Research Data Management
and FAIR Data Principles

Yuri Demchenko

SLICES Summer School

13-15 June 2023, Oulu, Finland

Outline

Context: Data Governance and Data Management

- Enterprise Data Governance and Data Management

A. Open Access, Open Data, Open Science

B. Research Data Management factors

C. Data Management basics

- Creating documentation and metadata, metadata for discovery
- Backing up your data

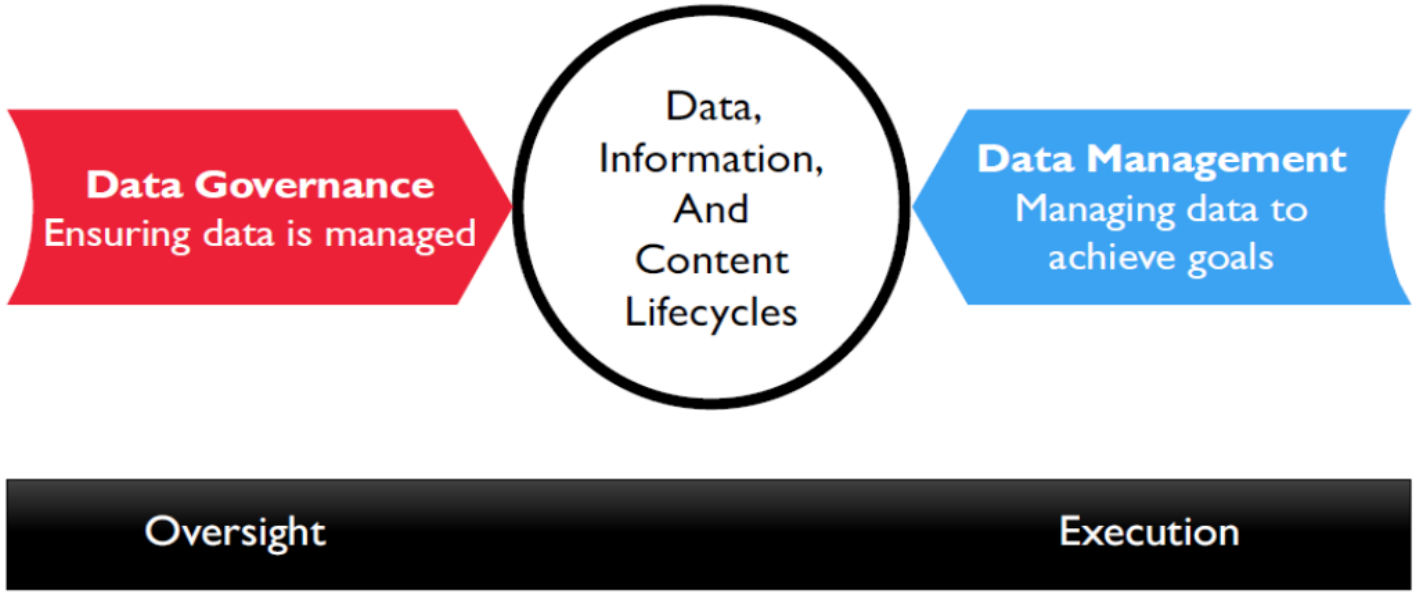
D. Responsible Data Use (citation, copyright, data restrictions)

- Handling sensitive data, Ethical issues

Workshop Materials

- https://drive.google.com/drive/folders/1mfoZs3OXOx_Klhy1r6-YVXlW_4MtadFh?usp=sharing

Data Governance and Data Management (Difficulties of translation to other languages)



Top-down Approach

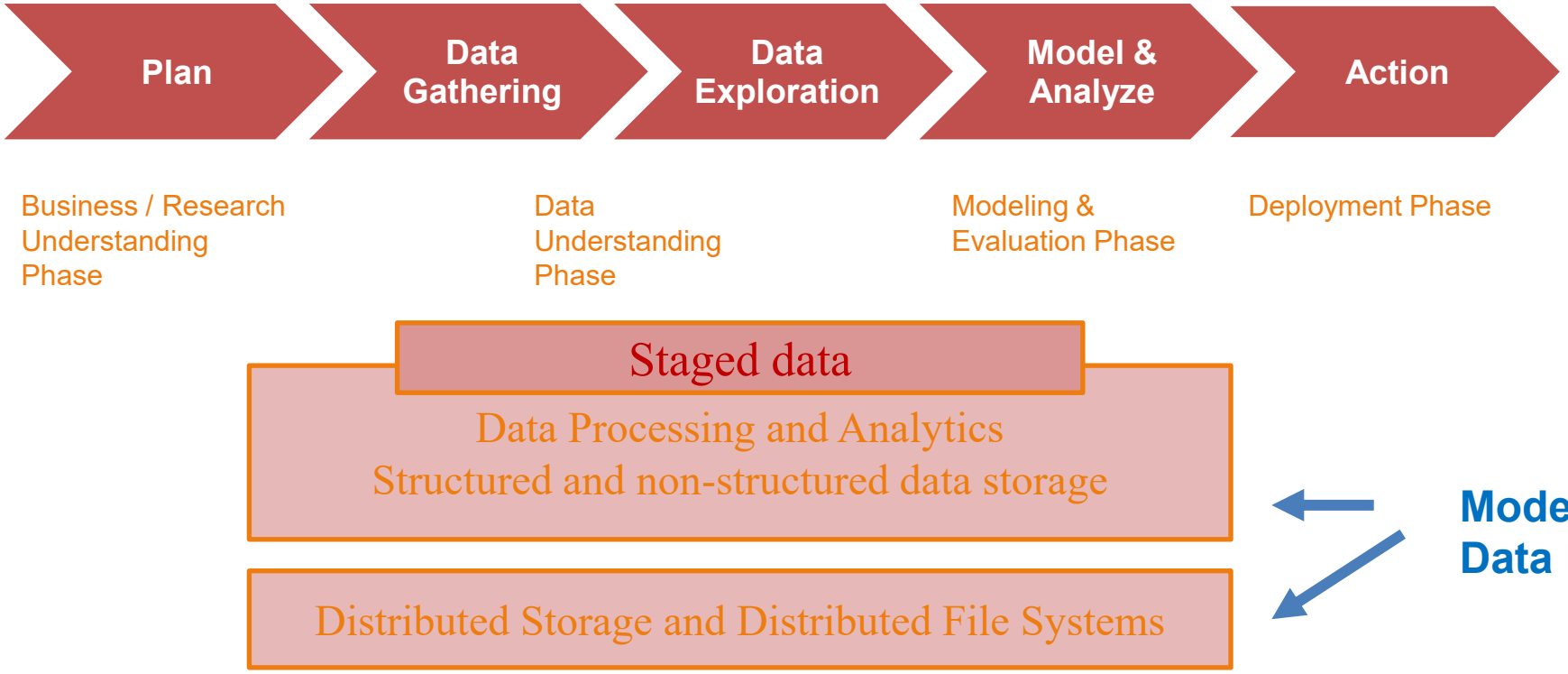
Bottom-up Approach

- Dutch: Gegevensbeheer en gegevensbeheer
- Polish: Zarządzanie danymi i zarządzanie danymi
- Spanish, Greek
- Ukrainian: Урядування даних та управління даними
- German: Datenverwaltung und Datenmanagement
- Finish: Tietojen hallinta ja tiedonhallinta

- Data Governance – Top down
  - *Data Governance is a collection of practices and processes which help to ensure the formal management of data assets within an organization.*
- Data Management – Bottom-up

Business Data Analytics Project Flow and Data Management

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

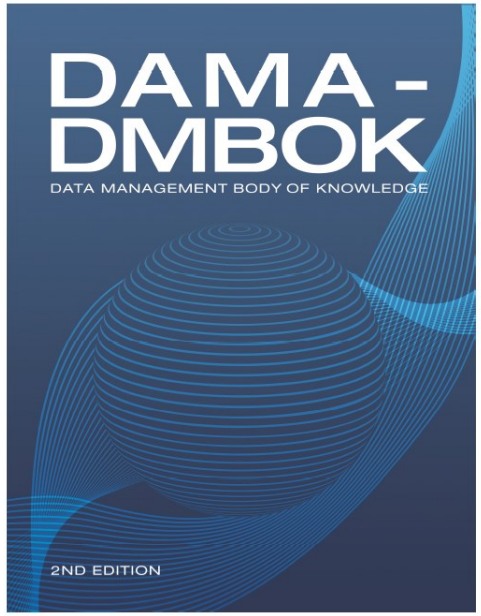


The DAMA-DMBOK Framework

The DAMA-DMBOK Framework goes into depth about the Knowledge Areas that make up the overall scope of data management.

- DAMA-DMBOK Guidelines describe DMBOK and provide re commendations for implementation

- The DAMA Wheel – 11 Knowledge Areas
- The Environmental Factors hexagon
- The Knowledge Area Context Diagram



Technics Publications
BASKING RIDGE, NEW JERSEY

626 pages

[ref] DAMA – DMBOK: Data Management Body of Knowledge, 2nd Edition, 2017. DAMA International, Technics Publications Llc

11 DMBOK Knowledge Areas

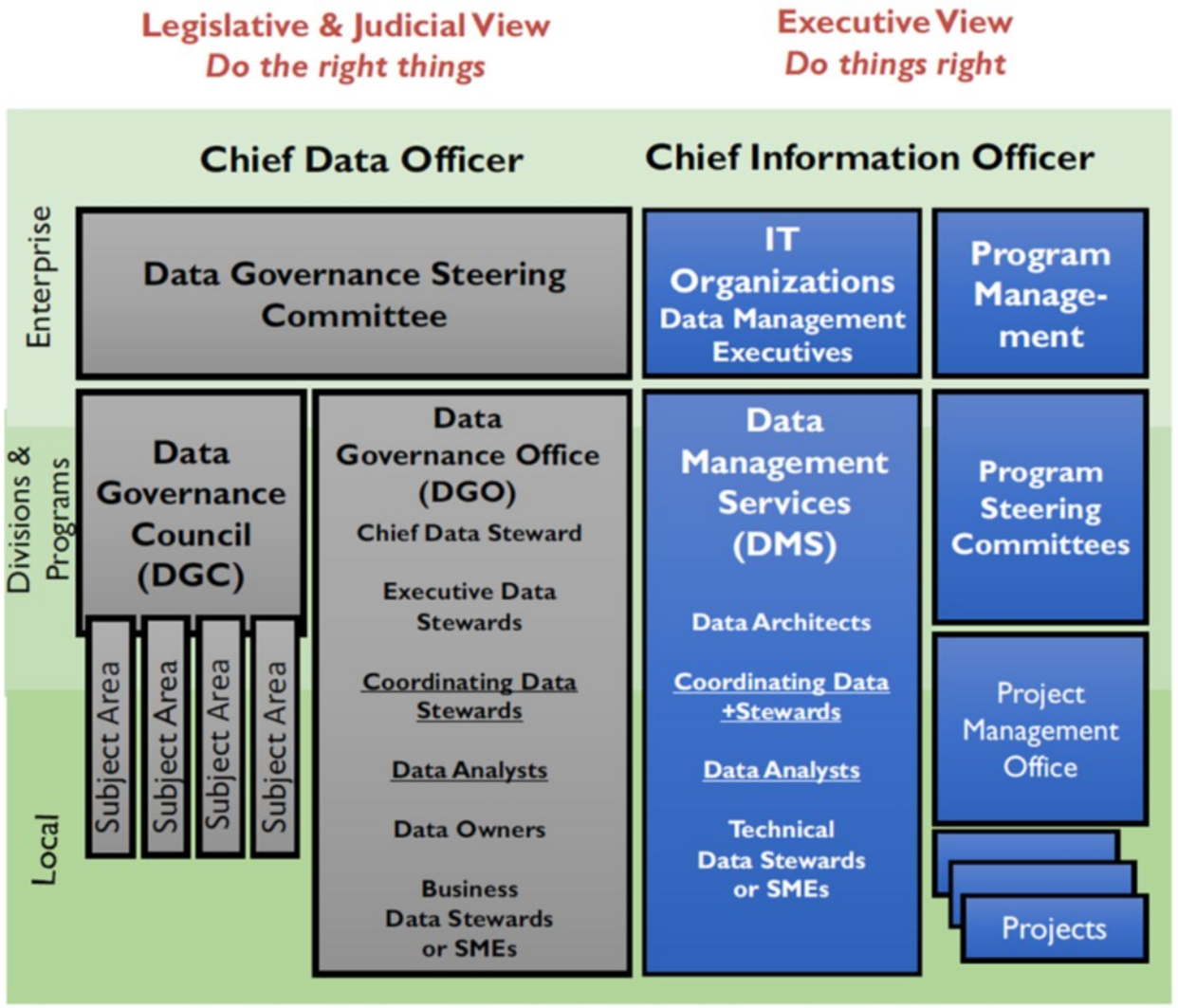
Knowledge Areas describe the scope and context of sets of data management activities.

1. **Data Governance** provides direction and oversight for data management by establishing a system of decision rights over data that accounts for the needs of the enterprise.
2. **Data Architecture** defines the blueprint for managing data assets by aligning with organizational strategy to establish strategic data requirements and designs to meet these requirements.
3. **Data Modeling and Design** is the process of discovering, analyzing, representing, and communicating data requirements in a precise form called the *data model*.
4. **Data Storage and Operations** includes the design, implementation, and support of stored data to maximize its value. Operations provide support throughout the data lifecycle from planning for to disposal of data.
5. **Data Security** ensures that data privacy and confidentiality are maintained, that data is not breached, and that data is accessed appropriately.
6. **Data Integration and Interoperability**
7. **Document and Content Management**
8. **Reference and Master Data**
9. **Data Warehousing and Business Intelligence**
10. **Metadata**
11. **Data Quality**

Data Management Principles

- Data is an asset with unique properties
- The value of data can and should be expressed in economic terms
- Managing data means managing the quality of data
- It takes Metadata to manage data
- It takes planning to manage data
- Data management requirements must drive Information Technology decisions
- Data management is cross-functional; it requires a range of skills and expertise
- Data management requires an enterprise perspective
- Data management must account for a range of perspectives
- Data management is lifecycle management
- Different types of data have different lifecycle characteristics
- Managing data includes managing the risks associated with data
- Effective data management requires leadership commitment

Data Governance Organisation Parts



- Separation or governance responsibilities
- Multi-layer
- CDO
- CIO
- Councils

Data Governance Office (DGO)

- Chief Data Steward
- Executive Data Steward
- Business Data Steward or SME

Research Data Management - Part 1

- Open Access, Open Data, Open Science
- Data Commons, Data Spaces
- 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/Horizon Europe Open Research Data (ORD) Pilot

- ORD pilot aims to improve and maximise access to and re-use of research data generated by EU funded 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 funded projects
  - 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***
  - *Fair use for educational purposes*
- 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.

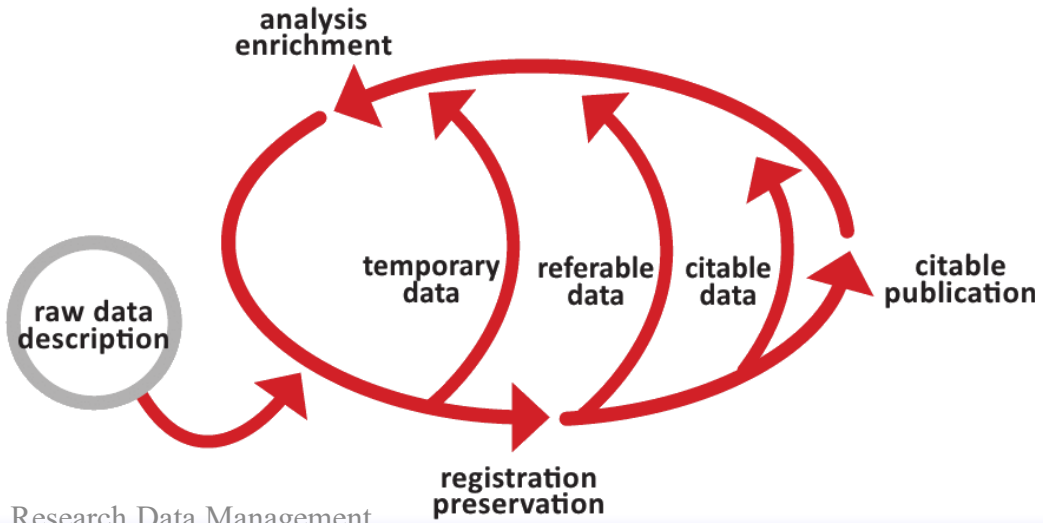
Self-archiving services

- 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.
  - Zenodo creates a unique DOI for all archived documents and 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 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Uses Arxiv DOI format.
- Figshare - https://figshare.com/
  - Data repository, datasets citation

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/)

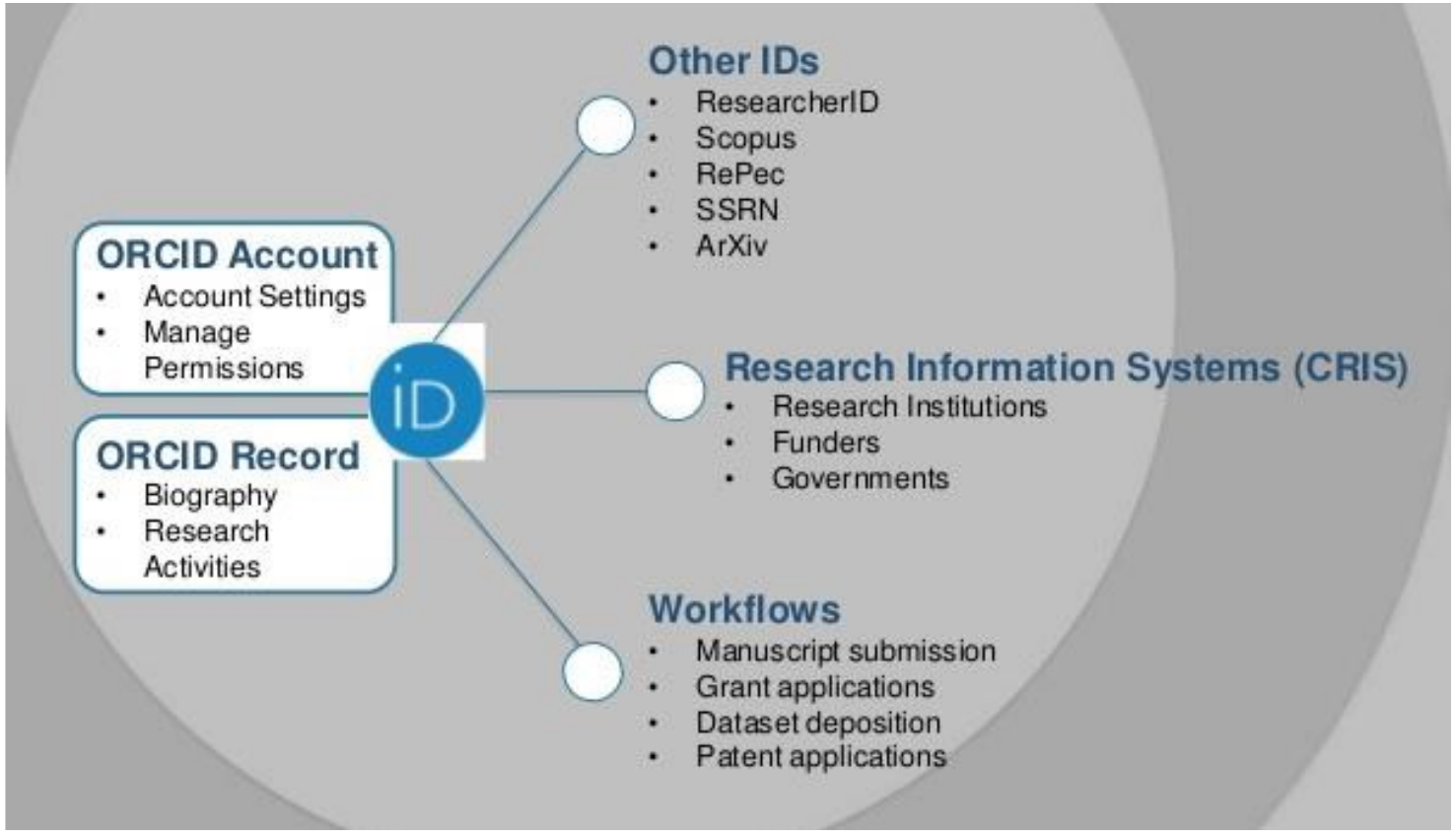
- PID provides a mechanism to link data during the whole research data transformation cycle
  - EPIC RESTful Web Service API published May 2013



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/

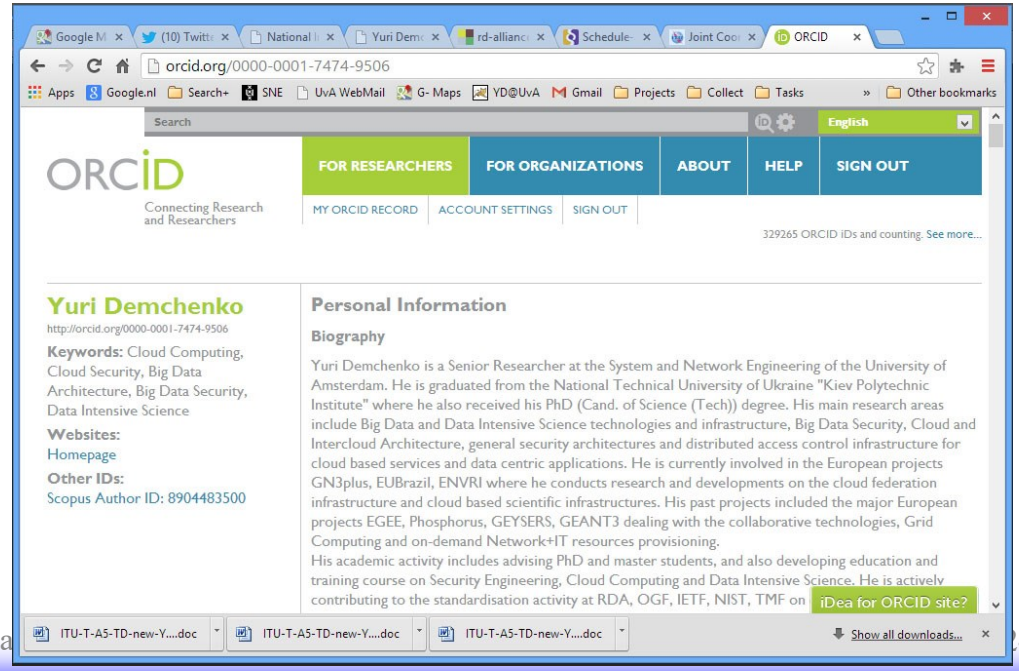
Connecting research and researchers



- 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



Research Data Management

- Data Governance Policy and Data Management
- Documenting data
-

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

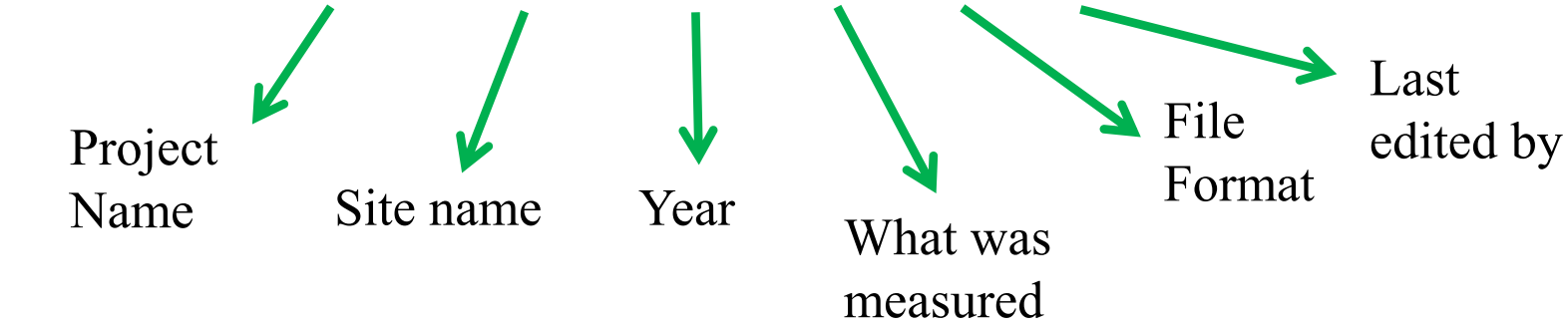
Assign descriptive file names

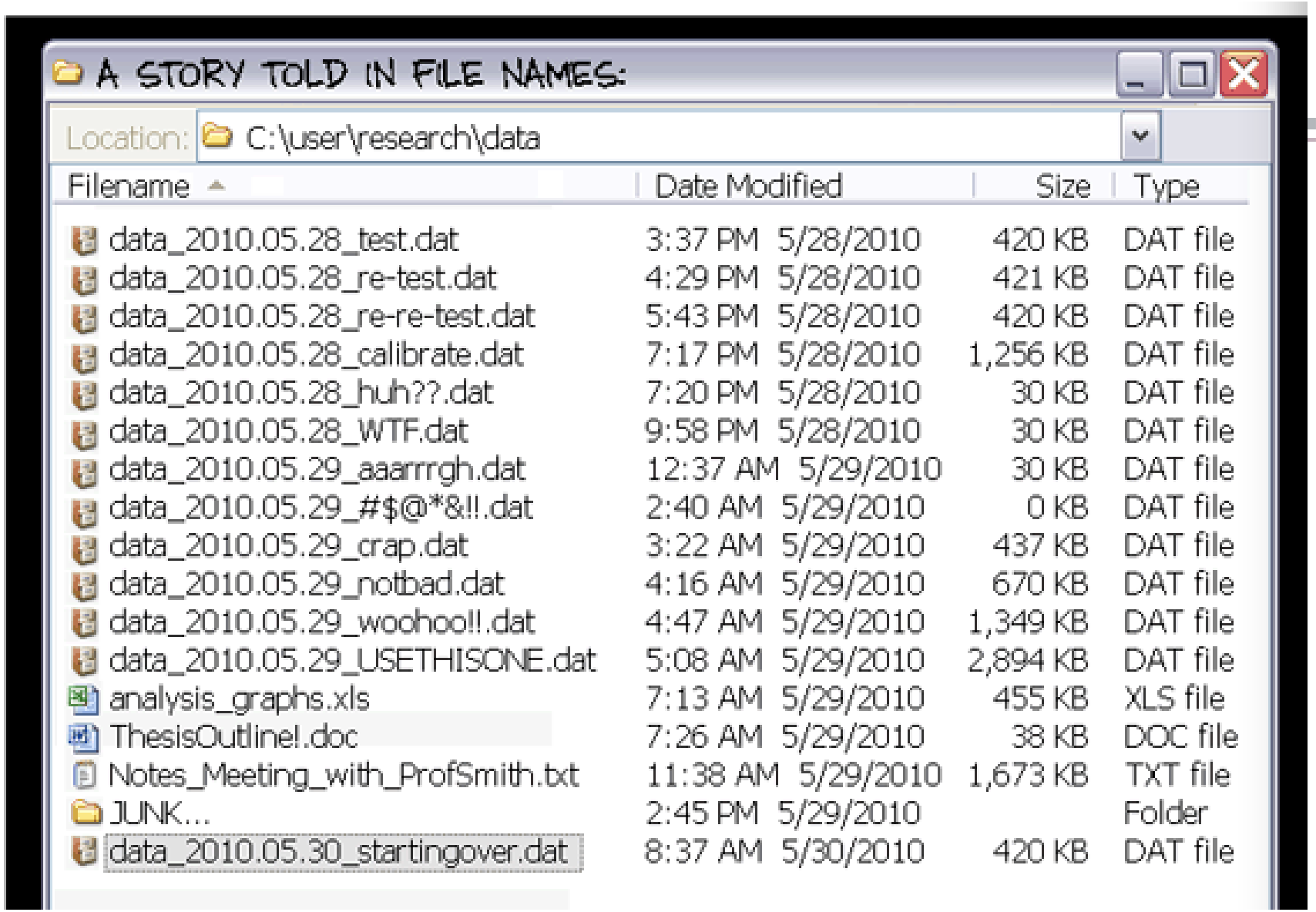
- Use descriptive file names
  - Unique
  - Reflect contents
  - ASCII characters only
  - Avoid spaces
- Provide an explanation of the convention used to name files

Bad: Mydata.xls

2001_data.csv
best version.txt

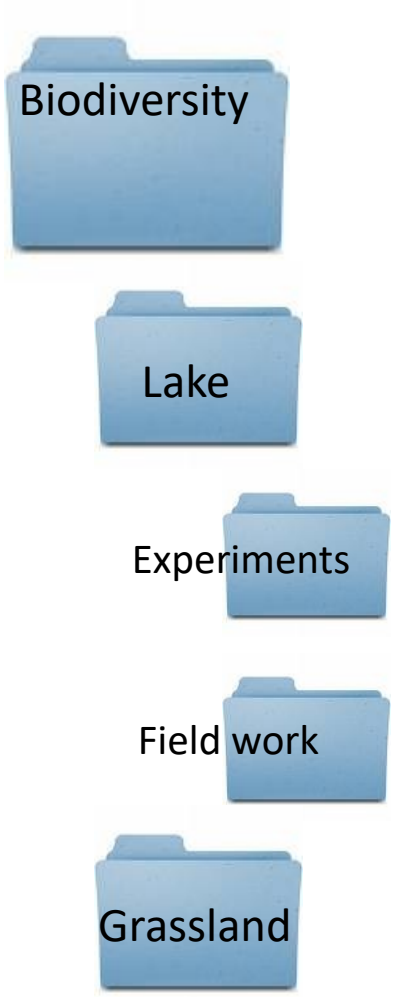
Better: bigfoot_agro_2000_gpp-v03-yd.tiff





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

Organize files logically



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

- 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 – document owner/editor: doesn't append initials
  - 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 Office365 Word – bad for tracking versioning
  - However you can track and restore individual edits

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

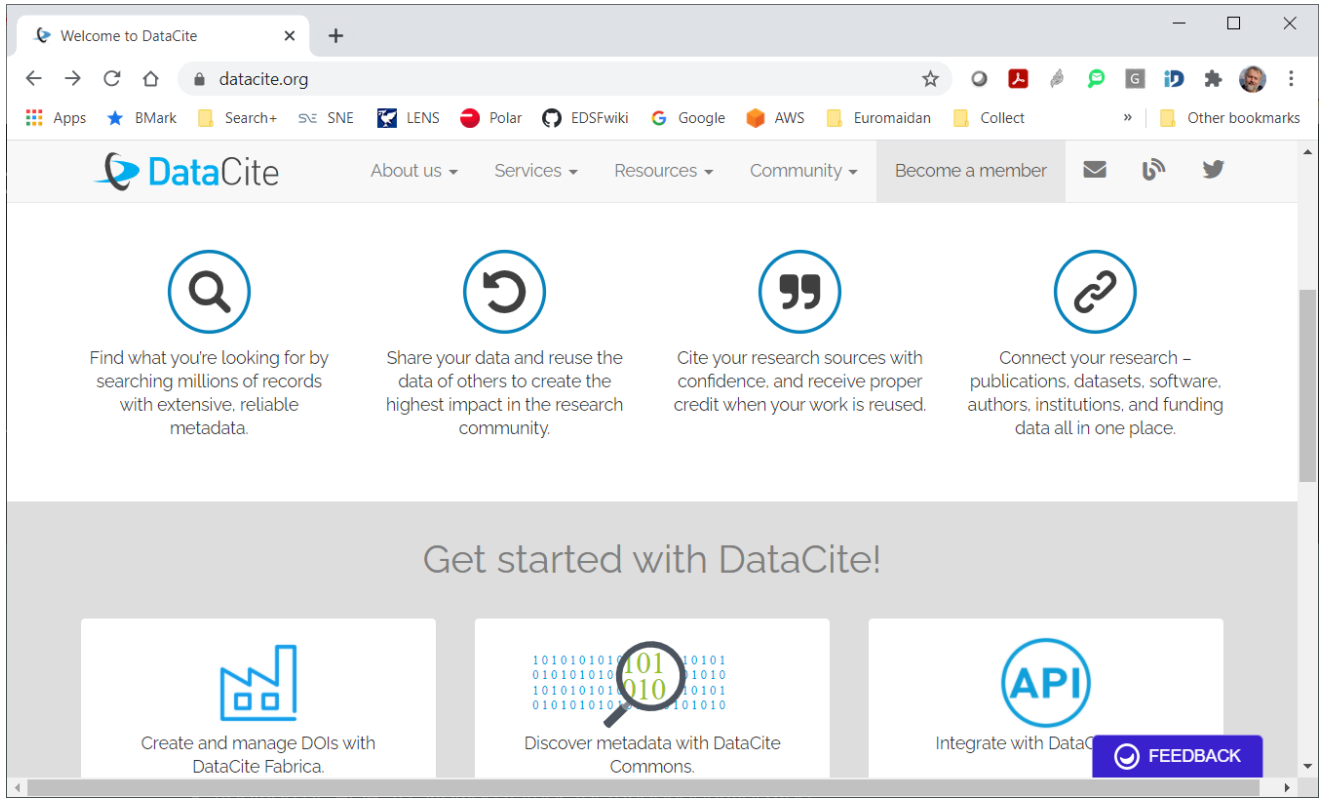
- Metadata definition
- Dublin Core
- Discovery Level Metadata
- Creating a Citation for Your Data
- Sharing your data

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** https://www.dublincore.org/
  - Data Documentation Initiative (DDI) https://ddialliance.org/
  - Metadata standards used by digital libraries
    - Metadata Encoding and Transmission Standard (METS) https://www.loc.gov/standards/mets/
    - Preservation Metadata Maintenance Activity (PREMIS) https://www.loc.gov/standards/premis/

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

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



- Metadata schema
  - https://schema.datacite.org/meta/kernel-4.4/metadata.xsd
  - https://schema.datacite.org/meta/kernel-4.4/

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, °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

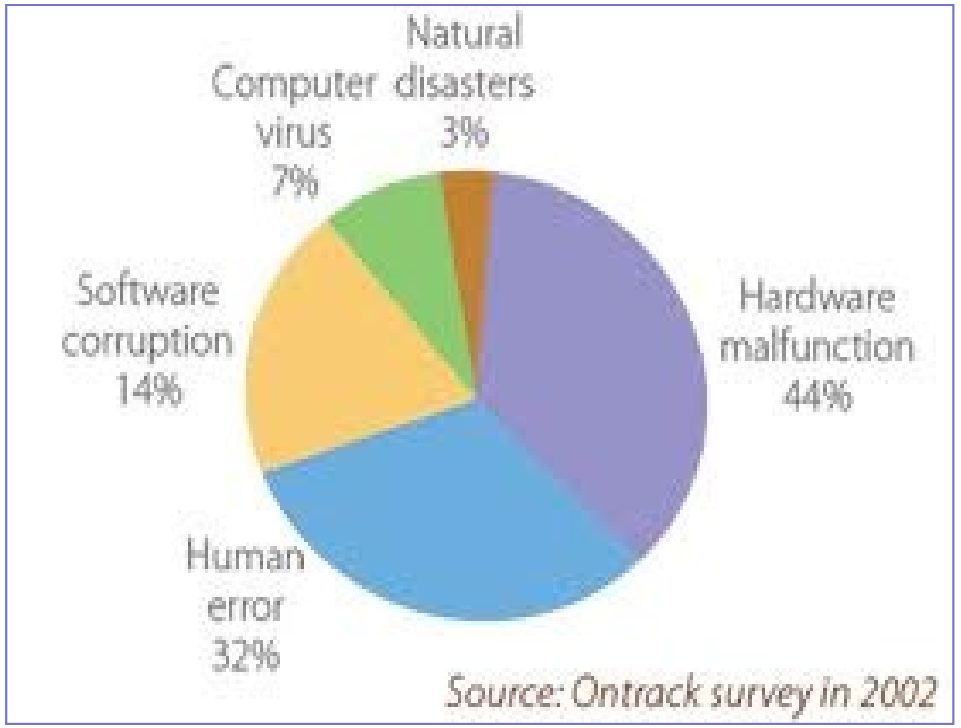
- Discovery level metadata makes it easier to find relevant data in portals, metadata registries, and data inventory systems.
- Being able to find and distinguish data from other similar data sets makes maintaining a data inventory easier because data managers have a better understanding of the content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



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

Compare (2022): Causes of Data Loss

[ref] 7 Causes of Data Loss and How to Combat Them, October 25, 2022 https://invenioit.c
causes-data-loss/

1. Human error
   - Roughly 75% of data loss is caused by human error
2. Natural disasters
   - 40 to 60% of small businesses never reopen their doors after a disaster
3. Hardware failure
   - Hard drives (lifespan 5yr) and SSD (unrecoverable failures)
4. Ransomware, viruses and other malware
5. Software failure
6. Migration errors
7. Malicious deletion

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
- Many organisations are considering 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’ services increasingly popular

- GoogleDrive, DropBox, Microsoft OneDrive, SURFdrive.nl 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
- 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 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

Discussion

- Discussion questions and comments

Part 3 Practice: Data Management Plan (DMP)

- Use template for DMP construction
- Consider GDPR issues
- Consider FAIR data principles

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