Open Science and Research Data Management

Core principles and Best practices

Dr. Yuri Demchenko
SLICES-RI, University of Amsterdam

SLICES-CONVERGE Tutorial
27 February 2024

Outline

A. Open Access, Open Data, Open Science, EU policy on Open Science

B. Research Data Management factors

C. Data Management basics

- Creating documentation and metadata, metadata for discovery
- Best practices in Research Data Management
- Backing up your data
- Data Security

D: Data Management Plan (DMP)

- DMP example
- Future development: Machine actionable DMP (maDMP)

Research Data Management - Part 1

- Open Access, Open Data, Open Science
  - EU policy on Open Access and Open Science
- PID, ORCID
- FAIR data principles

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

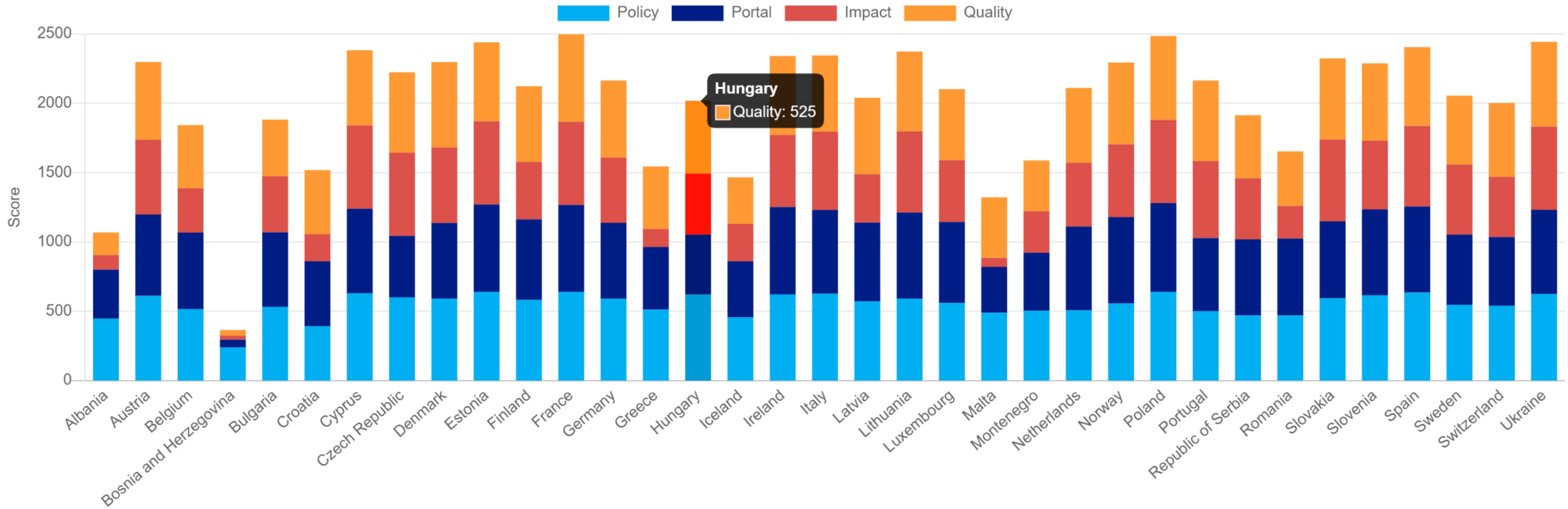
EU policy on Open Research Data – Since H2020

- 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.
  - Creative Common (CC) Open Source license

Horizon 2020 Open Research Data (ORD) Pilot - Historical

- 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.

Open Data in Europe – data.europa.eu



- Open Data Maturity Report 2023 - https://data.europa.eu/en/publications/open-data-maturity/2023
- European data: The official portal for European data - data.europa.eu - https://data.europa.eu/en
  - 1,674,520 Datasets, 183 Catalogues, 36 Countries, 216 Data stories
- Data Academy https://data.europa.eu/en/academy
  - 11 Courses

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
- Toll Access, as a traditional method, is different from Open 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***
  - *Is applicable for educational use*
- 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.

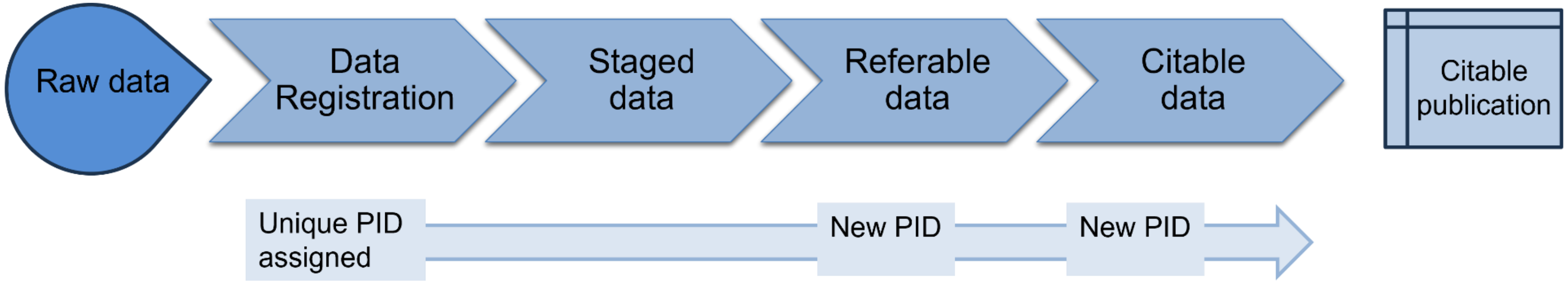
Self-archiving services

- 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.
  - Similar services: bioRxiv, SocArXiv, PsyArXiv
- Figshare - https://figshare.com/
  - Data repository, datasets citation, research workflows to support reproducibility
- PubMed Central (PMC) - https://www.ncbi.nlm.nih.gov/pmc/
  - A free digital repository that archives publicly accessible full-text scholarly articles in the biomedical and life sciences.

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

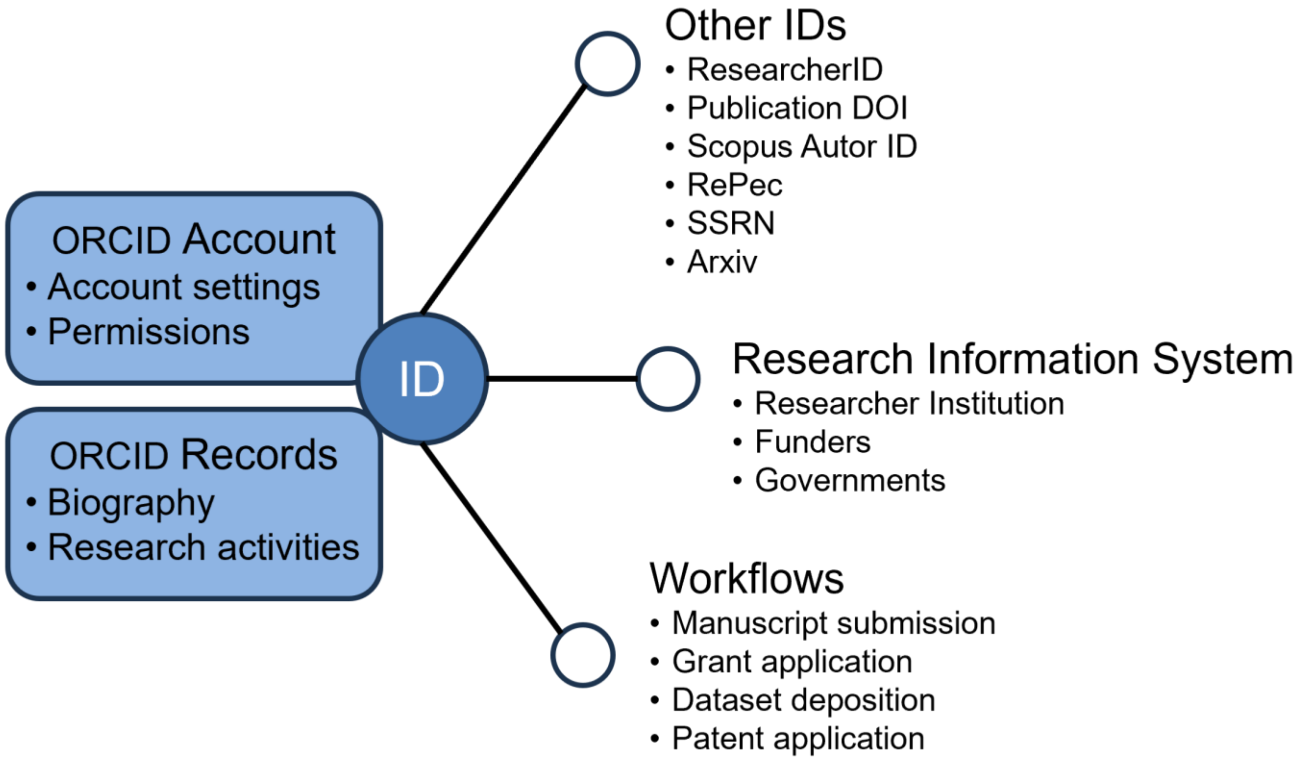


PID based data linking

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
- Permissions

ORCID Records
- Biography
- Research activities

Other IDs
- ResearcherID
- Publication DOI
- Scopus Autor ID
- RePec
- SSRN
- Arxiv

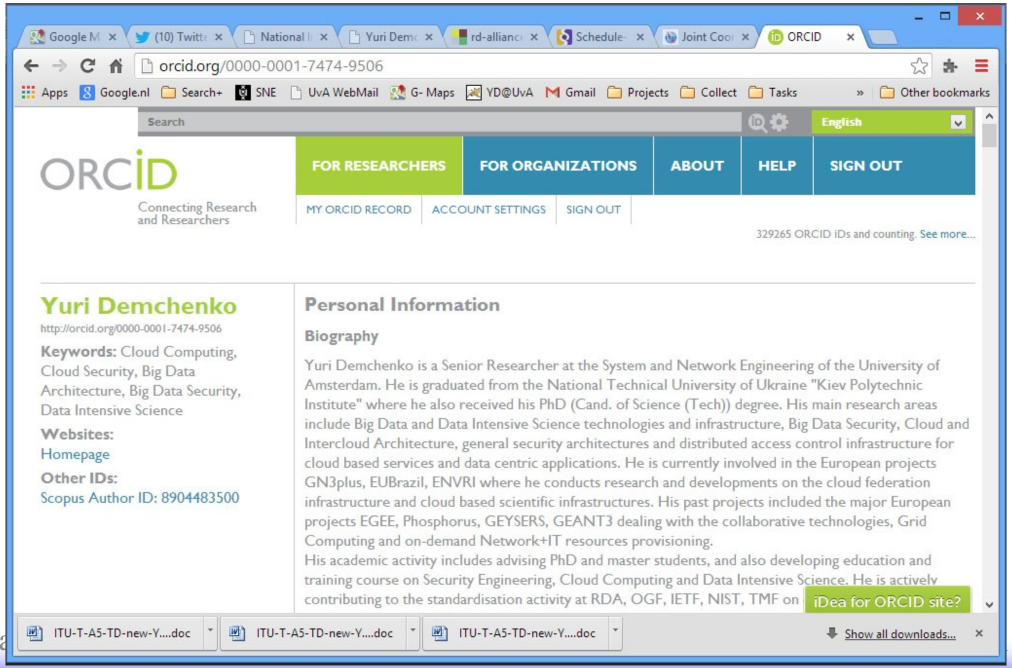
Research Information System
- Researcher Institution
- Funders
- Governments

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

- 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 – October 2020
  - Live ORCID IDs – 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;

- https://fairdataforum.org/

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 PID – 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

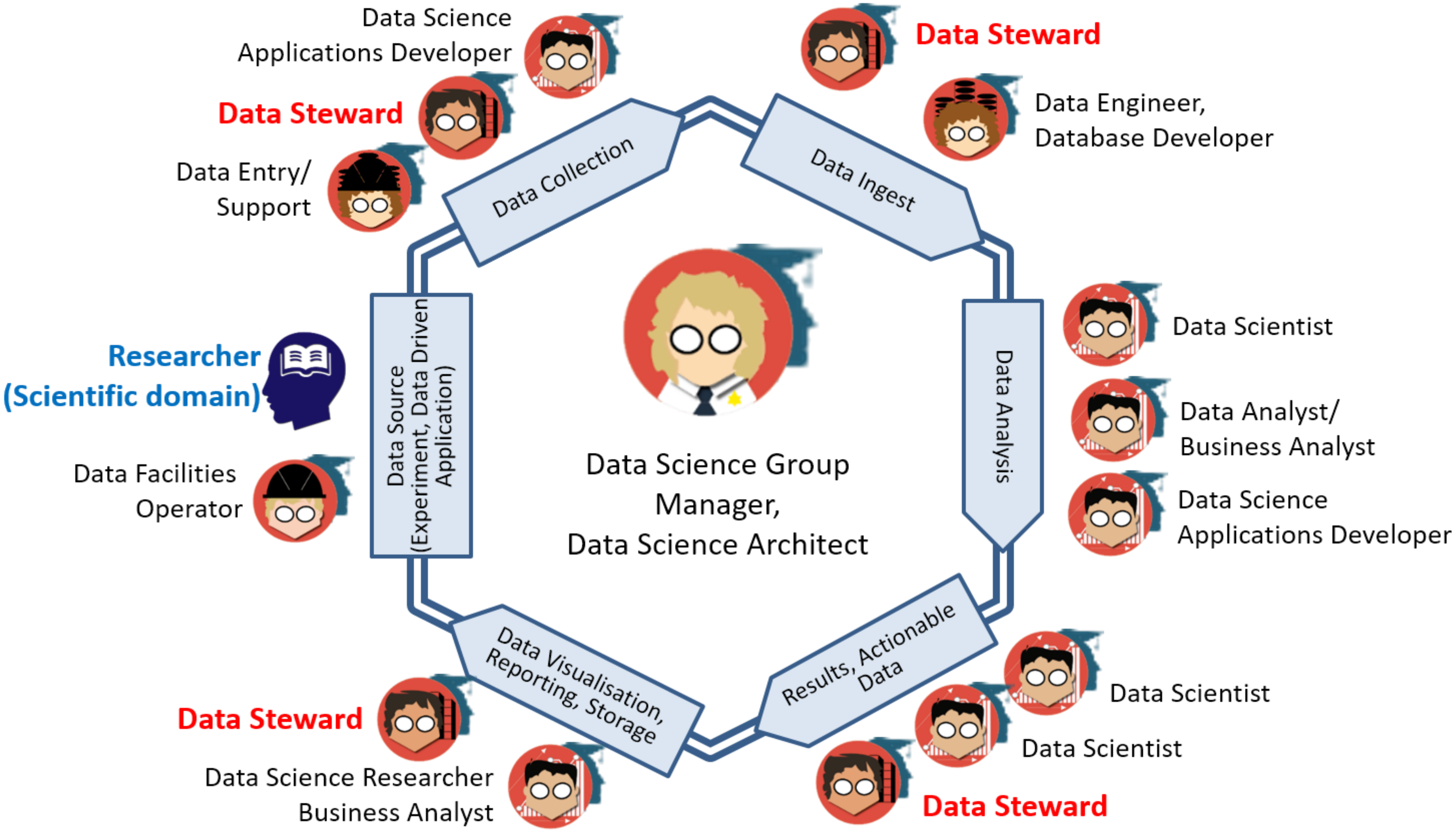
- Consequently related competences from Data Stewards are needed

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 Stewards,** 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
  - **Data Steward** as a link between researchers and ICT department

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

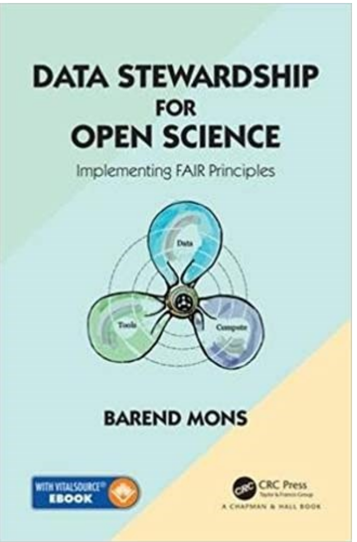
- FAIR Initiative by Dutch Techcentre for Life Science (DTLS) – Prof. Barend Mons
  - Part of Horizon 2020 Programme
- FAIR Principles for research data:
  Findable – Accessible – Interoperable - Reusable
- Data Stewards as a **key bridging role** between Data Scientists as (hard)core data experts and 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

Data Stewards – Job market review

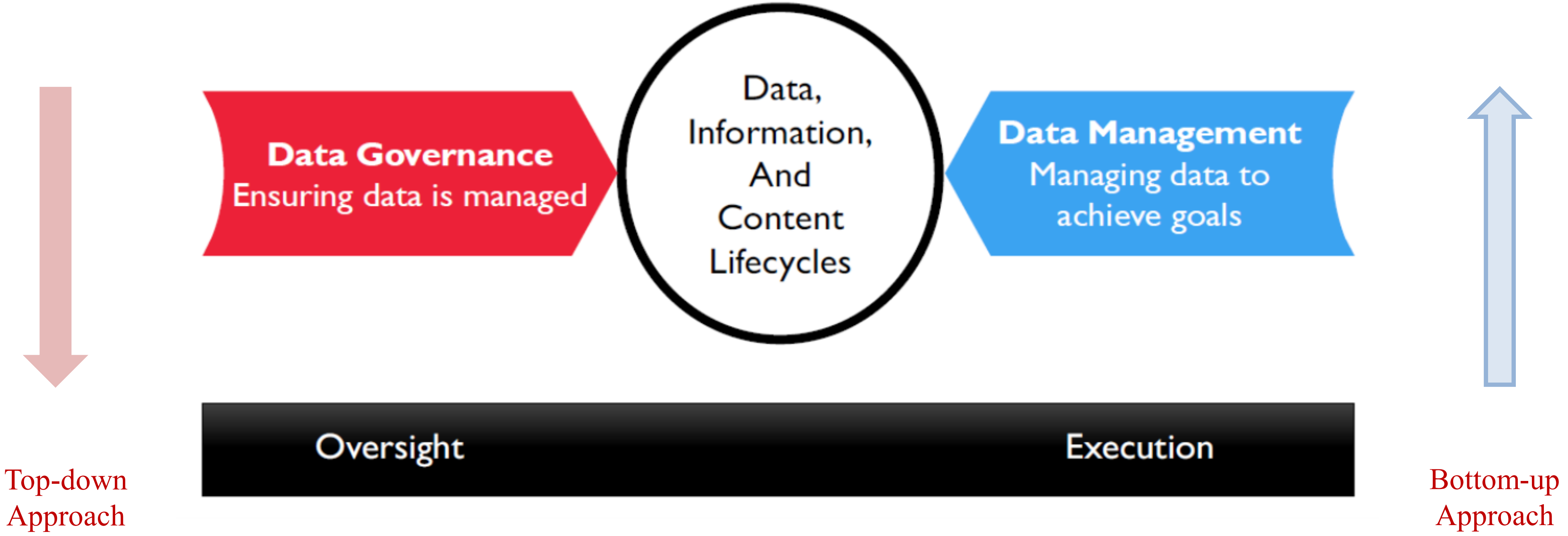
- Date 30 August – 1 September 2020 (revisited 2022)
- Indeed.com – NL, UK, DE, USA
- Days open: >50% more than 30 days
- Data Steward and related vacancies
  - NL – 51, UK – 30+, DE ~20, US – 300+
  - Key skills snapshot
- Sample vacancies detailed analysis
  - NL, UK – 12, US - 6

Data Management Practices

Data management practices targeted for researchers and technical staff supporting data handling during the research lifecycle.

- Working with data: the following aspects to be defined:
  - file naming and formatting
  - data formats and software
  - file transfers, file sharing and remote access
  - version control
- Administering data/datasets (produced in the research process)
  - documentation and metadata
  - back-ups
  - access controls
  - security
- Storing and sharing
- Ethical and legal aspects of data handling and data ownership

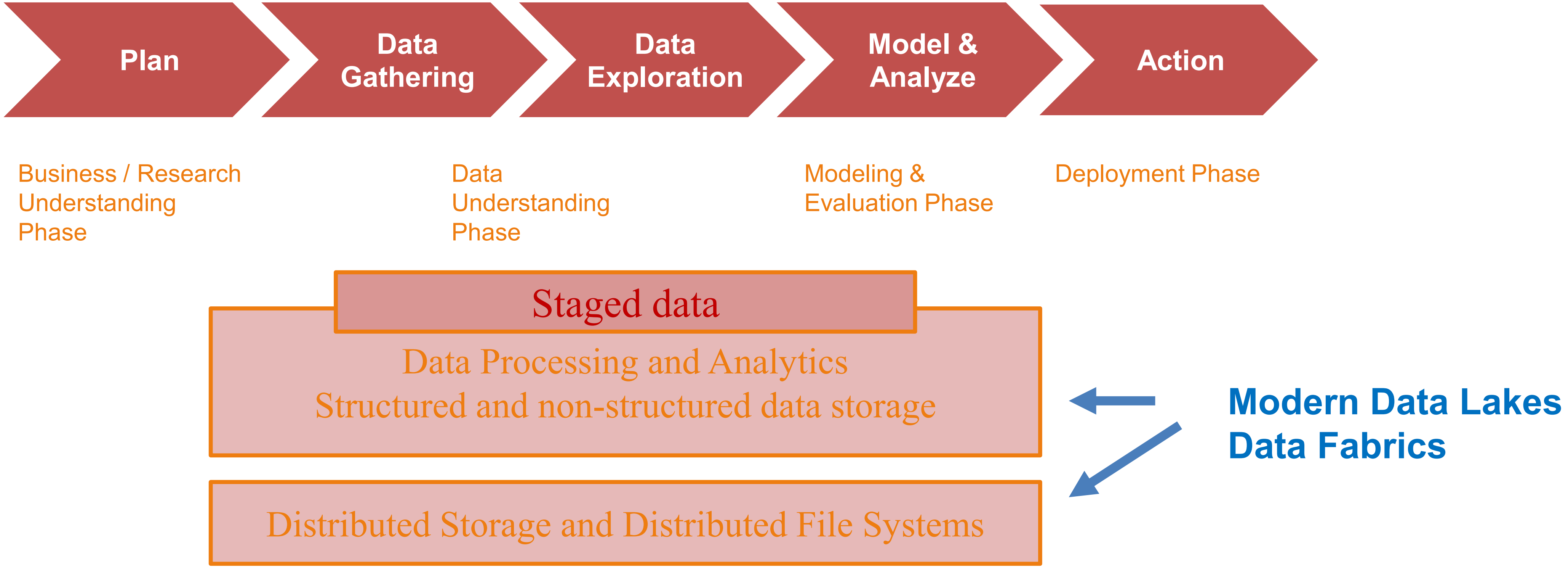
Data Governance and Data Management



- Data Governance – Top down
- Data Management – Bottom-up

The Analytics Project Flow and Data Management

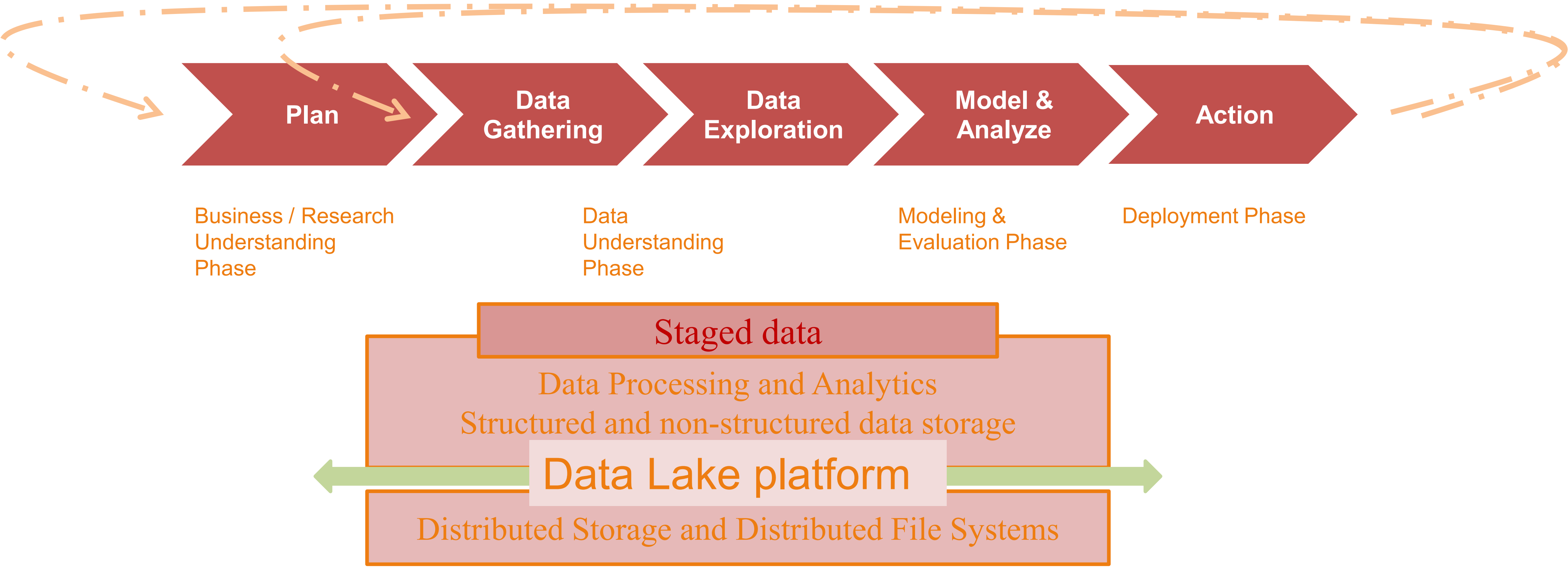
- From idea to action – Business View
- Consider documenting all data stages



Plan
Data Gathering
Data Exploration
Model & Analyze
Action
Business / Research Understanding Phase
Data Understanding Phase
Modeling & Evaluation Phase
Deployment Phase
Staged data
Data Processing and Analytics
Structured and non-structured data storage
Distributed Storage and Distributed File Systems
Modern Data Lakes
Data Fabrics

The Analytics Project Flow and Data Management

- From idea to action – Business View
- Consider iterative improvement process



Plan
Data Gathering
Data Exploration
Model & Analyze
Action
Business / Research Understanding Phase
Data Understanding Phase
Modeling & Evaluation Phase
Deployment Phase
Staged data
Data Processing and Analytics
Structured and non-structured data storage
Data Lake platform
Distributed Storage and Distributed File Systems

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

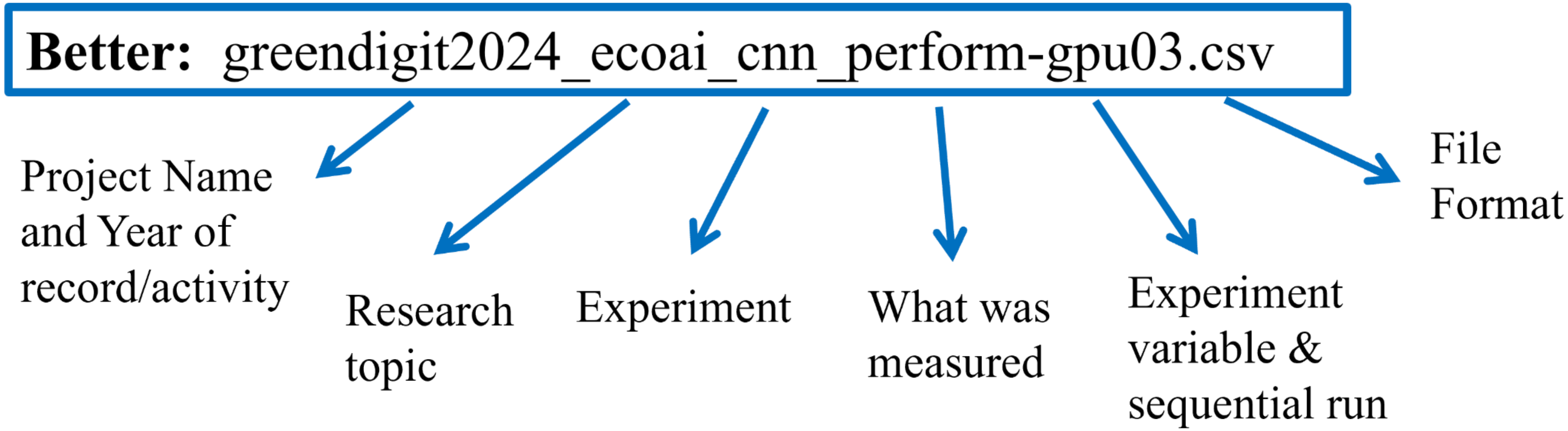
- Clear, descriptive, and unique file names may be important later when your data file is combined in a directory or FTP site with your own data files or with the data files of other investigators
- File name = principal identifier of file
  - use logical naming i.e. easy to identify and retrieve the 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: greendigit2024_ecoai_cnn_perform-gpu03.csv

Project Name and Year of record/activity

Research topic

Experiment

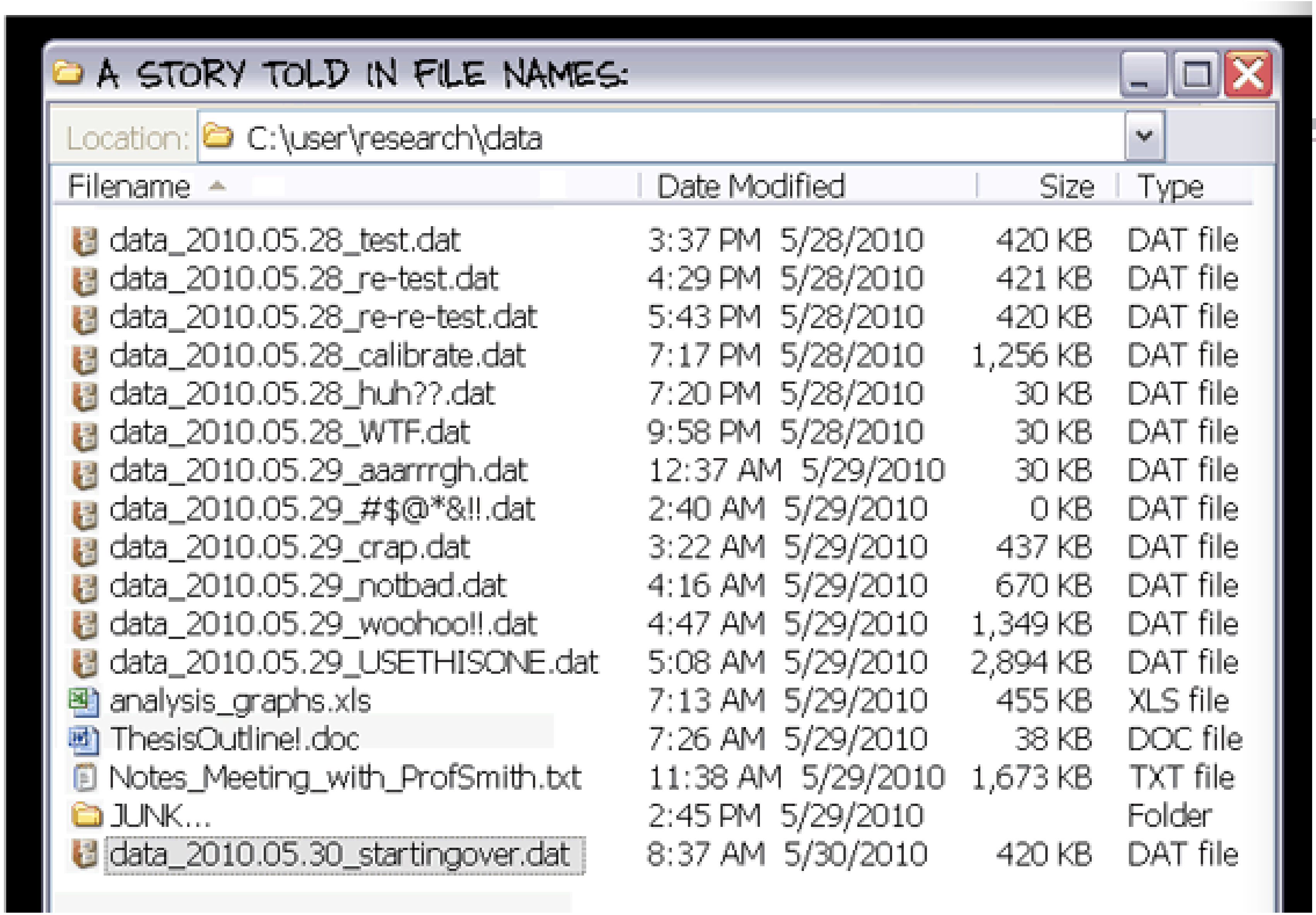
What was measured

Experiment variable & sequential run

File Format

Alternative to file naming approach

- Files that are not expected to be used outside of the system or directory structure
  - Software packages
  - GitHub directories
  - Directory related metadata, access control, etc
- Reserved filenames
- However, Java, Python class names
  - Variables in SQL databases



A STORY TOLD IN FILE NAMES:

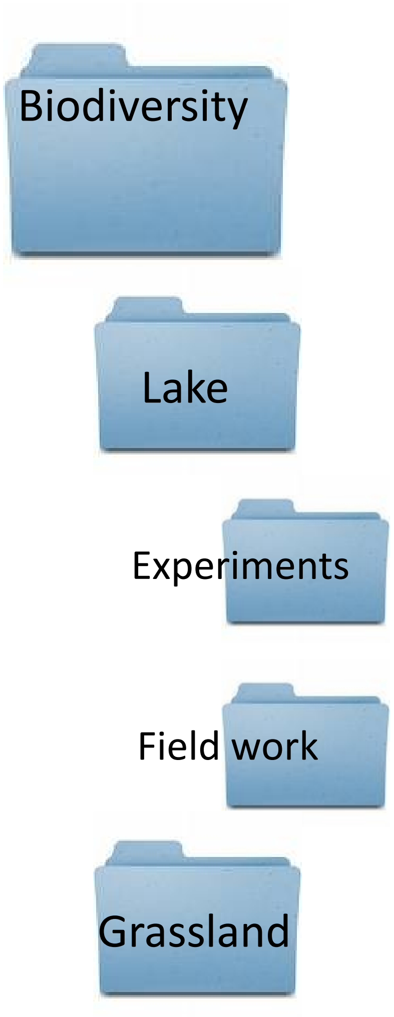
Location: C:\user\research\data

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

Make sure your directory system is logical and efficient



Biodiversity

Lake

Experiments

Field work

Grassland

Biodiv_H20_heatExp_2005_2008.csv
Biodiv_H20_predatorExp_2001_2003.csv
….

Biodiv_H20_planktonCount_start2001_active.csv
Biodiv_H20_chla_profiles_2003.csv
…

Data Version Control Principles and Benefits

- Traceability:
  - Version control systems track changes made to a dataset over time.
- Documentation:
  - Many version control systems allow (or even require) users to provide comments or "commit messages" when changes are made.
- Backup Strategy:
  - It is essential to have a backup strategy in place that would provide a snapshot of your project and data at a certain moment of time or at a certain project stage.
- Reproducibility:
  - Scientific research demands reproducibility. With version-controlled data, researchers can refer to or revert to specific versions of a dataset that were used in particular analyses, ensuring results can be replicated by others.
- Retention and Archiving Policy:
  - Decide on how long historical data versions will be retained and when/how archiving will take place.
- Accountability, Auditing and Compliance:
  - In with strict regulatory oversight, having a clear record of data changes is essential for compliance.
- Collaboration and Consistency:
  - Version control systems ensure that everyone involved in a project is working with the same set of data.

Version Control

- 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_2010_04_01; // good option for files ordering
  - HealthTest_06-04- 2008; // bad option for files ordering
  - 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-v02.doc – new version by document owner

- GoogleDoc or Word – bad for tracking versioning

Managing your personal documents and data collection

Finding necessary document in your personal archive

- It is important to distinguish what kind of resource you stored, when it was published, and possibly the authors.
- Using built-in search on your computer may not immediately help you.
- Searching on filename or directory is the quickest but it can be effective if you have created a clear directories structure and used a helpful file naming approach.
- Searching on the content is slow, works only for textual search patterns, and may produce too many hits.

It is a good practice to add additional information or provide initial sorting of downloaded information.

- Place downloaded file in appropriate directory, e.g. under a project directory tree, possibly create a special sub-directory named like “[project/topic]-refs[year-month]” where “refs” means references and adding month-year when you created references collection.
- Change/extend file name if necessary, to contain (1) type of resource: paper, article, blog, webpage, standards, industry review, (2) date year and optionally month, (3) content topic.

Applying the proposed recommendations will require certain discipline, but it will pay back in the future by saving time in finding necessary information and tracking down its origin or source.

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

- Generally defined, metadata are data about data.
- Provide 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
- Benefits of best practices in metadata management
  - Improves Data Discoverability
  - Enhances Data Quality and Reliability
  - Facilitates Data Interoperability
  - Supports Data Compliance and Governance
  - Enables Better Data Analysis and Utilization
  - Assists in Long-term Preservation of Data
  - Supports Knowledge Management and Sharing

International Metadata Standards

Metadata standards are digital containers for structured information about a data set

- **Dublin Core (DC)**: A simple and widely used metadata standard that can describe a variety of digital resources.
- DataCite   Metadata for research data and other research outputs to make them discoverable to the research community - http://schema.datacite.org/
- Data Documentation Initiative (DDI): Primarily used for describing data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences. DDI facilitates data comparison, sharing, and interoperability.
- Metadata Object Description Schema (MODS): Developed by the Library of Congress, MODS is a bibliographic element set that can be used for a variety of purposes, including library catalogs, archives, and digital libraries. It is richer in detail than Dublin Core.
- Metadata Encoding and Transmission Standard (METS) is a standard for encoding descriptive, administrative, and structural metadata regarding digital library objects (such as books, journals, newspapers, and archival materials). Developed by the Digital Library Federation and maintained by the Library of Congress.
- Preservation Metadata Maintenance Activity (PREMIS): A standard focused on preserving digital objects and ensuring their long-term usability, defining metadata necessary for preservation activities.

Metadata Standards by discipline

Biosciences:

- Darwin Core (DwC)
- Minimum Information About a Microarray Experiment (MIAME)
- Biological Data Profile (BDP)

Earth and Environmental Sciences:

- Federal Geographic Data Committee (FGDC)
- Climate and Forecast (CF) Metadata Conventions
- ISO 19115: An international standard for describing geographical information and services.
- Ecological Metadata Language (EML)

Health and Medicine:

- Health Level Seven (HL7)
- DICOM (Digital Imaging and Communications in Medicine)

Genomics:

- Minimum Information About a Sequence (MIxS): Standards for describing sequence data and associated environmental context.

Education:

- Learning Object Metadata (LOM):

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/

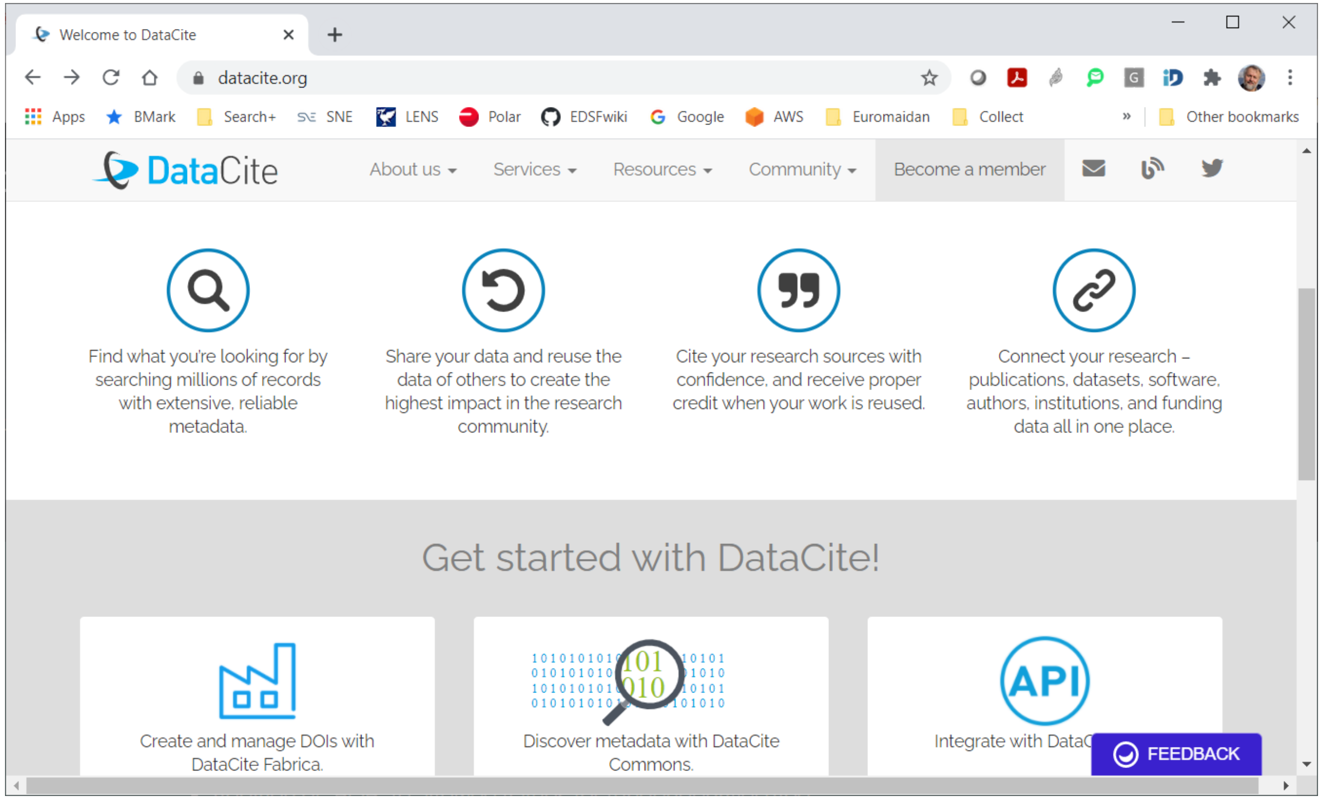
[DCTERMS]_http://dublincore.org/documents/dcmi-terms/

Example: Dublin Core metadata

```
<textbook>
   <dc:creator>Esterhuis, Rudy.</dc:creator>
   <dc:date>c2019</dc:date>
   <dc:description>Textbook and reading references for the course BDIT4DA</dc:description>
   <dc:format>137 pp, PDF</dc:format>
   <dc:identifier>Article No 974306176301</dc:identifier>
   <dc:language xsi:type="http://purl.org/dc/terms/ISO639-2">eng</dc:language>
   <dc:publisher>Triarty Press</dc:publisher>
   <dc:subject>Big Data Infrastructure Technologies</dc:subject>
   <dc:subject xsi:type="http://purl.org/dc/terms/LCSH">Textbook</dc:subject>
   <dc:title>Sustainable Architecture Design Principles</dc:title>
   <dc:type>Text</dc:type>
</textbook>
```

Metadata Standards: DataCite http://datacite.org/

- Services
  - Assign DOIs
  - Metadata search
  - Event data
  - Profiles
  - re3data
  - Citation formatter
  - Statistics
  - Service status
  - Content negotiation
- Resources
  - Metadata schema
  - Support



Find what you're looking for by searching millions of records with extensive, reliable metadata.

Share your data and reuse the data of others to create the highest impact in the research community.

Cite your research sources with confidence, and receive proper credit when your work is reused.

Connect your research – publications, datasets, software, authors, institutions, and funding data all in one place.

Get started with DataCite!

Create and manage DOIs with DataCite Fabrica.

Discover metadata with DataCite Commons.

Integrate with DataC

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, °F… lots of different names

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/ Time 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.

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

FAIR Core for EOSC – Tools First Release Announce

1. EOSC Research Discovery Graph (RDGraph) to deliver advanced Discovery tools across EOSC resources and communities;
2. EOSC PID Graph (PIDGraph) to improve the way of interlinking research entities across domains and data sources on the basis of persistent identifiers (PIDs);
3. EOSC Metadata Schema and Crosswalk Registry (MSCR) to support publishing, Discovery and access of metadata schemas and provide functions to operationalize metadata conversions by combining crosswalks;
4. EOSC Data Type Registry (DTR) to provide user friendly APIs for metadata imports and access to different data types and metadata mappings;
5. EOSC PID Meta Resolver (PIDMR) to offer users a single PID resolving API in which any kind of PID can be resolved through a single, scalable PID resolving infrastructure;
6. EOSC Compliance Assessment Toolkit (CAT) to support the EOSC PID policy compliance and implementation;
7. EOSC Research Activity Identifier Service (RAiD) to mint PIDs for research projects, allowing to manage and track project related activities;
8. EOSC Research Software APIs and Connectors (RSAC) to ensure the long-term preservation of research software in different disciplines;
9. EOSC Software Heritage Mirror (SWHM) to equip EOSC with a mirror of the Software Heritage universal source code archive.

EOSC Metadata Schema and Crosswalk Registry (MSCR)

https://faircore4eosc.eu/eosc-core-components/metadata-schema-and-crosswalk-registry-mscr

- The EOSC MSCR supports registering schemas/crosswalks hosted elsewhere as well hosting them in the repository
- It offers basic data management support: PIDs, metadata, versioning and provenance information.
- Supports a GUI for visually creating crosswalks between metadata schemas
  Provides an API and guidelines for organisations to register and maintain metadata schemas and crosswalks
- When registering metadata schema users are able to provide detailed data-type information for fields and attributes using the DTR
- Provides a (meta-)data interoperability service that facilitates conversion between metadata schemas
- The metadata schema and crosswalk registration process and governance is aligned with the EOSC Provider and Resource onboarding process (currently operated by EOSC Future)
- The MSCR will be integrated with all relevant EOSC-Core services: AAI, monitoring and helpdesk

Discovery Level Metadata

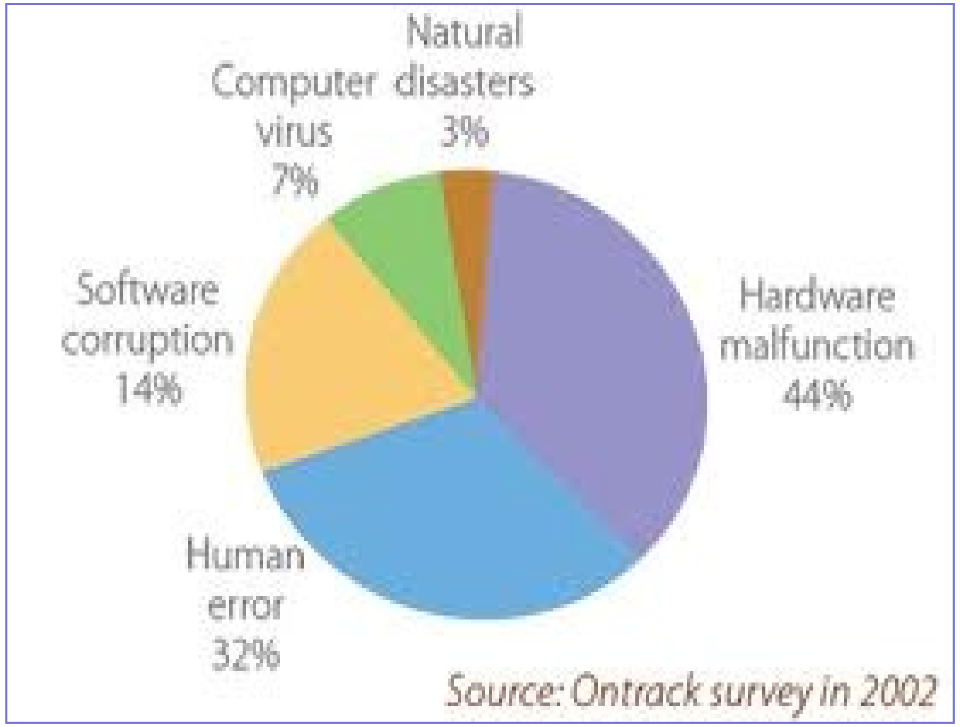
- A data set description (metadata) that provides information to determine if a particular data set meets the users' needs.
- Typically provides essential information to enable a user to find out if a particular dataset exists, the data's location, and ownership, and how to obtain further information.
- 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.

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

- All digital media are fallible
- Optical (CD, DVD) & magnetic media (hard drives, tape) degrade – lifespan even lower if kept in poor conditions
- Physical storage media become obsolete e.g. floppy disks
- Copy data files to new media two to five years after first created

Other storage options

- Your university or department may have options available e.g. secure backed up storage space
- VPN giving access to external researchers
  - locally managed Dropbox-like services such as ownCloud and ZendTo
  - 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
- SURFnet provides Research Data Storage service
  - University of Amsterdam is still in the process of establishing own data storage/backup services

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 – Online or Cloud

Online or ‘cloud’ services increasingly popular

- GoogleDrive, DropBox, Microsoft OneDrive, SURFdrive.nl, national services, etc.
- Accessible anywhere
- Background syncing
- Mirror files
- Mobile apps available
- Convenient
- Everyone uses them, and that’s ok BUT precautions must be taken
  - Consider if appropriate, as services can be hosted outside the EU (remember GDPR and personal data)
  - Encrypt anything sensitive or avoid services altogether
- SURFDrive is a file exchange services for NL academia and research

Verification and integrity checks

- Ensure that your backup method is working as intended
  - Possible issue: Long filenames and long paths: Typically you can archive but not restore
- Be wary when using sync tools in particular
  - Mirror/copy in the wrong direction or using the wrong method, and you could lose new files completely
- Applies to online DropBox-like 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
  - Also about potentially dangerous emails

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

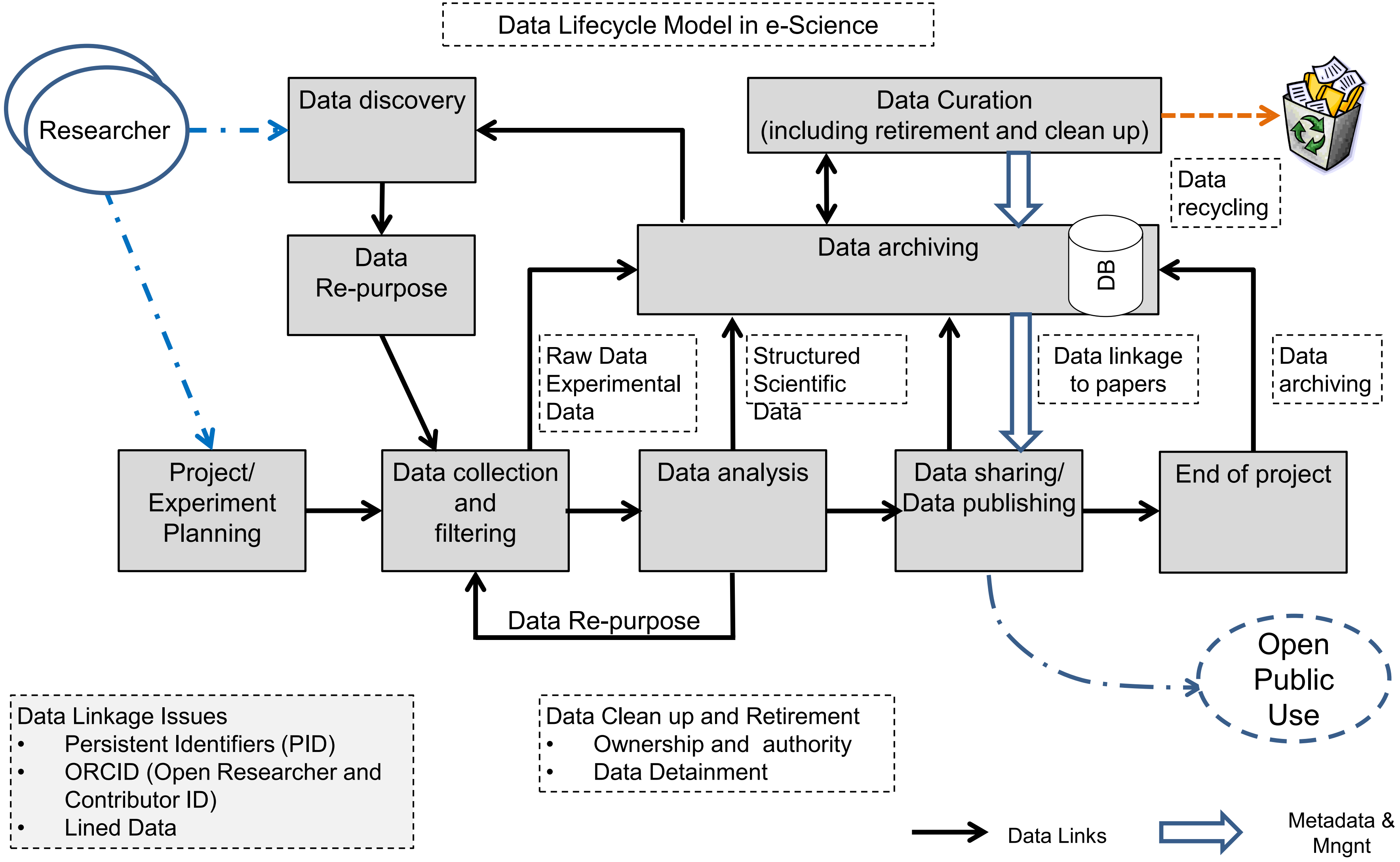
Data Management Plan (DMP)

- Research Lifecycle and Data Management
- Data Management Plan structure
- Using template and online tools for DMP construction
- Future development:

Research Lifecycle and Data Lifecycle

- General Research Lifecycle – from Experiment planning to results publication
- SLICES Data Lifecycle for reproducible research

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 Links
Metadata & Mngnt

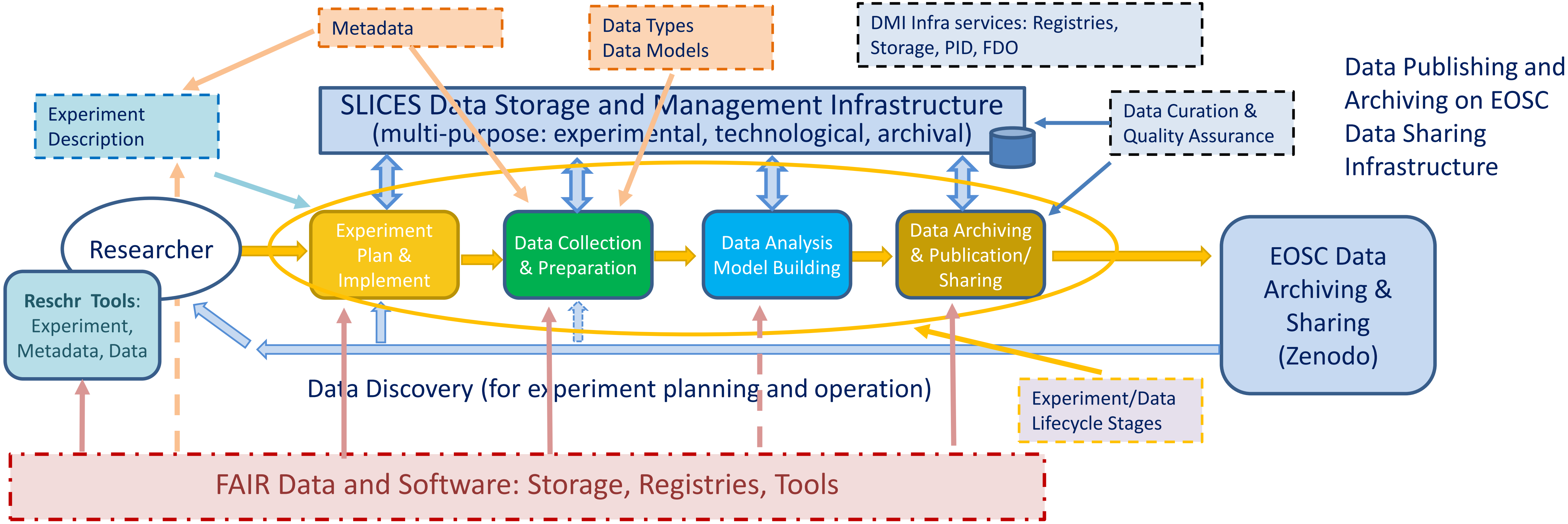
Data Linkage Issues

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

Data Clean up and Retirement

- Ownership and authority
- Data Detainment

SLICES Experimental Data Lifecycle Model and Dataflow


Metadata
Data Types
Data Models
DMI Infra services: Registries, Storage, PID, FDO
Data Publishing and Archiving on EOSC Data Sharing Infrastructure
Experiment Description
SLICES Data Storage and Management Infrastructure
(multi-purpose: experimental, technological, archival)
Data Curation & Quality Assurance
Researcher
Reschr Tools: Experiment, Metadata, Data
Experiment Plan & Implement
Data Collection & Preparation
Data Analysis Model Building
Data Archiving & Publication/ Sharing
EOSC Data Archiving & Sharing (Zenodo)
Data Discovery (for experiment planning and operation)
Experiment/Data Lifecycle Stages
FAIR Data and Software: Storage, Registries, Tools

- **Each Data Lifecycle stage** – experiment, data collection, data analysis, and finally data archiving, works with own **data set,** which must be **linked**.
  - All data sets need to be stored and possibly re-used in later processes.
- Many experiments and research require already existing datasets that will be available in SLICES data repositories or can be obtained/discovered in EOSC data repositories

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 required to be 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!

Data Management Plan Structure

1. Data Summary

    General Description: An overview of the data to be collected, processed, or generated.

    Purpose: The rationale behind data collection or generation and its relation to the objectives of the project.

2. FAIR Data principles support

    Making Data **Findable**, Including Provisions for Metadata: Details on how data will be made findable, such as metadata standards to be used, and the creation of a data catalogue if applicable.

    Making Data **Accessible**: Information on data storage, sharing, and access.

    Making Data **Interoperable**: Discusses the use of standards to ensure data compatibility and integration with other datasets.

    Increase Data **Reuse** (through clarifying licenses): Describes the licensing and any restrictions on data reuse.

3. Allocation of Resources

    Costs and Resources: Details on the budget and resources allocated for data management.

    Responsibility and Resources: Information on who will be responsible for data management tasks.

4. Data Security

    Data Security: Outlines measures for ensuring data security, confidentiality, and integrity.

5. Ethical Aspects

    Ethical Aspects: Addresses any ethical issues, including compliance with legislation and codes of conduct.

6. Data Sharing and Open Access

    Data Sharing and Access: Describes the strategy for data sharing, including timelines for data release and access conditions.

    Open Access: If applicable, details the open access provisions for the data, aligning with the EC's open access policies.

7. Archiving and Preservation

Future Developments: Machine Actionable DMP (maDMP)

- maDMP – challenges
- Principles and best practices: 10 principles of maDMP
- RDA Common Format for maDMP
  - Implementations
- maDMP for Scientific Workflows and Experimental Research
- maDMP and AI-aware DMP

maDMP for AI/ML projects - Challenges

- maDMP and automated provisioning as a part of scientific workflow
  - Including data models and resources (management)
- Data volume, distributed data, distributed computation and scalability
- ML model is a subject for data management and consistency
- Reproducibility and models sharing – to be supported by data linkage
- Data provenance and Versioning/Lineage
- Data Quality and model bias influenced by data changes/data skew
- Ethics and Privacy (personal data protection)
- Security and compliance aspects for data to be part of maDMP
- Continuous maDMP monitoring and monitoring services defined in maDMP

Principles of maDMP

10 principles of maDMP (2019)

1. Workflow integration
2. Automatic action on behalf of stakeholders
3. DMP for machines
4. Description both human readable and machine-readable
5. Use of PID and controlled vocabulary
6. Common Data Model
7. User interface and API for human and machines
8. Support for DMP evaluation and monitoring
9. Continuous update
10. Public availability

[ref] Tomasz Miksa, Stephanie Simms,Daniel Mietchen, Sarah Jones, Ten principles for machine-actionable data management plans. PLOS Compoutational Biology, 2029 https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006750

Demand of Experimental Automation and AI/ML projects

1. Standardisation
2. Interoperability
3. Automation and Workflow integration
4. FAIR actionable
5. Quality assurance
6. Privacy and Security Compliance
7. Ethical use of data (auto compliance)
8. Reproducibility
9. Flexibility
10. Accessibility
11. Transparency
12. User interface
13. Collaboration and sharing
14. Sustainable DM practices

Discussion

- Discussion questions and comments