Data Archiving and Networked Services
DANS

FAIR Data & Metadata

An introduction to good practices in Research Data Management

Cees H.J. Hof
Data Archiving and Networked Services (DANS-KNAW)

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

Contribution to: Data Acquisition & Analysis
1st year MSc AI and Engineering Systems
Eindhoven University of Technology

18 October 2022, Time slot: 09:45 - 10:30

A few words on DANS and me

Data Archiving and Networked Services
DANS

Dutch national centre of expertise and repository for research data.

- 15+ years of data experience
- ca. 60 staff members
- 3 technical data services
- Archiving over 200.000 scientific datasets

https://dans.knaw.nl/en/

Cees Hof

PhD in evolutionary biology
(Mantis Shrimps)

In past:15+ years working with biodiversity data

GBIF

- Trainer Research Data Management
- DANS Data Station for Life, Health & Medical Sciences
- DANS Data Station for Physical & Technical sciences

Learning objectives:

- Understand what the **FAIR data** concept means
- Understand what **metadata** are and recognise what important categories of metadata we use for scientific data (or software)
- Understand and recognise what **Knowledge Organisation Systems** (KOS) are and what they mean for your (meta)data
- Recognise how all these elements come together in (good) **Research Data Management**

Scientific research data….

Research data……

Research data constitute primary research data (the raw, rough measurements or observation) and secondary research data (the results after the data have been processed by a researcher (recoded, combined, categorised, visualised, etc.)).

Source: University policy framework for research data, Utrecht University (2016)

But…. this is one of the many definitions/ descriptions of research data that go around!

Scientific research data….

Research data……

**May be facts, observations, interviews, recordings, measurements, experiments, simulations and software;
Numerical (quantitative), descriptive (qualitative) and visual;
raw, cleaned up and processed;
they may or may not support an actual or intended publication;
and may be stored and exchanged in various formats on various storage media.**

Source: Berchum, M. van, & Grootveld, M.J. (2016).
In: Handboek Informatiewetenschap, IV B 475, Vakmedianet

Scientific research data....

Five Ways To Think About Research Data:

1) **Research data collection (where do they come from?)**
2) **Types of research data (how do they look like?)**
3) **Electronic storage/ formats (where and how stored?)**
4) **Size and complexity of datasets**
5) **(Research) Data Life Cycle**

Source: **Introducing Research Data** © University of Southampton 2016, Fourth edition
https://eprints.soton.ac.uk/403440/1/introducing_research_data.pdf

1) Research data collection

Reference data
Dataset for comparison or information lookup, for example a complete human genome.

Scientific experiments
Data generated by, e.g. instruments during a scientific experiment.

Models or simulations
Data generated on computer by an algorithm, mathematical model, or simulation.

Derived data
A data set created by taking existing data and performing some manipulation to it.

Observations:
Data generated by recording observations of a specific, **possibly unrepeatable**, event at a specific time or location.

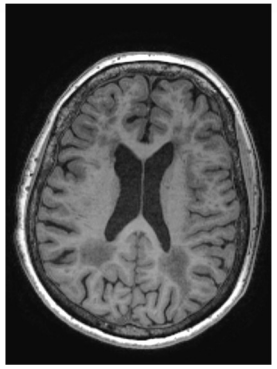
2) Types of research data & 3) Storage

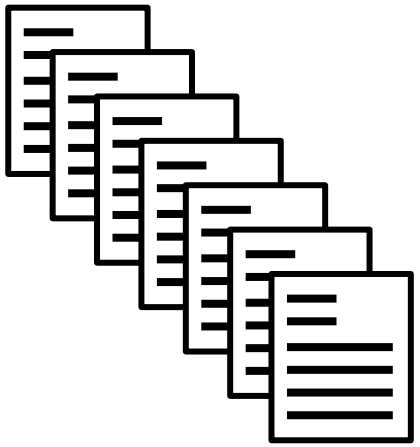
Electronic text documents	• TXT • DOC • PDF
Spreadsheets, numerical data	• Excel (.xls, .xlsx) • CSV
Audiotapes and videotapes, Photographs and films	• Image (JPEG, TIFF, DICOM, …) • Movie (MPEG, AVI, …) • Audio (MP3, WAV, OGG, …)
Specimens, samples, artefacts and slides	For example Digital Cinema Package (DCP), that includes the packaging of different file formats for cinematographic data
Databases	• Multi-purpose (XML) • Relational (MySQL database)
Models, algorithms and scripts	• Py (Python) • R
Discipline, software or instrument specific data files	Software specific formats Discipline specific formats Instrument specific formats

4) Size and complexity of data sets

For example in MRI imaging…..

4) Size and complexity of data sets





- Raw images (e.g. hundreds of stacked images, up to 20 GB for each image, ISMRMRD-standard in h5-format)
- Worked images (DICOM files)
- Algorithm for worked images (often derived from proprietary software)
- Machine metadata
- Machine settings metadata
- Subject data (including medical reference data)
- Patient data
- Consent information
- Research(er) information
- …..

5) Research data life cycle

- During its lifetime data go through a number of phases
- Different disciplines have different ways of thinking about this life cycle
- Transitions between the phases require validation
- Research Data Management is required for each phase
- Each phase (can) come with its own data formats and specific platforms or repositories

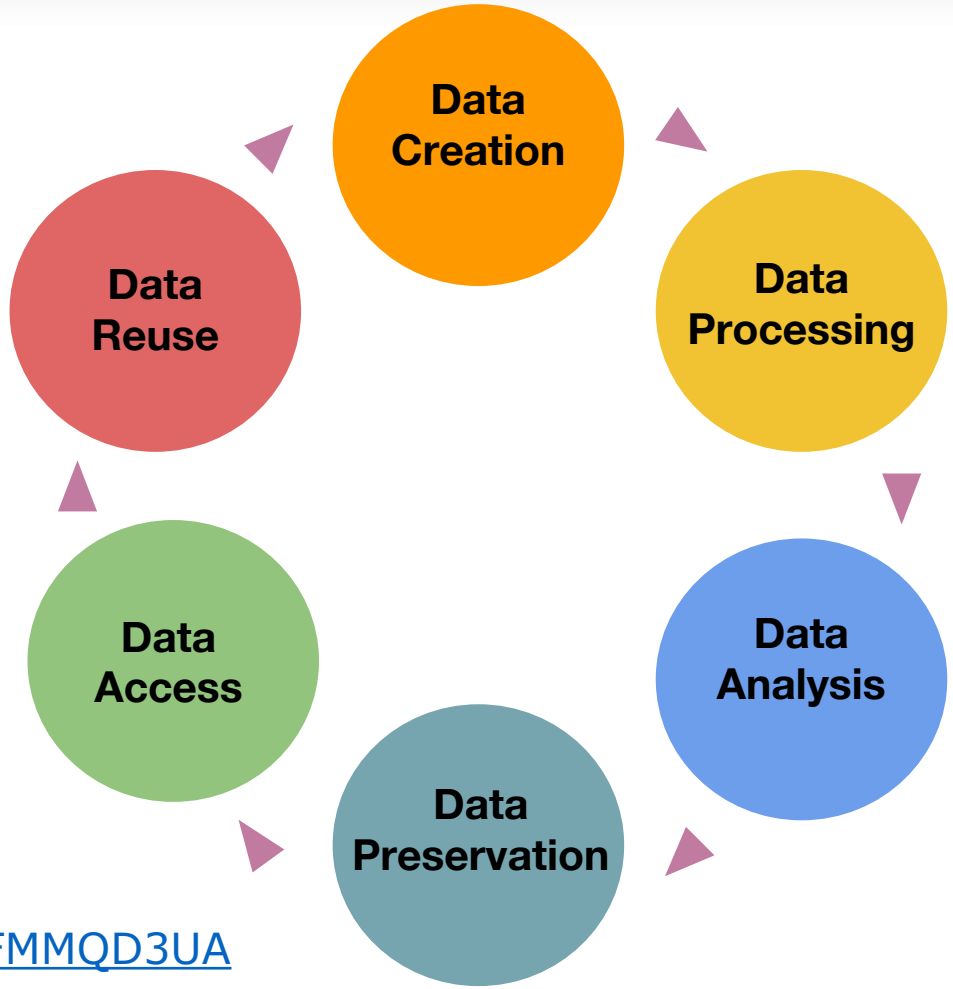
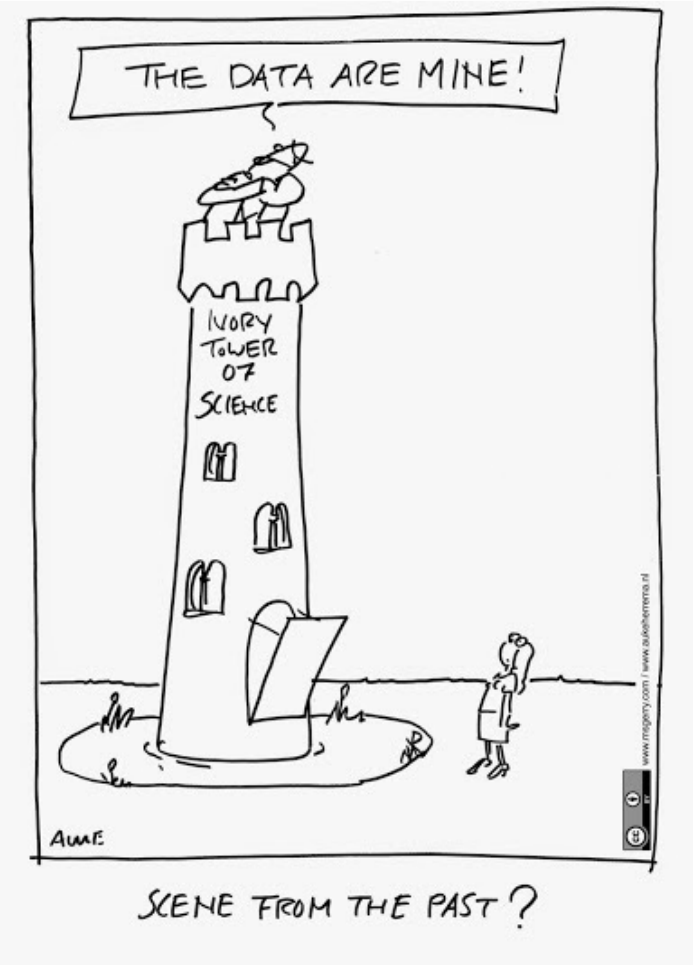


Image derived from UK Data Service clip: https://youtu.be/-wjFMMQD3UA

Why is it good to share science data?



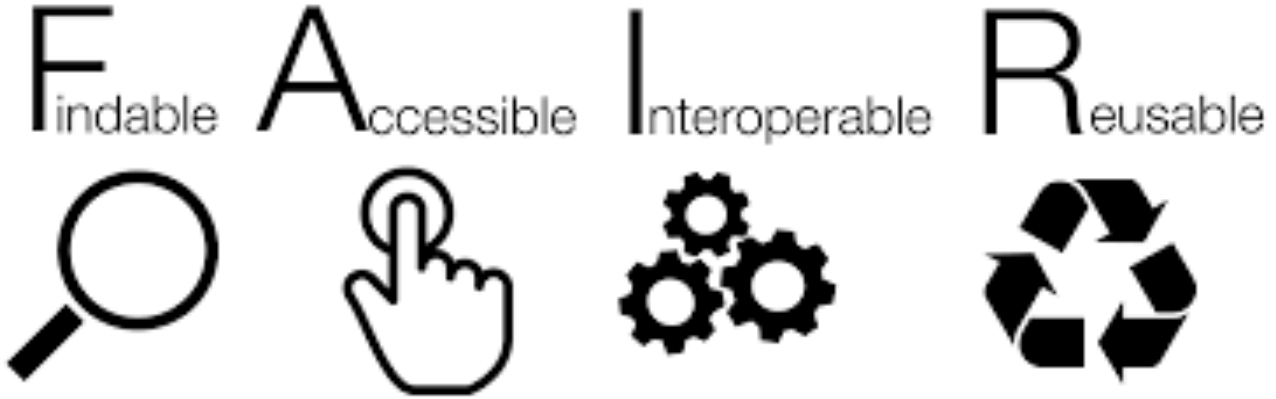
From the OECD 2007 report…..

- Return on public **investments**
- The creation of strong value chains of **innovation**
- Promotes international **co-operation**
- Reinforces open **scientific inquiry**
- Encourages **diversity** of analysis and opinion
- Promotes **new research**
- Facilitates the **education** of new researchers
- Enables the exploration of **topics** not envisioned by the initial investigators
- Permits the creation of **new data sets** when data from multiple sources are combined

OECD Principles and Guidelines for Access to Research Data from Public Funding (2007)
https://www.oecd.org/sti/inno/38500813.pdf

How to make full use of the potential of research data?

Make your data….



The FAIR principals originated from a Life Science Lorentz Workshop in Leiden in 2014

Lorentz center

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).

The FAIR Data Principles

In order to make full use of the potential of research data, it is necessary to include them in the research eco-system as **Findable**, **Accessible**, **Interoperable** and **Reusable** as possible

The FAIR principles consist of 15 facets. The main thing is that research data should not only be FAIR for **people**, but also for **computers/ machines**.

The FAIR principles are now an integral part of the data management landscape and form the basis of the construction plan for the European Open Science Cloud.

Source(s): GO FAIR https://www.go-fair.org/fair-principles/ & EOSC https://eosc.eu

FAIR Data Principles (selection of….)

Findable (metadata and data should be easy to find for both **humans and computers**):

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata

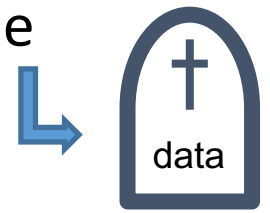
F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible (including authentication and authorisation):

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A2. Metadata are accessible, even when the data are no longer available



Digital tombstones provide metadata for dead data

Source: https://www.go-fair.org/wp-content/uploads/2022/01/FAIRPrinciples_overview.pdf

FAIR Data Principles (selection of….)

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

Reusable:

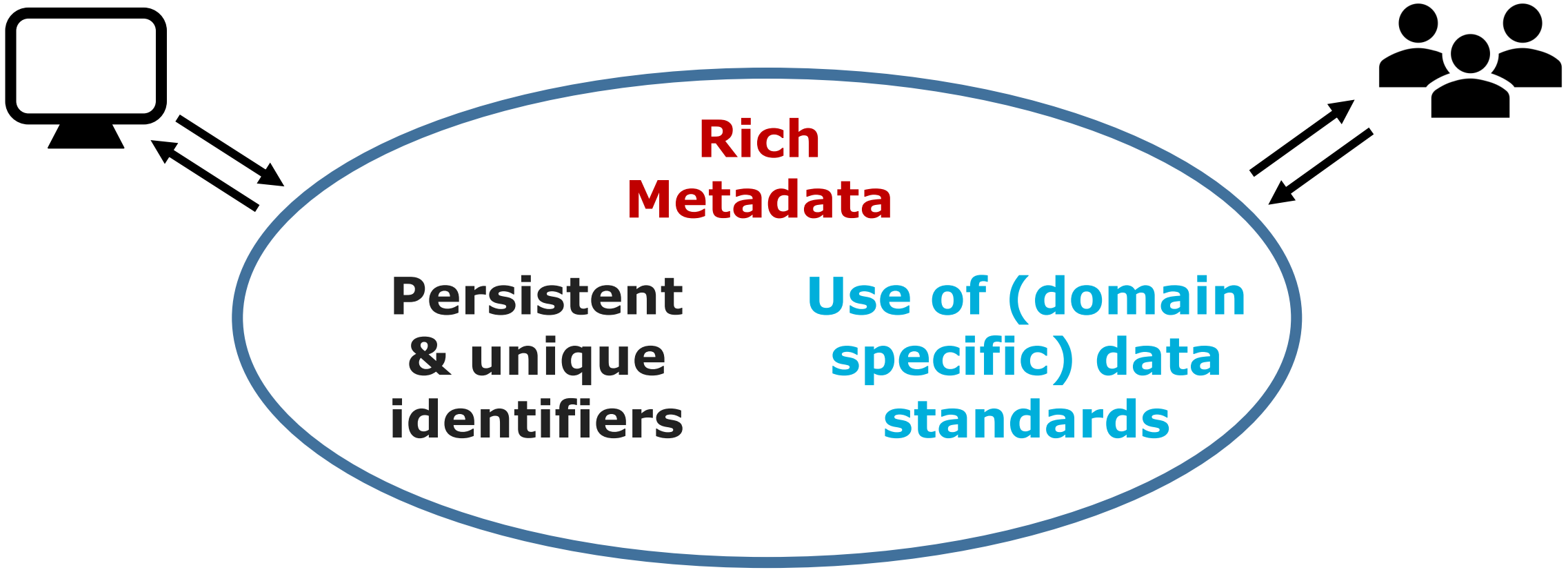
R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

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 Data Principles, pivotal elements



FAIR & Metadata.....

Metadata is "data that provides information about other data"

FAIR & categories of metadata…..

Three main types of metadata are recognised:

Administrative metadata: data about a project or resource that are relevant for managing it; E.g. project/resource **owner**, **principal investigator**, project **collaborators**, **funder**, project period, etc. They are usually assigned to the data, before you collect or create them.

Descriptive or citation metadata: data about a dataset or resource that allow people to discover and identify it; E.g. **authors**, title, abstract, **keywords**, **persistent identifier**, related publications, etc.

Structural metadata: data about how a dataset or resource came about, but also how it is **internally structured**. E.g. the unit of analysis, collection method, sampling procedure, sample size, categories, **variables**, etc. Structural metadata have to be gathered by the researchers according to **best practice in their research community**…

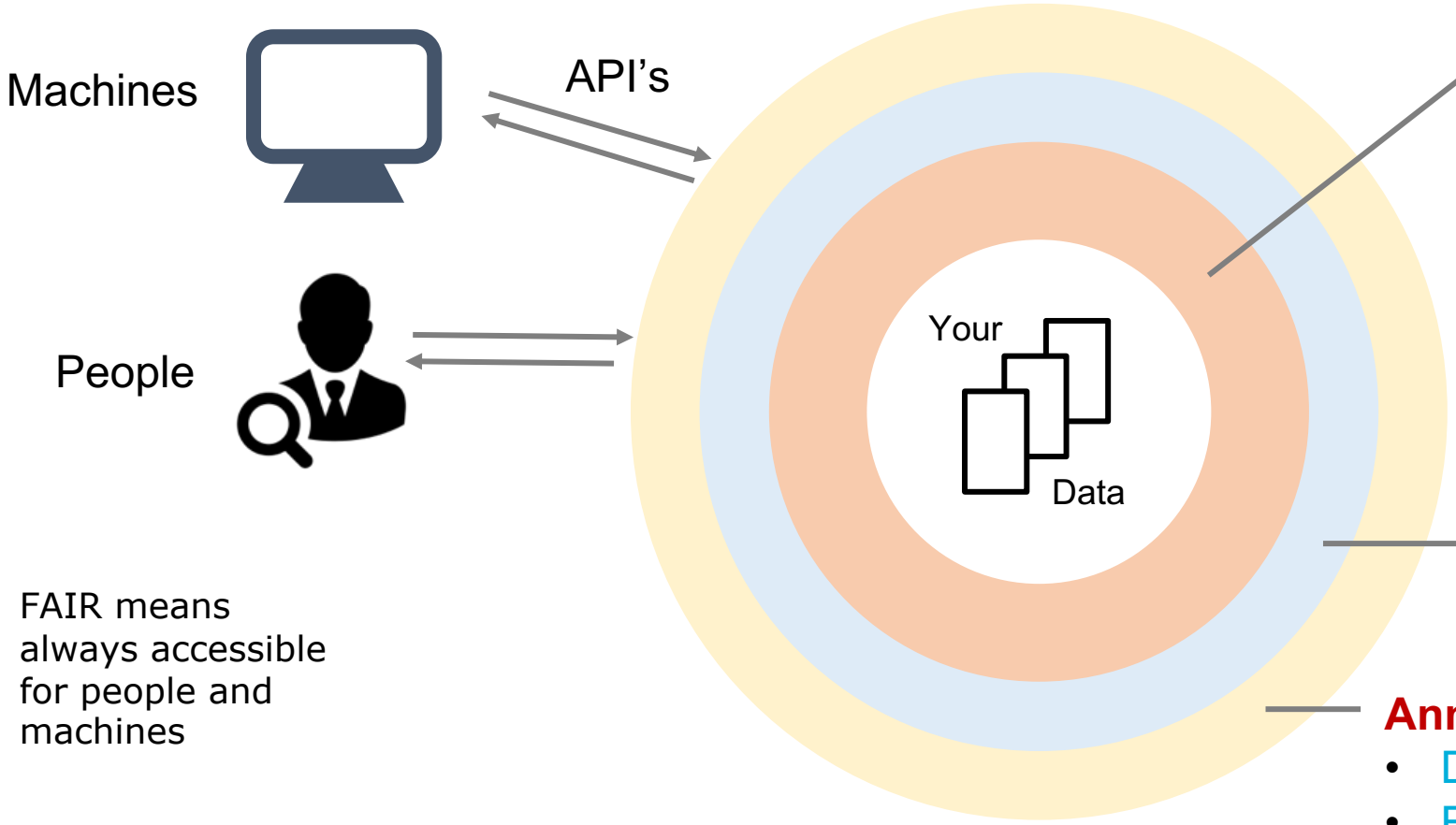
Source: https://merlinone.com/types-of-metadata/

FAIR & categories of metadata.....

Examples of different metadata standards:

- **Dublin Core** - domain agnostic, basic and widely used metadata standard
- **DDI** (Data Documentation Initiative) - common standard for social, behavioral and economic sciences, including survey data
- **EML** (Ecological Metadata Language) - specific for ecology disciplines
- **ISO 19115** and FGDC-CSDGM (Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata) - for describing geospatial information
- **DCAT** standard used by the Dutch government: DCAT-DONL
- **FITS** (Flexible Image Transport System) - Astronomy digital file standard that includes structured, embedded metadata

FAIR & categories of metadata.....



FAIR means always accessible for people and machines

Core metadata (descriptive & administrative)

- Author
- Place
- Time
- PID
- Description
- Access
- Terms of use
- Statistics

Domain (structural)

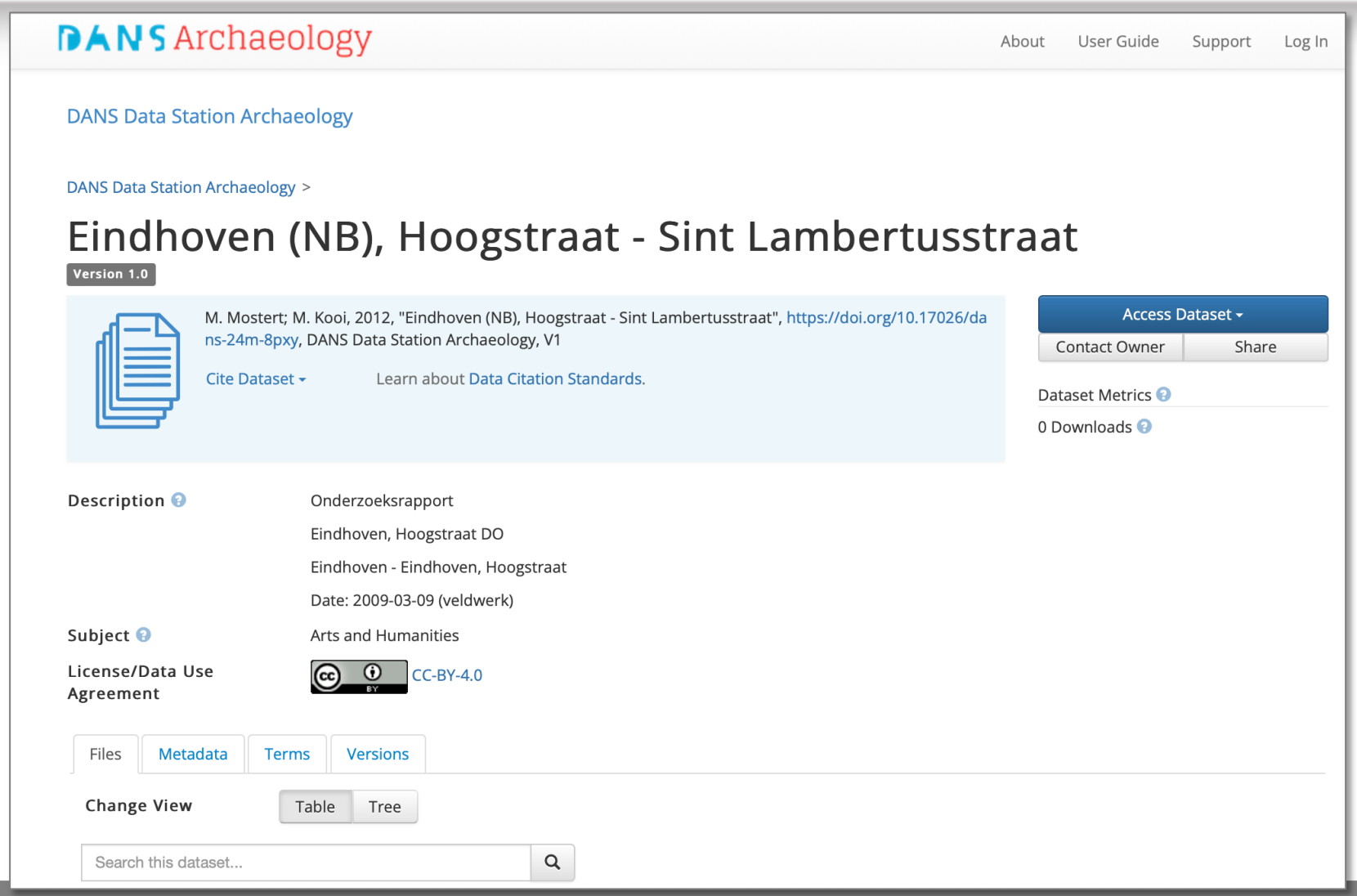
- Standards
- Methods
- protocols

Annotations:

- Data use
- Enhancements
- Errors

Example…. metadata at work

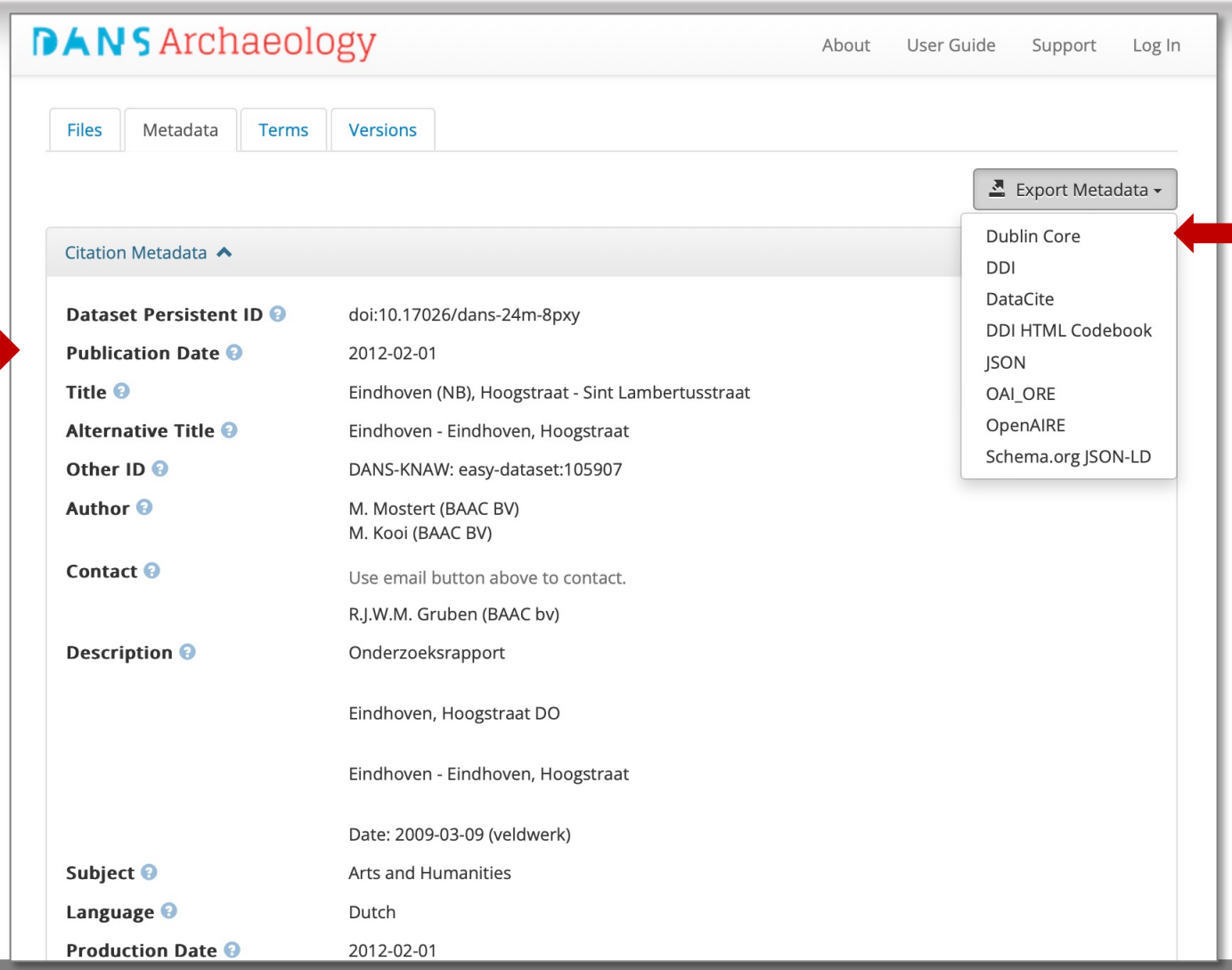
Archaeological data example city of Eindhoven



https://doi.org/10.17026/dans-24m-8pxy

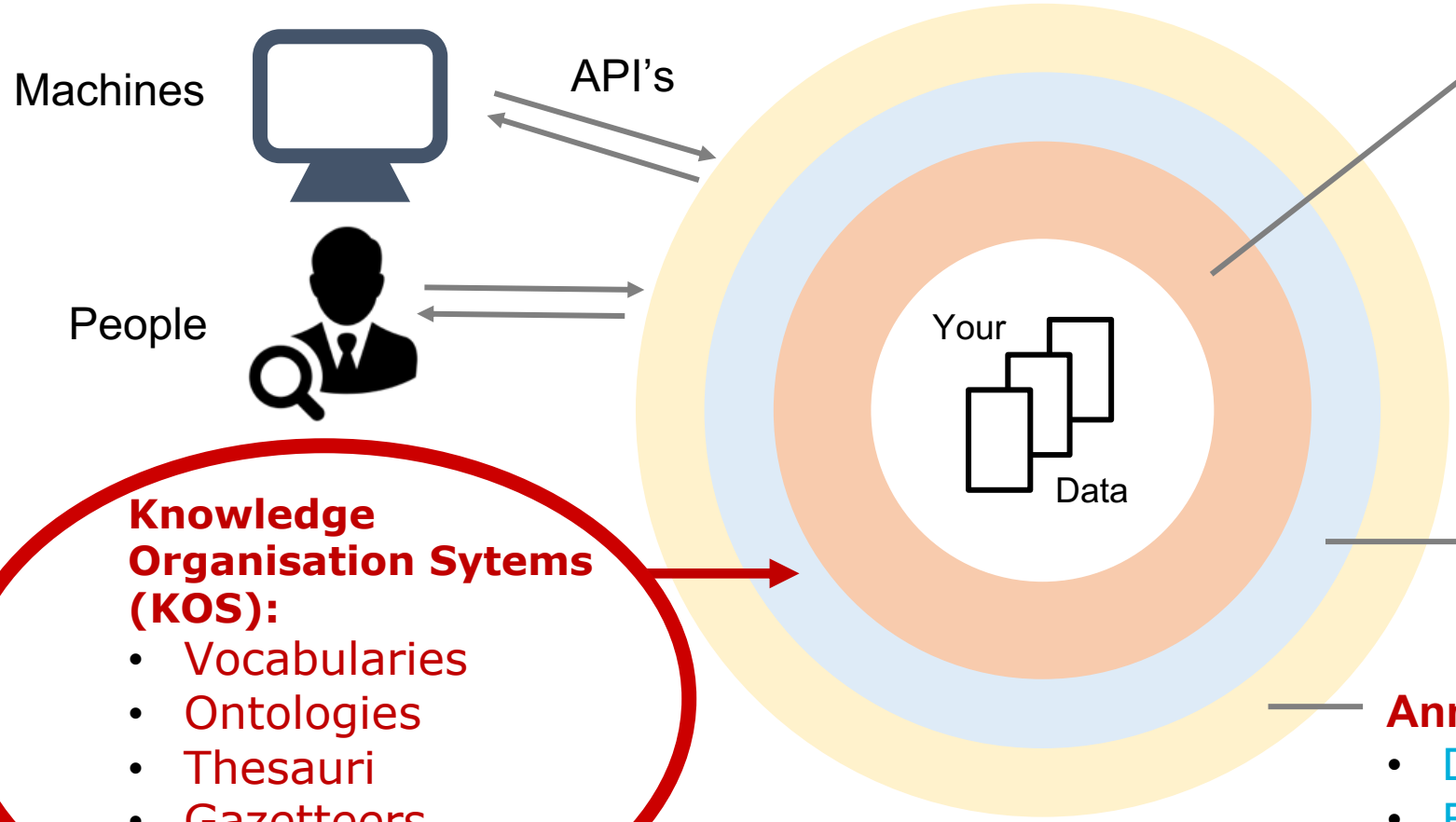
Example…. metadata at work

An example with metadata following the (very generic) Dublin Core metadata standard



Machine-readable metadata export formats (also available through APIs)

FAIR & means to enhance your (meta)data



Core metadata (descriptive & administrative)
- Author
- Place
- Time
- PID
- Description
- Access
- Terms of use
- Statistics

Domain (structural)
- Standards
- Methods
- protocols

Knowledge Organisation Sytems (KOS):
- Vocabularies
- Ontologies
- Thesauri
- Gazetteers
- ….

Annotations:
- Data use
- Enhancements
- Errors

FAIR & Knowledge Organisation Systems

The term **knowledge organization systems (KOS)** is intended to encompass all types of schemes for organizing information and promoting knowledge management.

Council on Library and Information Resources

What they all have in common is that they have been designed to support the organization of knowledge and information in order to make the **management and retrieval of data and information easier**.

International Society for Knowledge Organization

Types:

- Vocabularies: organized words and phrases representing unique concepts, for indexing and cataloguing purposes. Example: the LTER Controlled Vocabulary for ecological data.
- Glossaries: alphabetical lists of terms with definitions. Example: the list of environmental terms used by the European Environment Agency.
- Thesauri: reference work that lists words grouped together according to similarity of meaning, usually with a cross-reference system. E.g. the Getty Thesaurus of Geographic Names ®.
- Ontologies: logic-based organizational structures for knowledge, allowing the creation of a large number of relationships. (Most complex form of KOS.)

FAIR & Knowledge Organisation Systems

Eindhoven in the Getty Thesaurus of Geographic Names

https://www.getty.edu/research/tools/vocabularies/tgn/

Exports in different formats

Specific PID for placename

Standardised coordinates

Geographical hierarchy

Research

Research Home ▸ Tools ▸ Thesaurus of Geographic Names ▸ Full Record Display

Getty Thesaurus of Geographic Names® Online
Full Record Display

New Search | Previous Page | Help

Vernacular Display | **English Display**

Click the icon to view the hierarchy.

Semantic View (JSON, JSONLD, RDF, N3/Turtle, N-Triples)

ID: 7006842 **Record Type: administrative**
Page Link: http://vocab.getty.edu/page/tgn/7006842

Eindhoven (inhabited place)

Coordinates:
Lat: 51 27 00 N *degrees minutes* Lat: 51.4500 *decimal degrees*
Long: 005 28 00 E *degrees minutes* Long: 5.4667 *decimal degrees*

Note: Located on the Dommel river in the Kempen heathland, the small village grew dramatically after 1900 to become one of the largest industrial centers of The Netherlands. Known as the 'town of light,' as it is the home of the Philips light bulb factory.

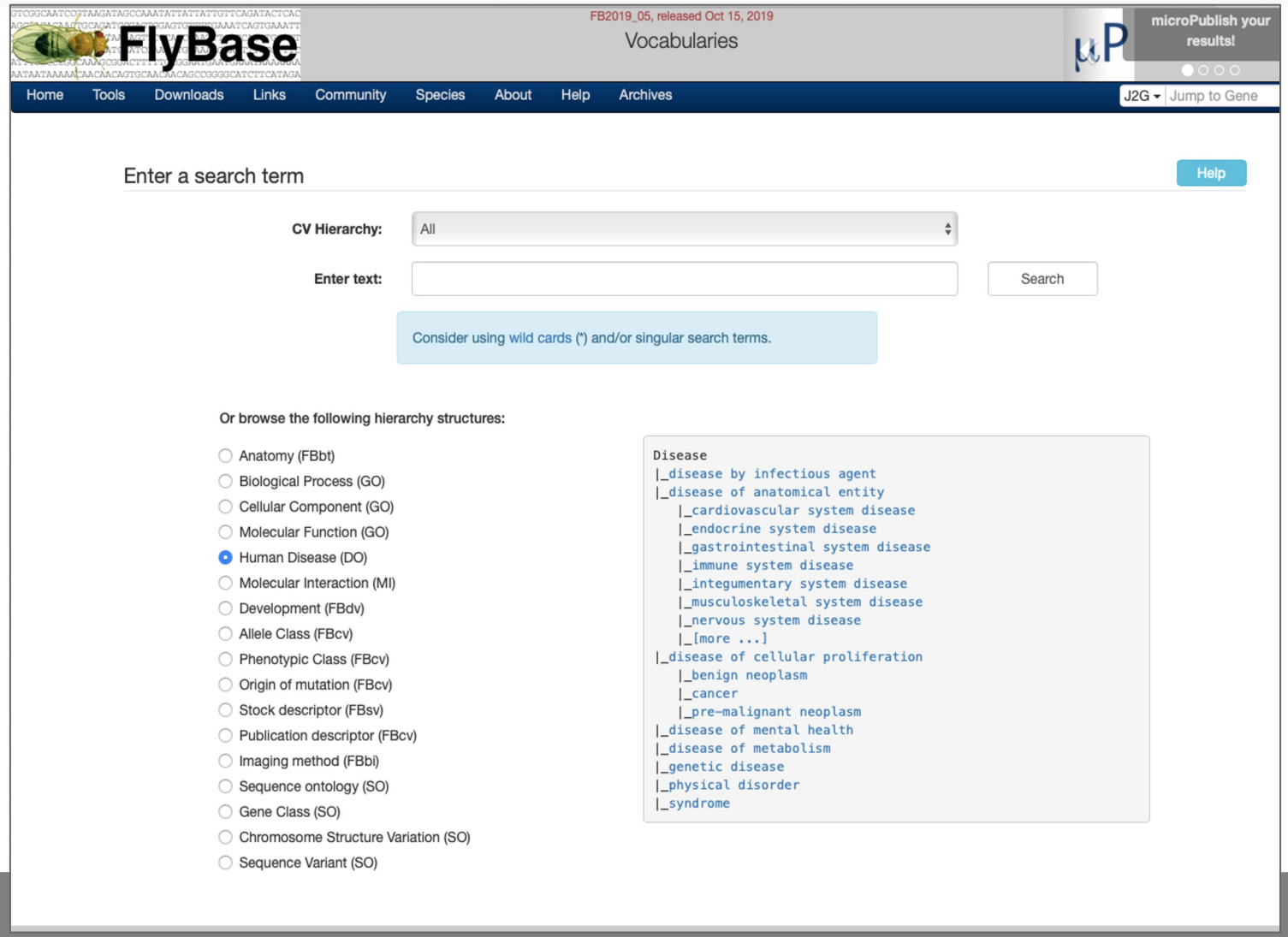
Names:
Eindhoven (**preferred**,C,V,Dutch,U)

Hierarchical Position:
World (facet)
.... Europe (continent) (P)
..... Netherlands (nation) (P)
..... North Brabant (province) (P)
..... Eindhoven (inhabited place) (P)

FAIR & Knowledge Organisation Systems

Within certain research communities there are very advanced (controlled) vocabulary systems. A very good example is FlyBase (genetic research related to fruit flies). Here the controlled vocabalary (CV) consists of a list of terms that all have been used to annotate genetic expression.

https://flybase.org/vocabularies



The Knowledge Organisation Systems (KOS) landscape

KOS terminology is not used consequently and interpretations differs significantly across scientific disciplines and information domains (libraries, data management, computing and software).

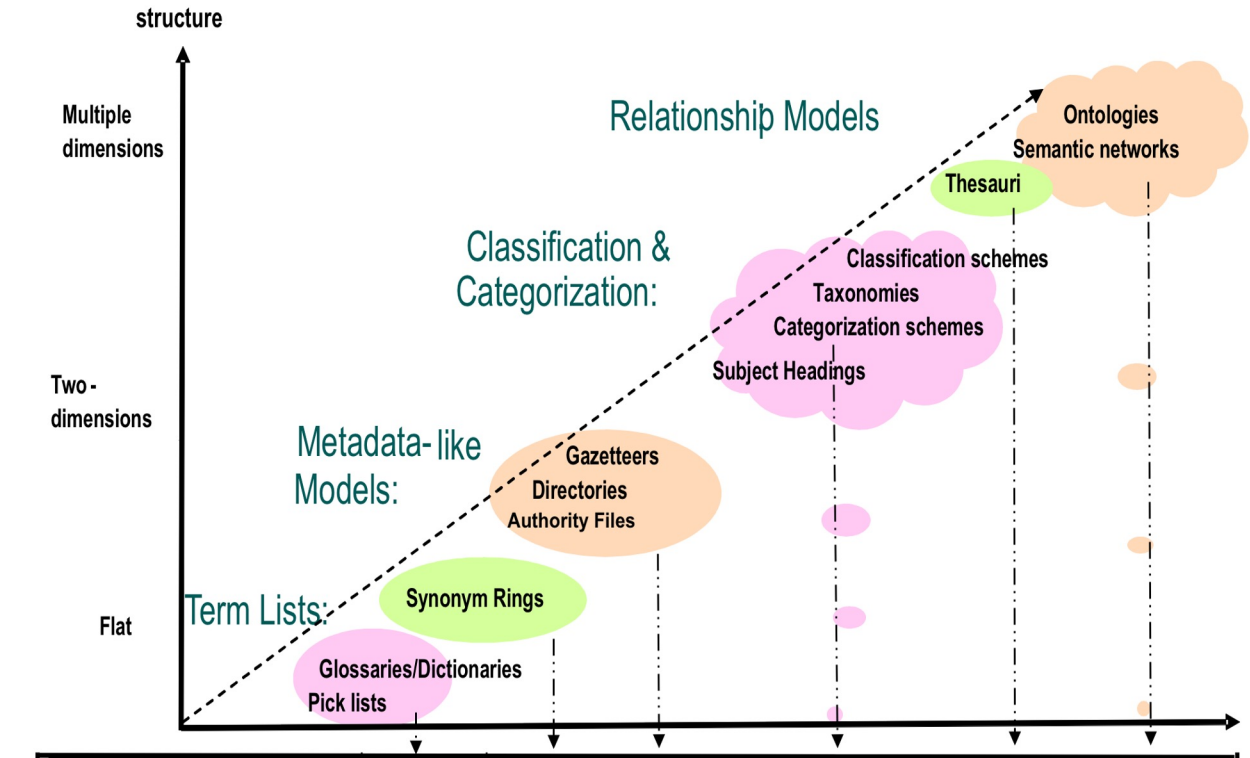
What is clear is that there is a scale of complexity when it comes to types of Knowledge Organisation Systems

Generic info sources:
https://www.clir.org/pubs/reports/pub91/

Zeng, M.L., Knowledge Organization Systems (KOS) Knowl. Org. 35(2008) No.2/No.3

Various Types of KOS

Zeng 2008 p. 161



Major functions						
eliminating ambiguity	xxx		xxx	xx	xxxx	xx
controlling synonyms		xxxx	xxx	xx	xxxx	xx
establishing relationships: hierarchical			x	xxxx	xxx	xxx
establishing relationships: associative					xxxx	xxxxx
presenting properties						xxxxx

Discussion….

In this lecture we basically addressed some of the key aspects of FAIR data at a very generic level. For a true understanding of subjects such as metadata and metadata standards, the use of identifiers, how to use Knowledge Organisation Systems and the role of repositories in making data FAIR, additional teaching and training is required. Basic (online) Research Data Management courses are a good start to familiarise yourself with the skills and knowledge needed to make (your) research data truly FAIR…..

FAIR & how to bring everything together

Research data management (RDM) refers to how you handle, organise, and structure your research data throughout the research process. Data management:

- Begins with your initial considerations regarding what will be necessary for using or collecting your particular type of data;
- Includes measures for maintaining the integrity of the data, making sure that they are not lost due to technical mishaps, and that the right people can access the data at the appropriate time;
- Looks forward to the future, making it clear that you should provide detailed and structured documentation to be able to share your data with other colleagues and prepare them for long-term availability.

From: CESSDA Data Management Expert Guide
https://dmeg.cessda.eu/Data-Management-Expert-Guide

Want to learn more?

Start the free online course “Essentials 4 Data Support” in Dutch or English!

https://datasupport.researchdata.nl/en/