  - Get an ORCID (**Open Researcher and Contributor ID**) and identify yourself as a unique researcher
  - ORCID provides a persistent digital identifier that distinguishes you from every other researcher i.e. that Dr. John Smith
  - Looks something like: http://orcid.org/xxxx-xxxx-xxxx-xxxx
  - Simple and free to register at: http://orcid.org/

Connecting research and researchers



ORCID Account
- Account Settings
- Manage Permissions

ORCID Record
- Biography
- Research Activities

Other IDs
- ResearcherID
- Scopus
- RePec
- SSRN
- ArXiv

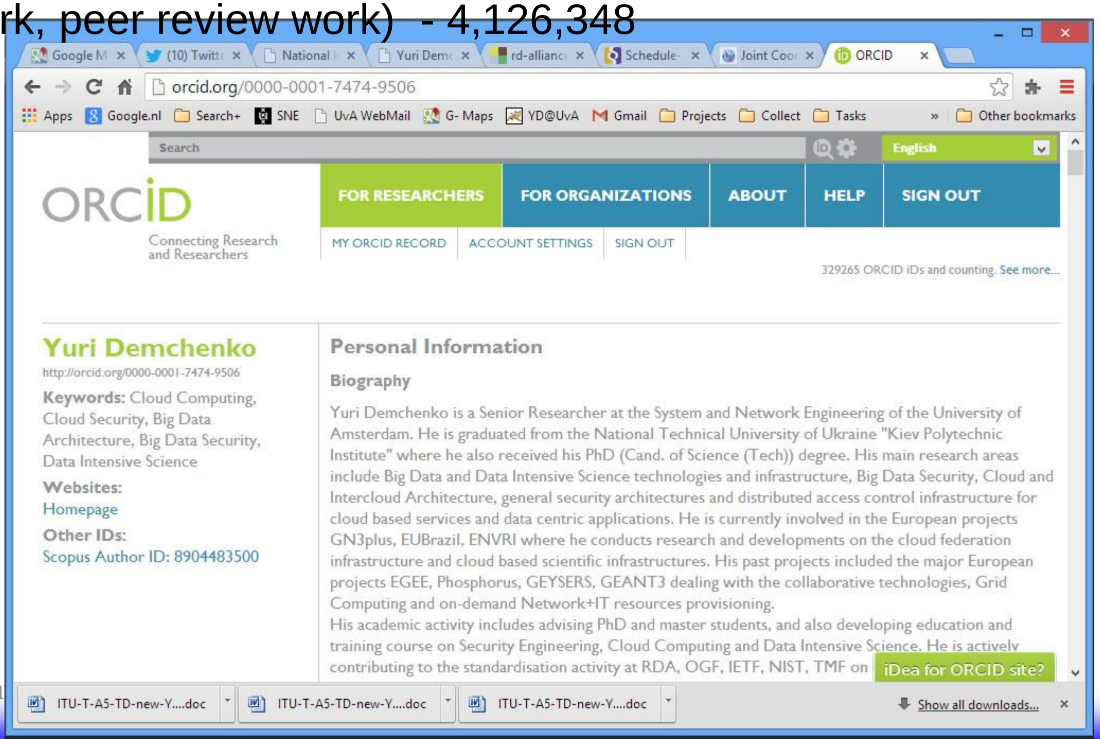
Research Information Systems (CRIS)
- Research Institutions
- Funders
- Governments

Workflows
- Manuscript submission
- Grant applications
- Dataset deposition
- Patent applications

- Link together your research
- Source: ORCID: Connecting Research and Researchers, Biblioteca del Campus Terrassa on Jul 11, 2013

ORCID (Open Researcher and Contributor ID)

- ORCID is a nonproprietary alphanumeric code to uniquely identify scientific and other academic authors
  - Launched October 2012
- ORCID Statistics – February 2021
  - Live ORCID IDs – **10783465 (Feb 2021),** 9 745 841 (May 2016 - 511, 203; October 2013 - 329,265)
  - ORCID IDs with at least one work 121,529 (October 2013 - 79,332)
  - IDs with external identifiers (person, org, funding, work, peer review work) - 4,126,348
  - Works 62,229,838
  - Works with unique DOIs 22,703,095
- Personal ORCID
  - ORCID **0000-0001-7474-9506**
  - **http://orcid.org/0000-0001-7474-9506**
  - Scopus Author ID 8904483500



329265 ORCID iDs and counting. See more...

Yuri Demchenko

http://orcid.org/0000-0001-7474-9506

Keywords: Cloud Computing, Cloud Security, Big Data Architecture, Big Data Security, Data Intensive Science

Websites:
Homepage

Other IDs:
Scopus Author ID: 8904483500

Personal Information

Biography

Yuri Demchenko is a Senior Researcher at the System and Network Engineering of the University of Amsterdam. He is graduated from the National Technical University of Ukraine "Kiev Polytechnic Institute" where he also received his PhD (Cand. of Science (Tech)) degree. His main research areas include Big Data and Data Intensive Science technologies and infrastructure, Big Data Security, Cloud and Intercloud Architecture, general security architectures and distributed access control infrastructure for cloud based services and data centric applications. He is currently involved in the European projects GN3plus, EUBrazil, ENVRI where he conducts research and developments on the cloud federation infrastructure and cloud based scientific infrastructures. His past projects included the major European projects EGEE, Phosphorus, GEYSERS, GEANT3 dealing with the collaborative technologies, Grid Computing and on-demand Network+IT resources provisioning.
His academic activity includes advising PhD and master students, and also developing education and training course on Security Engineering, Cloud Computing and Data Intensive Science. He is actively contributing to the standardisation activity at RDA, OGF, IETF, NIST, TMF on

RDM Focus: FAIR Data Principles

Findable:

- F1 (meta)data are assigned a globally unique and persistent identifier;
- F2 data are described with rich metadata;
- F3 metadata clearly and explicitly include the identifier of the data it describes;
- F4 (meta)data are registered or indexed in a searchable resource;

Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles;
- I3. (meta)data include qualified references to other (meta)data;

Accessible:

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol;
  - A1.1 the protocol is open, free, and universally implementable;
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary;
- A2 metadata are accessible, even when the data are no longer available;

Reusable:

- R1 meta(data) are richly described with a plurality of accurate and relevant attributes;
- R1.1 (meta)data are released with a clear and accessible data usage license;
- R1.2 (meta)data are associated with detailed provenance;
- R1.3 (meta)data meet domain-relevant community standards;

FAIR from the technical point of view

- Findable
  - Metadata and PDI – infrastructure and tools
  - Registries and handles resolution, API
  - Policies and SLA
- Accessible
  - Repositories and data storage: infrastructure and management
  - Policy and access control: infrastructure and API management
  - Data access protocols
  - Usage Policy and Sovereignty
  - Data protection, compliance, privacy and GDPR
- Interoperable
  - Standard data formats
  - Metadata and API
  - **FAIR maturity level and certification**
- Reusable
  - Data provenance and lineage
  - Preservation
  - Metadata, PID and API – linked or embedded into datasets

This motivates Data Stewards' interaction with both **Data Analytics and Applications developers** roles and **Data Infrastructure** roles

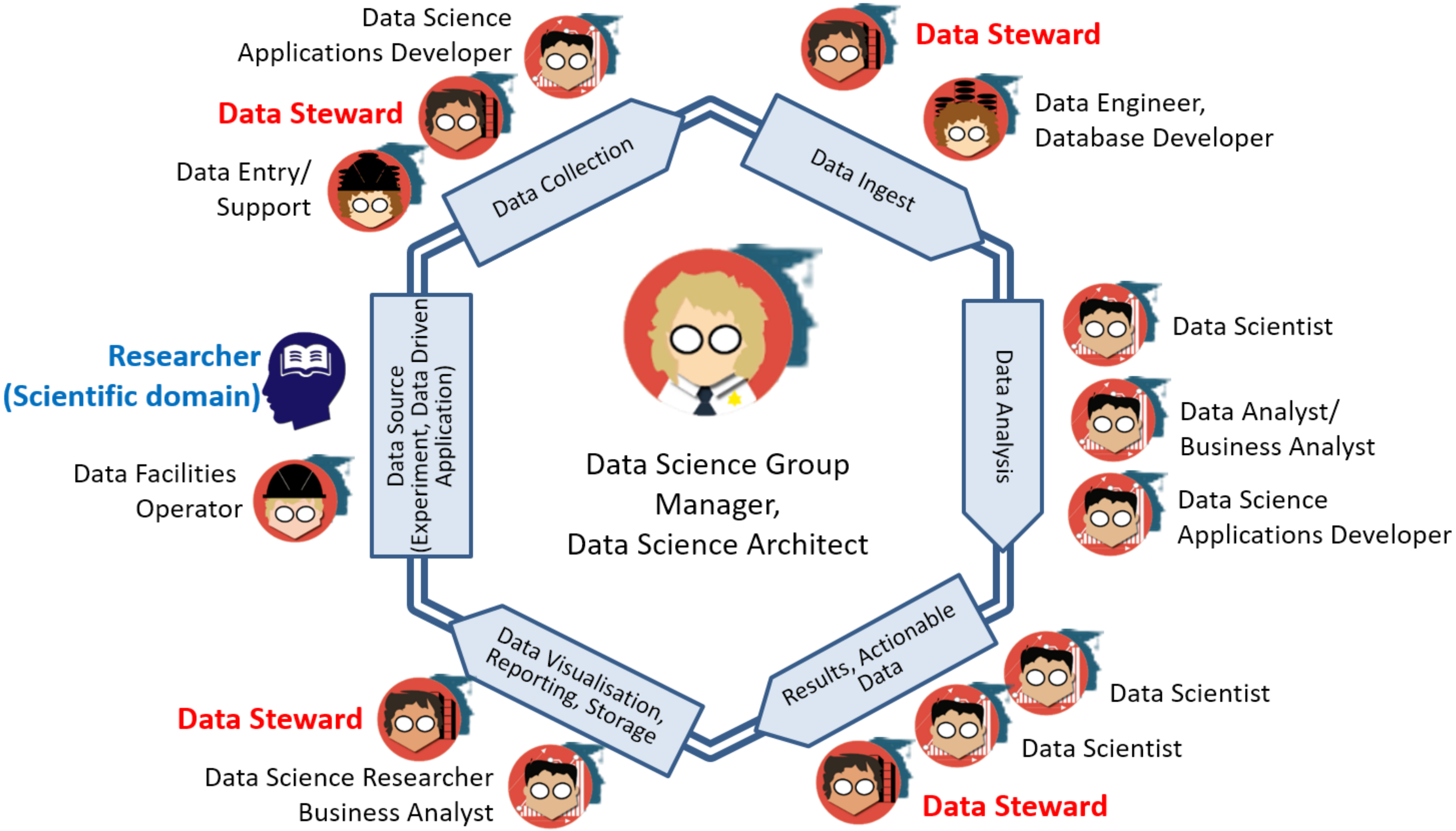
- Consequently related competences from Data Stewards are needed

FAIR Data Management and Organisational Roles

FAIR data principles to be adopted cross organisation for the whole data lifecycle

- Data collection
  - Researchers, Data Engineers, data entry workers
- Data preservation and curation
  - Data curators, Data Custodians/Archivists
- Data Analysis
  - Data Scientists, Data Architects, Application developers
- Data publication, sharing access
  - Data Stewards, Data Curators
- Data Governance and Data management
  - Data Stewards and CDO
    - Data policy and data delivery agreements
- Data Infrastructure and tools for data storage and handling
  - Storage, database engineers/managers
    - Metadata and PID services, Master data and Reference data

Building Data Science Team – Variety of Roles (Example)



Data Science Applications Developer
Data Steward
Data Entry/ Support
Data Steward
Data Engineer, Database Developer
Data Collection
Data Ingest
Researcher (Scientific domain)
Data Facilities Operator
Data Source (Experiment, Data Driven Application)
Data Science Group Manager, Data Science Architect
Data Analysis
Data Scientist
Data Analyst/ Business Analyst
Data Science Applications Developer
Data Visualisation, Reporting, Storage
Results, Actionable Data
Data Steward
Data Science Researcher Business Analyst
Data Scientist
Data Scientist
Data Steward

Data Stewardship in Research and FAIR Principles – GO FAIR and GO TRAIN

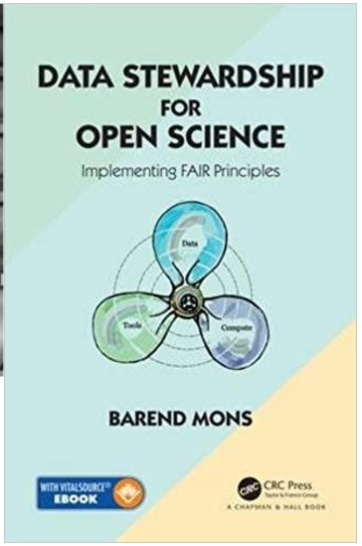
- FAIR Initiative by Dutch Techcentre for Life Science (DTLS) – Prof. Barend Mons
  - Part of Horizon 2020 Programme
- FAIR Principles for research data:
  Findable – Accessible – Interoperable - Reusable
- Data Stewards as a **key bridging role** between Data Scientists as (hard)core data experts and scientific domain researchers (HLEG EOSC report)
- Current definition of the Data Steward (part of Data Science Professional profiles - EDSF)
  - Data Steward is a **data handling and management professional** whose responsibilities include planning, implementing and managing (research) data input, storage, search, and presentation.
  - Data Steward creates data model for **domain specific data**, support and advice domain scientists/ researchers during the whole research cycle and data management lifecycle.

Realising the European Open Science Cloud

First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud

HLEG report on European Open Science Cloud

(October 2016)



DATA STEWARDSHIP FOR OPEN SCIENCE

Implementing FAIR Principles

BAREND MONS

Structure of the Data Governance Policy Document

- Data Flows
- Inventory of Data Assets
- Data Sensitivity Classification
- Data Quality
- Data Standards
- Data Sharing and/or Linking
- Data Governance and Organisational Roles
- Data Stewardship
- Awareness and Training

- Appendix A. Data Management Plan (developed per department or project)

Data management: Everything but analysis

- Organising
  - file naming and formatting
  - data formats and software
  - file transfers, file sharing and remote access
  - version control
- Administering
  - back-ups
  - documentation and metadata
  - access controls
  - security
- Storing and sharing
- Ethical and legal aspects of data handling and data ownership

Documenting Data – Importance

A crucial part of making data user-friendly, shareable and with long-lasting usability is to ensure they can be understood and interpreted by any user. This requires clear and detailed data description, annotation and contextual information.

- Areas of coverage
  - Study-level documentation and context
  - Data-level documentation
  - Metadata
  - Context debate
- Data doesn’t mean anything without documentation
  - a survey dataset becomes just a block of meaningless numbers
  - an interview becomes a block of contextless text
- Data documentation might include:
  - a survey questionnaire
  - an interview schedule
  - records of interviewees and their demographic characteristics in a qualitative study
  - variable labels in a table
  - published articles that provides background information
  - description of the methodology used to collect the data

What should be captured

- Contextual information about project and data
  - background, project history, aims, objectives, hypotheses
  - publications based on data collection
- Data collection methodology and processes
  - data collection process and sampling
  - instruments used - questionnaires, showcards, interview schedules
  - temporal/geographic coverage
  - data validation - cleaning, error-checking
  - compilation of derived variables
  - weighting: factors and variables, weighting process
  - secondary data sources used
- Data confidentiality, access and use conditions
  - anonymisation carried out
  - consent conditions/procedures
  - access or use conditions of data

Consider documentation early on

- Good data documentation and metadata depends on what you as the creator can provide
- Start gathering meaningful information from as early on in the research process as possible
- This consideration forms an important part of data management planning (which you will hear more on later in the course)

- Quantitative study
  - Smaller-scale study – single user guide may contain compiled survey questionnaire, methodology information
  - For complex studies - many documents presented separately

- Qualitative study
  - A user guide could contain a variety of documents that provide context: interview schedule, transcription notes, even photos
  - Data listing provides an at-a-glance summary of interview sets

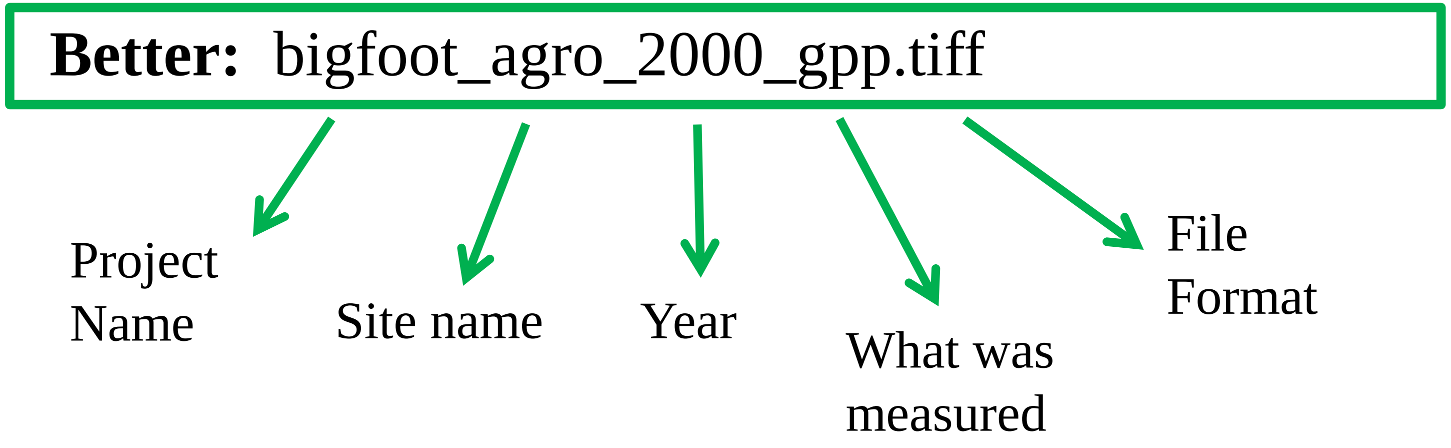
Managing Data: Assign Descriptive File Names

- Clear, descriptive, and unique file names may be important later when your data file is combined in a directory or FTP site with your own data files or with the data files of other investigators
- File name = principal identifier of file
  - use logical naming i.e. easy to identify and retrieve the file
  - naming provides organisation, context & consistency
  - name elements: version number, date, content description, creator name
- Best practice
  - name independent of location (i.e. domain/server, directory)
  - relevant to content
  - no special characters, dots or spaces
  - for separation use underscores _
  - versioning via filename: ascending, decimal version numbers
  - avoid very long file names

Assign descriptive file names

- Use descriptive file names
  - Unique
  - Reflect contents
  - ASCII characters only
  - Avoid spaces
- Provide an explanation of the convention used to name files

Bad: Mydata.xls
2001_data.csv
best version.txt



Better: bigfoot_agro_2000_gpp.tiff

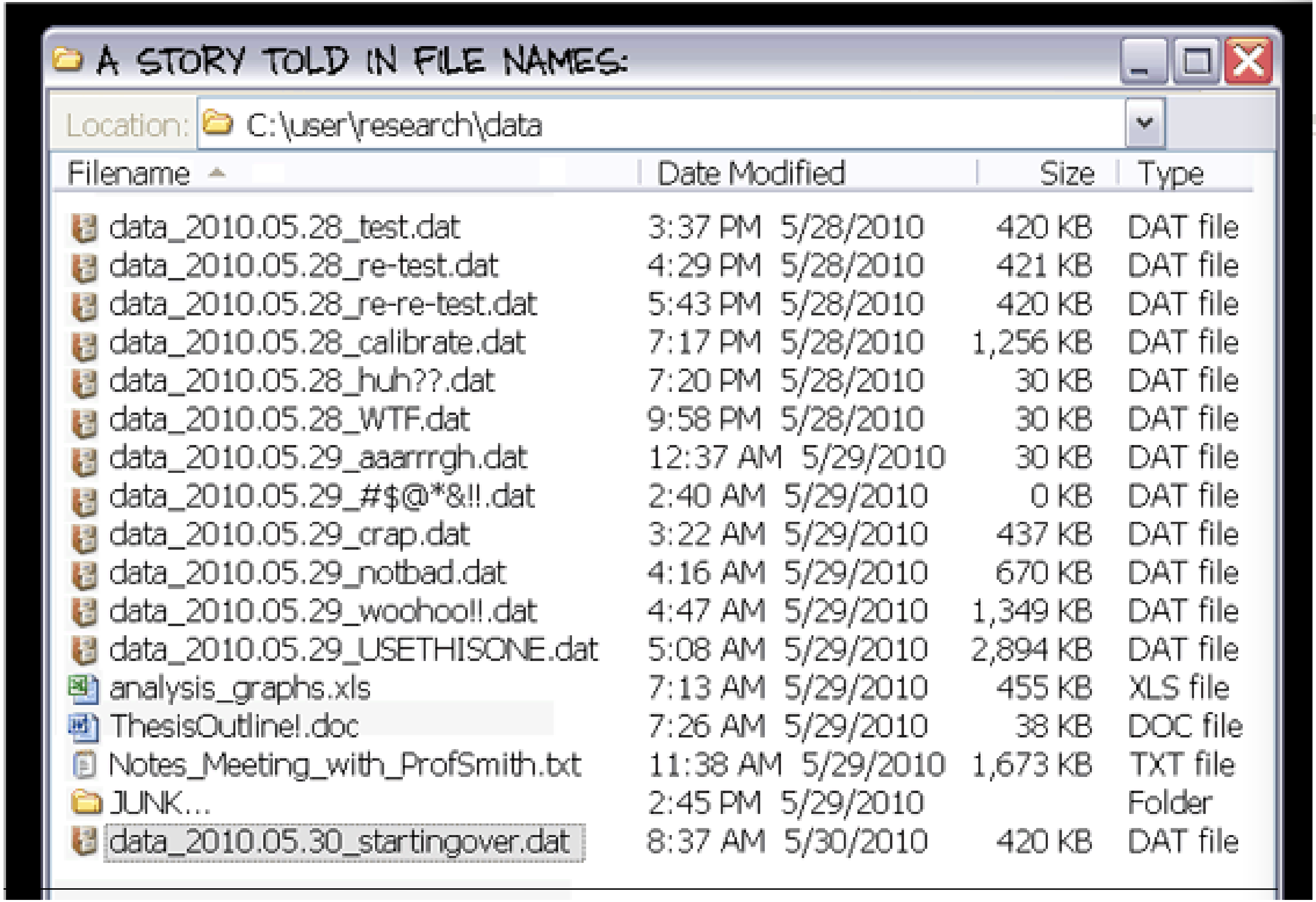
Project Name

Site name

Year

What was measured

File Format



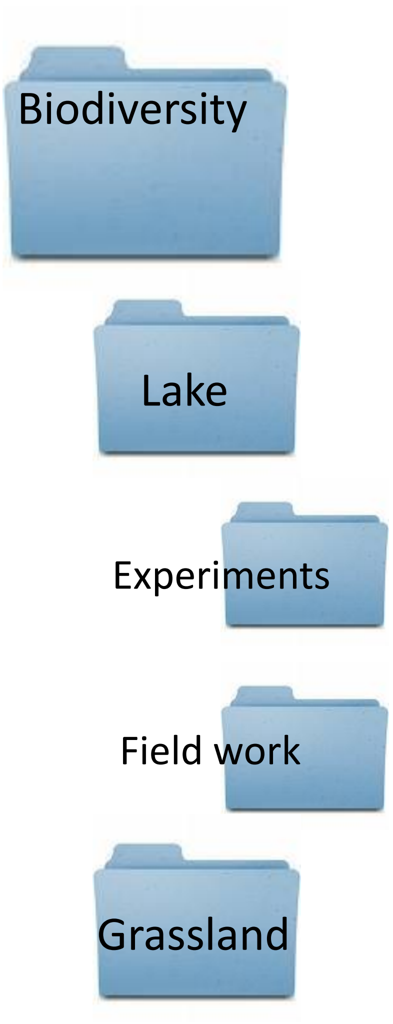
A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Courtesy of PhD Comics: http://www.phdcomics.com/comics.php

Organize files logically



Biodiversity

Lake

Experiments

Field work

Grassland

Make sure your directory system is logical and efficient

Biodiv_H20_heatExp_2005_2008.csv
Biodiv_H20_predatorExp_2001_2003.csv
….

Biodiv_H20_planktonCount_start2001_active.csv
Biodiv_H20_chla_profiles_2003.csv
…

Version Control and Document “ownership”

- Keep track of different copies or versions of data files
  - useful for files kept in multiple locations
  - or which have multiple users
  - a way to safeguard against accidental changes
  - collaboratively edit documents in ‘the cloud’ while tracking version history
    - Vs GoogleDoc change tracking and cooperative editing
  - Use CVS, Subversion or WebDAV platforms
- File names are a good way to do this
  - unique descriptive names for files
  - include date and/or version number in name
  - indicate relationships between files
  - e.g. FoodInterview_1_draft; FoodInterview_1_final; HealthTest_06-04- 2008; BGHSurveyProcedures_00_04

- Example: Document versioning best practice

Example: Document versioning best practice

- Document owner assigns version number
- Contributors provide contribution, edit – append their initials to current version

- Example:
  - cyclon-D5.2-data-management-v01.doc
  - cyclon-D5.2-data-management-v01-jd.doc – contribution by John Doe
  - cyclon-D5.2-data-management-v01-mc.doc – contribution by Mary Claire
  - cyclon-D5.2-data-management-v01-mc01.doc – 2nd contribution by Mary Claire to own version
  - cyclon-D5.2-data-management-v01-mc-tvdb.doc – contribution by Tom van den Berg
  - cyclon-D5.2-data-management-v02.doc – new version by document owner

- GoogleDoc or Word – bad for tracking versioning

Archiving non digital content

- Create searchable PDF
  - collate TIFFs and convert to PDF
  - bookmark PDF file for navigation: contents page, headings & metadata
- Create rich text using Optical Character Recognition (OCR)
  - automatically convert TIFF to RTF format
  - requires rigorous proof reading and checking
- Transcribe manually
  - represent the original material as closely as possible
  - avoid using formatting in data files
- Data transcription
  - translation between forms
  - transcription to be
    - representational
    - selective – can be multiple-perspective for video
    - interpretive
    - theoretical

Metadata

- Metadata definition
- Dublin Core
- Discovery Level Metadata
- Creating a Citation for Your Data
- Sharing your data

Metadata – Data about data

- Highly structured machine readable documentation
- Standard data collection metadata includes:
  - Components of a bibliographic reference
  - Core information that a search engine indexes to make the data findable
- Metadata standards are digital containers for structured information about a data set
- International standards/schemes for metadata
  - ISO 19115 http://www.fgdc.gov/metadata/geospatial-metadata-standards#nap
  - GCMD DIF http://gcmd.nasa.gov/User/difguide/difman.html
  - DataCite http://schema.datacite.org/
  - **Dublin Core**
  - Data Documentation Initiative (DDI)
  - Metadata Encoding and Transmission Standard (METS)
  - Preservation Metadata Maintenance Activity (PREMIS)

Dublin Core Metadata

The original Dublin Core Metadata Element Set consists of 15 metadata elements:

- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights

Each Dublin Core element is optional and may be repeated.

Example of code

```
<meta name="DC.Format" content="video/mpeg; 10 minutes">
<meta name="DC.Language" content="en" >
<meta name="DC.Publisher" content="publisher-name" >
<meta name="DC.Title" content="HYP" >
```

[RFC5013] http://www.ietf.org/rfc/rfc5013.txt

[NISOZ3985]_http://www.niso.org/apps/group_public/download.php/10256/Z39-85-2012_dublin_core.pdf

[ISO15836]_http://www.iso.org/iso/search.htm?qt=15836&searchSubmit=Search&sort=rel&type=simple&published=on

[TRANSLATIONS] http://dublincore.org/resources/translations/

[DCTERMS]_http://dublincore.org/documents/dcmi-terms/

Why Metadata - Example

Temperature

31.5

Why Metadata – Example

Temperature

31.5

Of what?

According to whom?

For what purpose?

Precision/accuracy?

Has anyone checked the quality of this value?

Collected when?

In what units?

Location?

When was the sensor last cleaned/calibrated?

Collected how?

Is this value averaged? Calculated?

AKA – T, Temp, degC, C, ^{o}F… lots of different names

How to create metadata for data - Tools

- Can be compiled using data deposit forms/tools
  - Currently not many available that are user friendly and maintained
  - May be better to create a spreadsheet
- Data Documentation Initiative (DDI) documentation can be created in software packages using certain DDI tools:
  http://tools.ddialliance.org – rich catalog
- Colectica Designer for survey data – Paid software
  http://www.colectica.com/software/designer
  - Create and publish metadata
- Nesstar Publisher 4.0 convert SPSS internal metadata to DDI using
  http://www.nesstar.com/software/publisher.html

Discovery Level Metadata

- A data set description (metadata) that provides information to determine if a particular data set meets the users’ needs.
- Typically provides essential information to enable a user to find out if a particular dataset exists, the data’s location, and ownership, and how to obtain further information.
- The metadata includes the science discipline of the data, data location, spatial coverage, data provider, data resolution, data quality, etc.
- Discovery level metadata is found in “portals” and metadata registries.
- A controlled keyword vocabulary helps provide a consistent search and discovery of data.

Categories of Discovery Level Metadata

What: Title of Data Set and Keywords Describing the Data Set

Why: Description and Purpose of the Data Set

When: Temporal Coverage of the Data Set

Who: Data Set Creator and Contact

Where: Geographic Extent and Location of Data Set Coverage

How: How the Data Set was Created and How to Access the Data

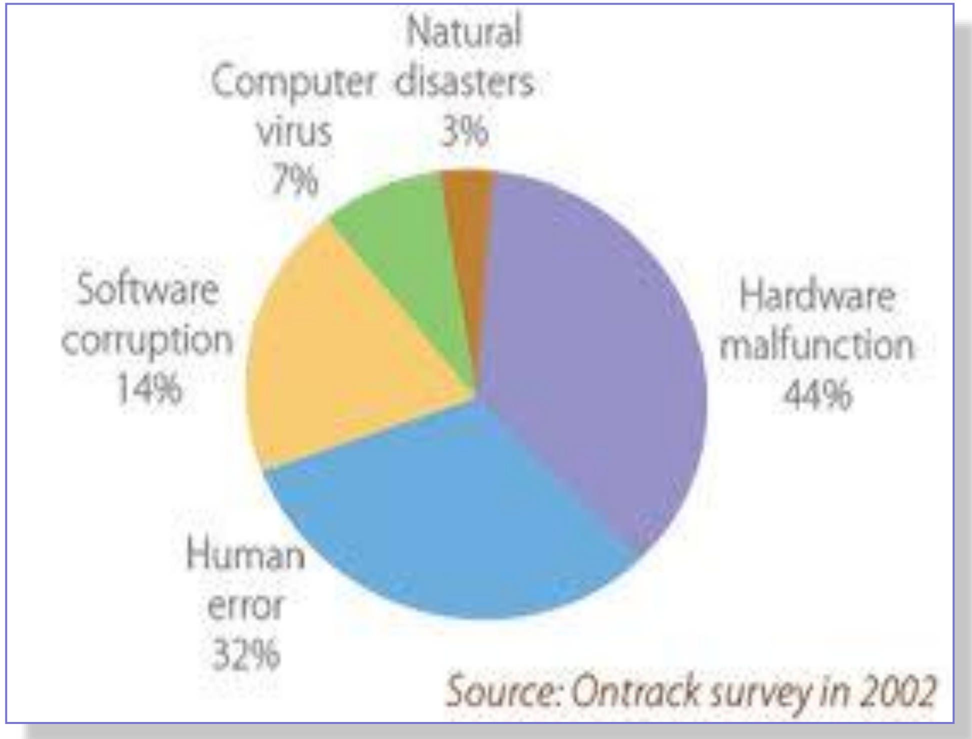
- Discovery level metadata makes it easier to find relevant data in portals, metadata registries, and data inventory systems.
- Being able to find and distinguish data from other similar data sets makes maintaining a data inventory easier because data managers have a better understanding of the content in their system.
- Creating and maintaining metadata is part of a data management lifecycle.

Storing and Backing up your Data

- Backup strategy
- Storage options
- Data security strategy

Backing Up Your Data

- Valuable data and information can be lost
- Limit loss of data, some of which may not be reproducible
  - Save time, money, productivity
- To protect against data loss, create multiple copies of files located in several sites
  - These files can be used to replace lost files
- Automatically test backup copies of files frequently to ensure they are viable
  - **Media degrade over time**
  - Annually test copies using checksums or file compare



Source: Ontrack survey in 2002

http://www.netpipe.com/NBS/web/page_1.html

Backing-up strategy

Consider:

- **What needs to be backed-up?** All, some, just the bits you change?
- **What media?** External hard drive, DVD, online etc.
- **Where?** Original copy, external local and remote copies
- **What method/software?** Duplicating, syncing, mirroring
- **How often?** Assess frequency and automate the process
- **For how long?** How long you will manage these backups for
- **How can you be sure?** Never assume, regularly test a restore, and use verification methods

Storage options - Local data storage (1)

Local data storage

- All digital media are fallible
- Optical (CD, DVD) & magnetic media (hard drives, tape) degrade – lifespan even lower if kept in poor conditions
  - CD/DVD storage time typically 20+ yrs
- Physical storage media become obsolete e.g. floppy disks

- Copy data files to new media two to five years after first created
- USB drives
- RAID and NAT

Storage options – Other storage services (2)

Other storage options

- Many organisations are establishing own data storage/backup services
  - Your university or department may have options available e.g. secure backed up storage space
  - Recently, organisations outsource data storage to cloud providers as part infrastructure or Office services
- VPN giving access to external researchers
  - locally managed Dropbox-like services such as ownCloud – sharing files and folders, and ZendTo – Web based file transfer
  - secure file transfer protocol (FTP) server
- Data repository or archive
  - a repository acts as more of a ‘final destination’ for data
  - many universities have data repositories now catering to its researchers

Storage services, Cloud storage

Online or ‘cloud’ services increasingly popular

- GoogleDrive, Dropbox, Microsoft OneDrive etc.
- Accessible anywhere
- Background syncing
- Mirror files
- Mobile apps available
- Very convenient
- Everyone uses them, and that’s ok BUT precautions must be taken
  - Consider if appropriate, as services can be hosted outside the EU (GDPR for personal data)
  - Encrypt anything sensitive or avoid services altogether
    - Key management (key escrow) aspects
- National NREN services:
  - SURFDrive is a file exchange services for NL academia and research

Verification and integrity checks

- Ensure that your backup method is working as intended
- Be wary when using sync tools in particular
  - mirror in the wrong direction or using the wrong method, and you could lose new files completely
- Applies to cloud based syncing services too
- You can use checksums to verify the integrity of a backup
  - Also useful when transferring files
  - Checksum is a kind of a files’ fingerprint
  - To be updated when the file changes

Data security strategy

Data security

- Protect data from unauthorised access, use, change, disclosure and destruction
- Personal data need more protection – always keep separate and secure

- Control access to computers and storage
  - use passwords, lock your machine when away from it
  - anti-virus and firewall protection, power surge protection
  - all devices: desktops, laptops, memory sticks, mobile devices
  - all locations: work, home, travel
  - restrict access to sensitive materials e.g. consent forms, patient records
- Control physical access to buildings, rooms, cabinets
- Proper disposal of data and equipment
  - Even reformatting the hard drive is not sufficient

Data Destruction

- When you delete a file from a hard drive, the chances are it’s still retrievable – even after emptying the recycle bin
- Files need to be overwritten (ideally multiple times) with random data to ensure they are irretrievable
- Destructing infected files, drives

Data destruction software

- BCWipe - uses ‘military-grade procedures to surgically remove all traces of any file’
  - Can be applied to entire disk drives
- AxCrypt* - free open source file and folder shredding
  - Integrates into Windows well, useful for single files
- If in doubt, physically destroy the drive using an approved secure destruction facility
- Physically destroy portable media, as you would shred paper

Summary of best practice in data storage and security

- Have a personal backup/storage strategy – original local copy, external local copy and external remote copy
- Copy data files to new media two to five years after first created
- Know your institutional back-up strategy
- Check data integrity of stored data files regularly (checksum)
- Create new versions of files using a consistent, transparent system
- Encrypt sensitive data – crucial if using web to transmit/share
- Know data retention policies that apply: funder, publisher, home institution – and remove sensitive data securely where necessary

Practice: Data Management Plan (DMP)

- Scientific Data Lifecycle
- Use template for DMP construction
- Consider GDPR issues

What is a Data Management Plan?

A brief plan written at the start of a project to define:

- What data will be collected or created?
- How the data will be documented and described?
- Where the data will be stored?
- Who will be responsible for data security and backup?
- Which data will be shared and/or preserved?
- How the data will be shared and with whom?

Why develop a DMP?

DMPs are often submitted with grant applications, but are useful whenever researchers are creating data.

They can help researchers to:

- Make informed decisions to anticipate & avoid problems.
- Develop procedures early on for consistency.
- Ensure data are accurate, complete, reliable and secure.
- Avoid duplication, data loss and security breaches.
- Save time and effort to make their lives easier!

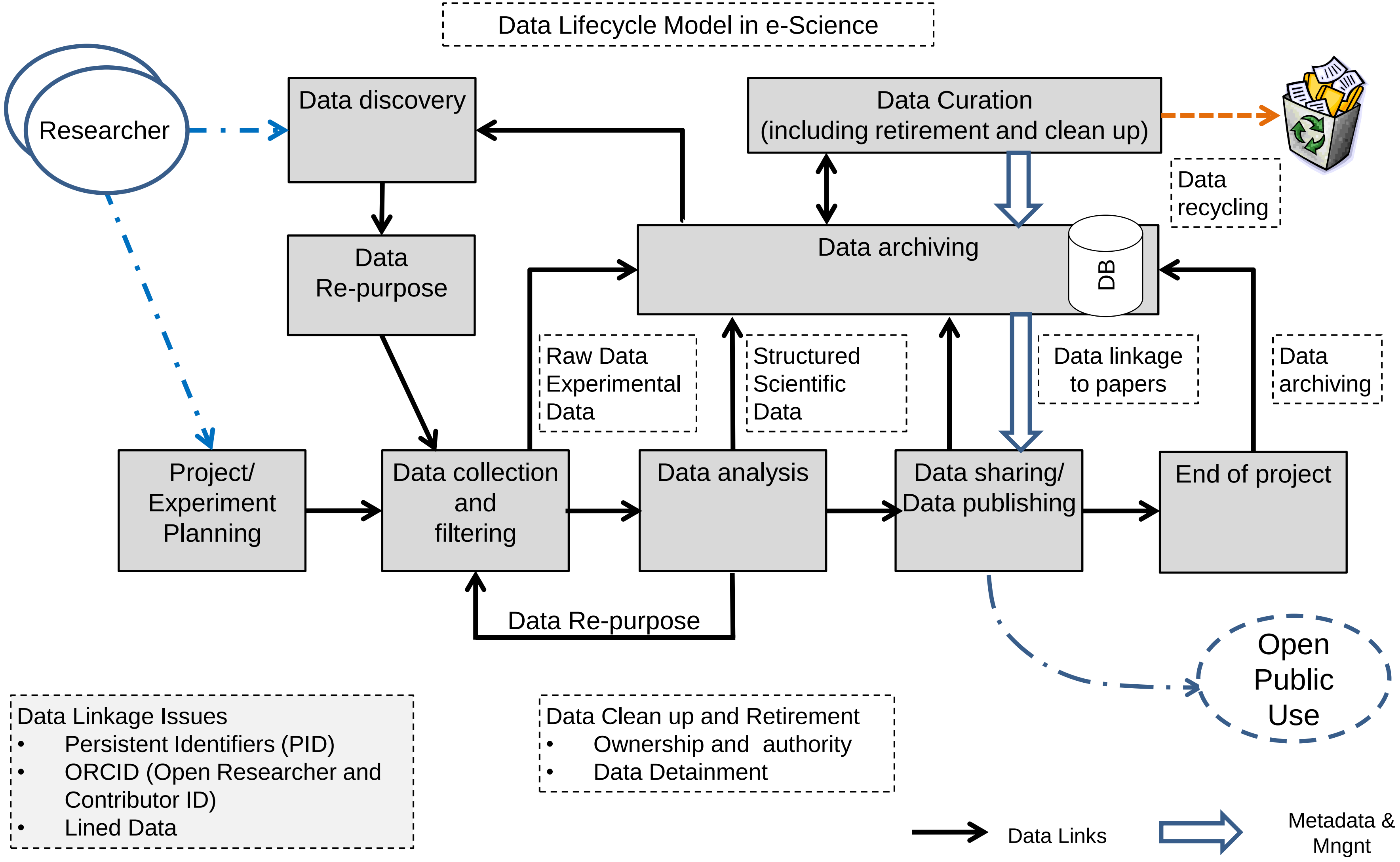
Topics to address in DMPs

- Data collection
- Documentation and metadata
- Ethics and legal compliance
- Storage and backup
- Selection and preservation
- Data sharing
- Responsibilities and resources

Data Lifecycle and RDM

- Scientific Data Lifecycle Model
- DCC Curation Lifecycle Model

Scientific Data Lifecycle Model



Data Lifecycle Model in e-Science

Researcher

Data discovery

Data Re-purpose

Data Curation (including retirement and clean up)

Data recycling

Data archiving

DB

Raw Data Experimental Data

Structured Scientific Data

Data linkage to papers

Data archiving

Project/ Experiment Planning

Data collection and filtering

Data analysis

Data sharing/ Data publishing

End of project

Data Re-purpose

Open Public Use

Data Linkage Issues
- Persistent Identifiers (PID)
- ORCID (Open Researcher and Contributor ID)
- Lined Data

Data Clean up and Retirement
- Ownership and authority
- Data Detainment

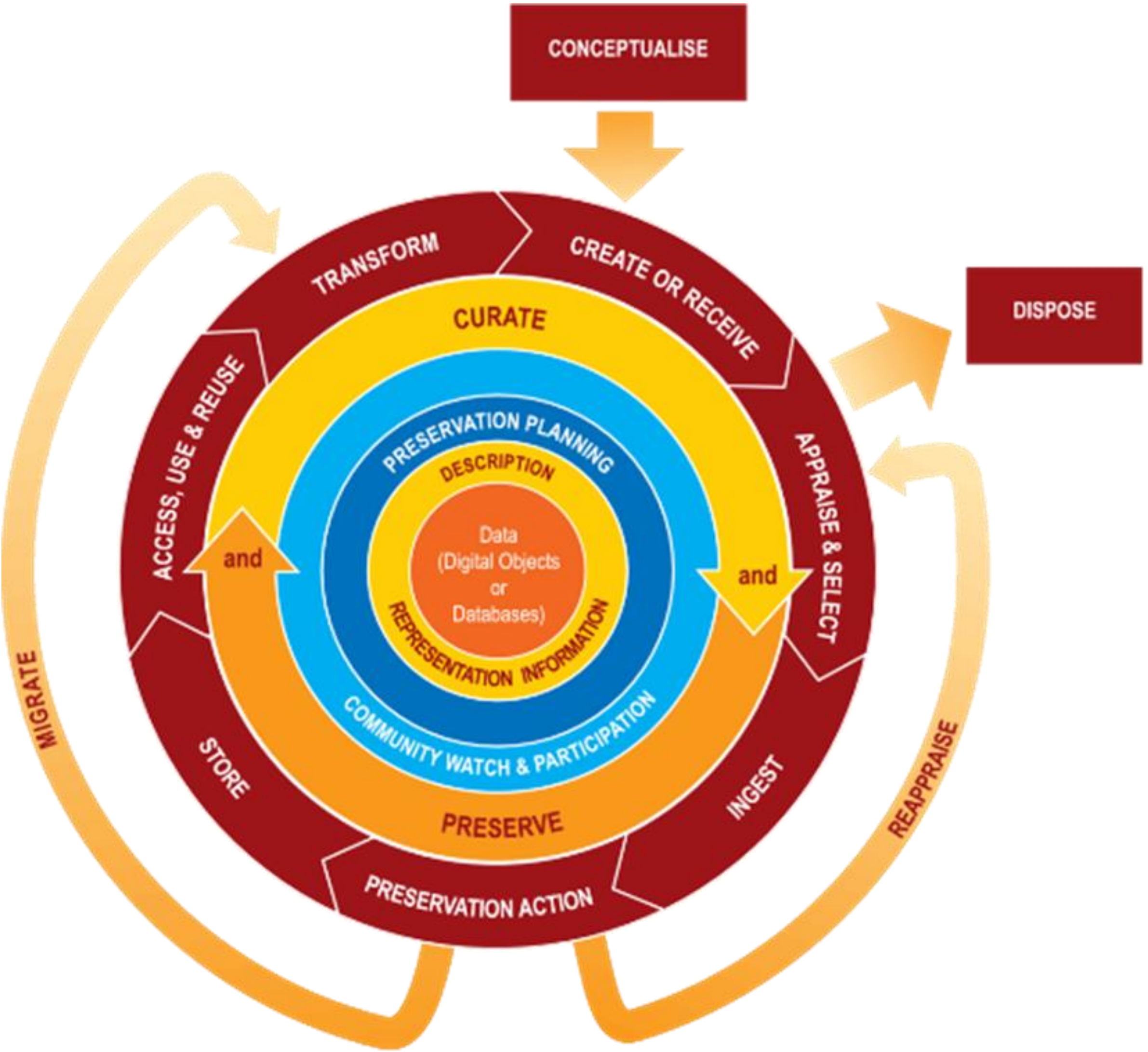
Data Links

Metadata & Mngnt

DCC Curation Lifecycle Model - Actions

Three sets of actions:

- Sequential Actions (7+1): key actions needed as data move through their lifecycle
- Occasional Actions (3): only occur when special conditions are met, but they do not apply to all data
- Full Lifecycle Actions (4): apply to all stages in the lifecycle



The DCC Curation Lifecycle Model is based on the OAIS Reference Model

- OAIS = Open Archival Information System (pictured)
- OAIS is a model that defines a generic framework for building a digital archive

Data Licensing

- Creative Commons licenses are not always suitable for data because data have different IPR than generic digital content
  - http://creativecommons.org/licenses/

cc BY   cc BY SA   cc BY ND   cc BY NC   cc BY NC ND   cc BY NC SA

- Open Data Commons have specific licenses for data that conform to the Open Knowledge Foundation's definition of Open Data: http://opendatacommons.org/guide/
- They have 2 basic options:
  - Public Domain: puts all the materials in the public domain
  - Share-Alike (plus Attribution): similar to the Creative Commons Attribution Share-Alike license
- Also see the following guides:
  - http://datalib.edina.ac.uk/mantra/preservation.html: an online learning unit from Mantra on "Sharing, preservation, and licensing", go to slides 15-17
  - http://infteam.jiscinvolve.org/wp/2012/10/09/opendatalicensing: Open Data licensing animation video
  - http://opendefinition.org/guide/data/

Practical exercise

- Data Management Plan checklist
  - University of Amsterdam DMP template
- See GoogleDrive folder for DMP and DGP Exampes

Discussion

- Discussion questions and comments

Acknowledgement

- This work is supported by the ERASMUS+ MATES project
- The work is committed to the Open Source under Creative Commons 4.0 CC BY License

Co-funded by the
Erasmus+ Programme
of the European Union

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.