



Module 1 – Introduction to data

Nene Djenaba Barry

Digital Learning Hub

27 June 2024



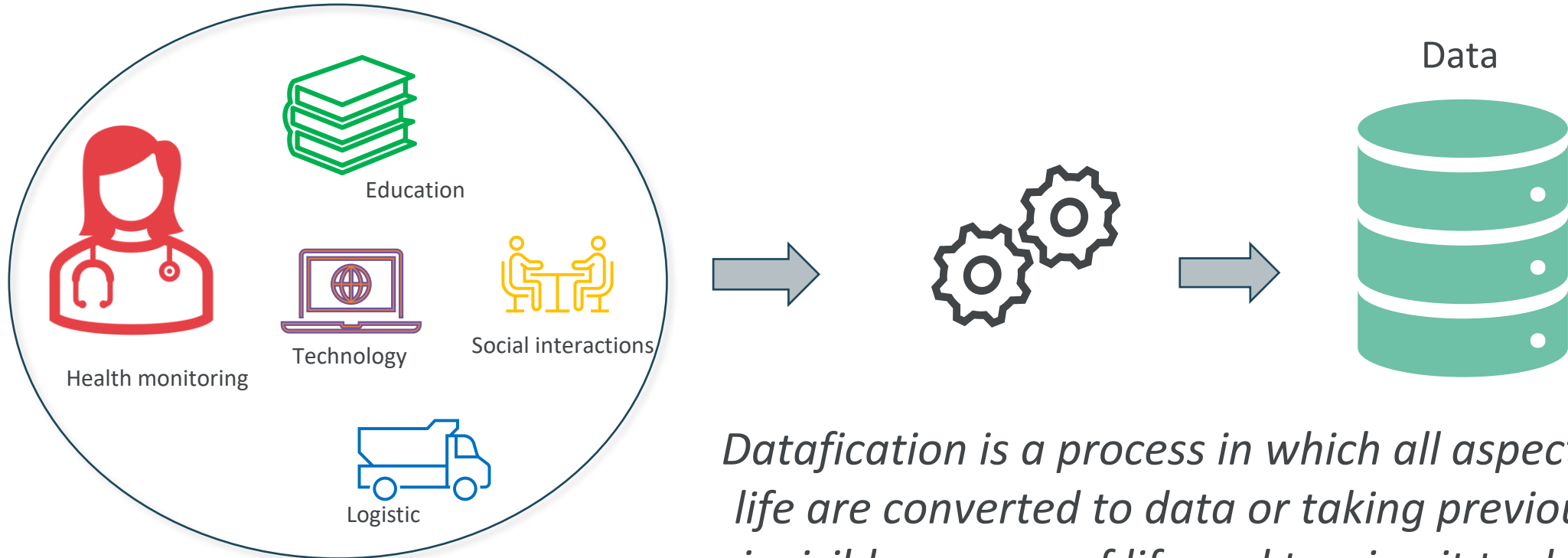
Learning objectives

- Understand the importance of data in our society
- Define data and distinguish between data, information, and knowledge
- Define metadata, list the types of metadata and understand its role in documenting data
- Identify data structure and the associated formats
- Define data store and differentiate between various types of data stores
- Define data classification and understand its importance in data management
- Understand the concept and purpose of data management and stewardship
- Define and identify the FAIR principles for data management



Introduction

Data is a key to understand and navigate our complex world!



Datafication is a process in which all aspects of life are converted to data or taking previously invisible process of life and turning it to data.



> 90% of the data in the world has been created in the last two years ¹

1

3

¹ <https://explodingtopics.com/blog/data-generated-per-day>

Introduction

The world's most valuable resource is no longer oil, but data¹

The data economy demands a new approach to antitrust rules



The role of data in health diplomacy: A case study on global vaccination governance²

Pichelstorfer, Anna ; Paul, Katharina T. 

Editors: Mays, C; Laborie, L; Griset, P

Journalism Education for Datafied Society: Fostering Data (infrastructural) literacy³

Milojevic Ana¹ 

Show affiliations

Researcher: Milojevic Ana¹ 

Show affiliations

¹ <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>

⁴ ² <https://zenodo.org/records/7410061>

³ <https://zenodo.org/records/6582712>



Data

Data definition

Facts or information, especially when examined and used to find out things or to make decisions -
"Oxford Learner's Dictionary"



Information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer -
"Cambridge Dictionary"



Data, Information, Knowledge are closely related concepts

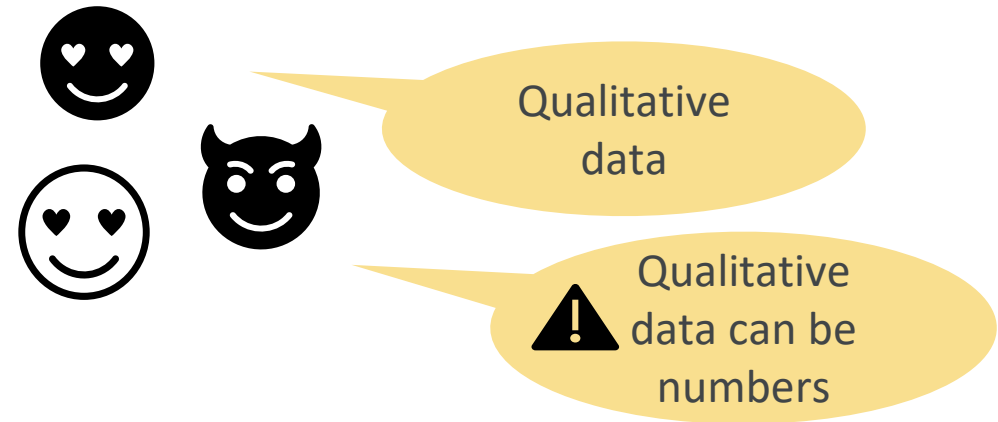


Data, Information, Knowledge

Definitions



20, 35, and 45 are the numbers of students registered for the next data stewardship courses



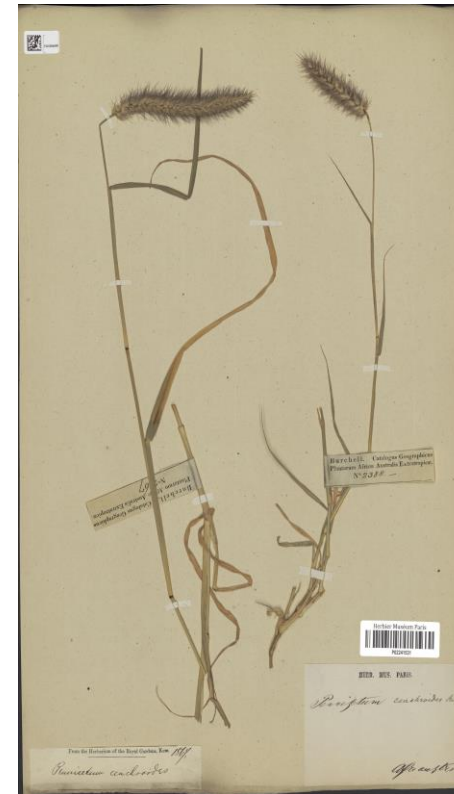
From this information, the average enrollment for data stewardship courses is 33.3%



Data documentation - Metadata

Data about data

- Structured information that describes, explains, locates, or otherwise makes retrieving, using, or managing an information resource easier.” NISO 2004
- Three common types of metadata
 - Descriptive: title, authors, subjects, keywords, and publisher
 - Structural: data dictionary, schema
 - Administrative: technical and rights metadata



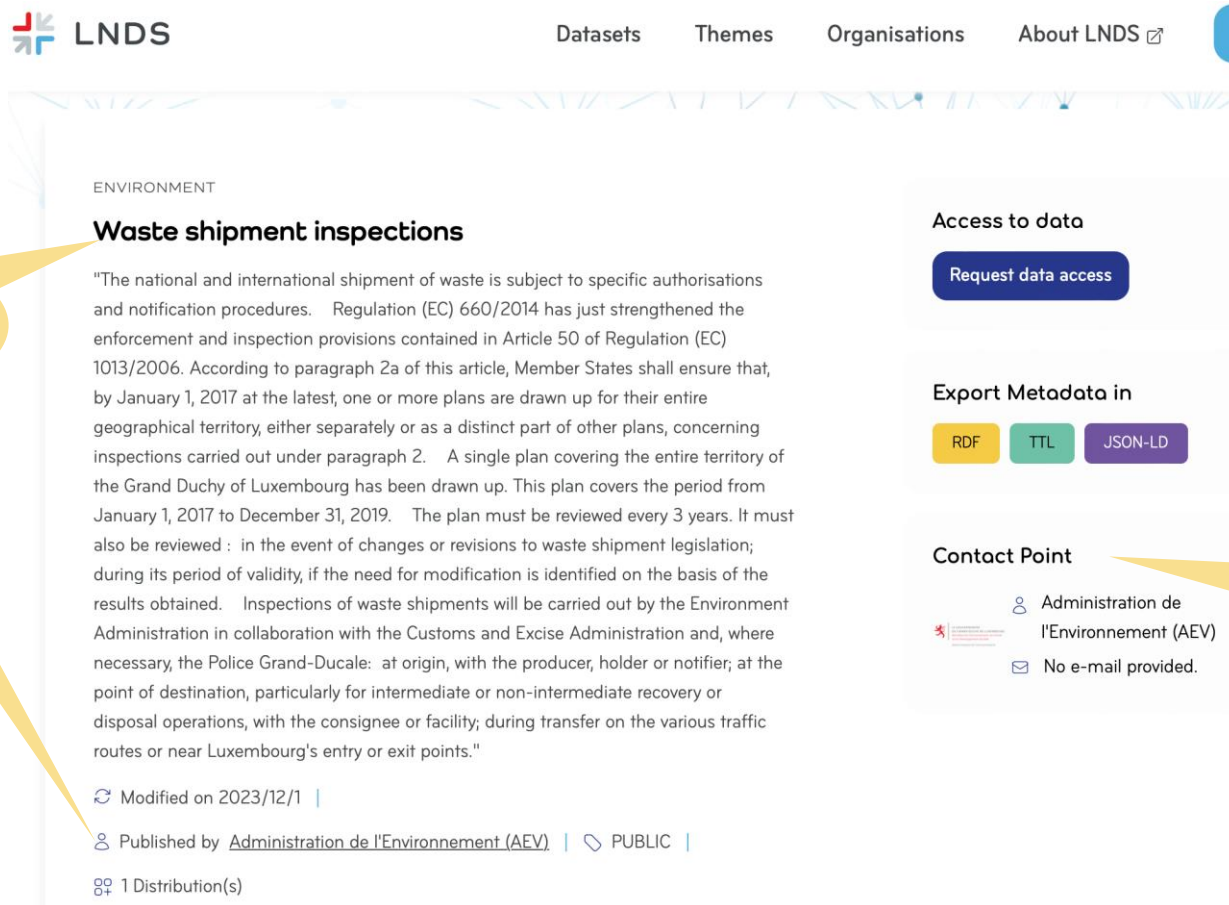
¹ [Sphagnum papillosum at Sphagnum cultivation at Universität Greifswald 2023-06-11 01.jpg](https://www.greifswald.de/wordpress/wp-content/uploads/2023/06/11_01.jpg)

² <https://en.wikipedia.org/wiki/Metadata#/media/File:Schlagwortkatalog.jpg>

³ https://en.wikipedia.org/wiki/Metadata#/media/File:Cenchrus_ciliaris_L._438045083.jpg

Metadata at the LNDC

As of 12 Jan 2024



The screenshot shows the LNDS website interface. At the top, there is a navigation bar with the LNDS logo and links for 'Datasets', 'Themes', 'Organisations', and 'About LNDS'. The main content area is titled 'ENVIRONMENT' and features a dataset entry for 'Waste shipment inspections'. The entry includes a detailed description of waste shipment regulations in Luxembourg, a 'Request data access' button, and options to export metadata in RDF, TTL, or JSON-LD formats. A contact point is listed as 'Administration de l'Environnement (AEV)' with a note that no email is provided. The page also shows a modification date of 2023/12/1 and is published by the 'Administration de l'Environnement (AEV)' as a public dataset with one distribution.

ENVIRONMENT

Waste shipment inspections

"The national and international shipment of waste is subject to specific authorisations and notification procedures. Regulation (EC) 660/2014 has just strengthened the enforcement and inspection provisions contained in Article 50 of Regulation (EC) 1013/2006. According to paragraph 2a of this article, Member States shall ensure that, by January 1, 2017 at the latest, one or more plans are drawn up for their entire geographical territory, either separately or as a distinct part of other plans, concerning inspections carried out under paragraph 2. A single plan covering the entire territory of the Grand Duchy of Luxembourg has been drawn up. This plan covers the period from January 1, 2017 to December 31, 2019. The plan must be reviewed every 3 years. It must also be reviewed : in the event of changes or revisions to waste shipment legislation; during its period of validity, if the need for modification is identified on the basis of the results obtained. Inspections of waste shipments will be carried out by the Environment Administration in collaboration with the Customs and Excise Administration and, where necessary, the Police Grand-Ducale: at origin, with the producer, holder or notifier; at the point of destination, particularly for intermediate or non-intermediate recovery or disposal operations, with the consignee or facility; during transfer on the various traffic routes or near Luxembourg's entry or exit points."

Modified on 2023/12/1 |

Published by [Administration de l'Environnement \(AEV\)](#) | PUBLIC |

1 Distribution(s)

Access to data

Request data access

Export Metadata in

RDF TTL JSON-LD

Contact Point

Administration de l'Environnement (AEV)

No e-mail provided.

Descriptive

Administrative

Data dictionary

Add meaning to your data

Structural

PRO_HDR_TBL (purchase Orders)

Column		Data type	References	Descriptions	Status	PII
ID		NUMBER		Row ID	VALID	Non-PII
PO_DT		DATE		Purchase order date	VALID	Non-PII
PO_REF	Reference	VARCHAR		Document reference	VALID	Non-PII
T_CODE		VARCHAR	PO_TYPES	Document type	VALID	Non-PII
STAT	Status	VARCHAR		A-approved, P-pending	VALID	Non-PII
VENDOR_ID		NUMBER	PO_VENDORS	Vendor	VALID	Non-PII
TERMS	payment terms	NUMBER	PY_TERMS	Payment terms	VALID	Non-PII
CCY	(Currency)			Currency	VALID	Non-PII
BILL_TO			PO_ADDR	Bill-to address	VALID	PII
ATTR1					DEPR.	
ATTR2		VARCHAR			DEPR.	
ATTR3	Buyer	VARCHAR			VALID	PII

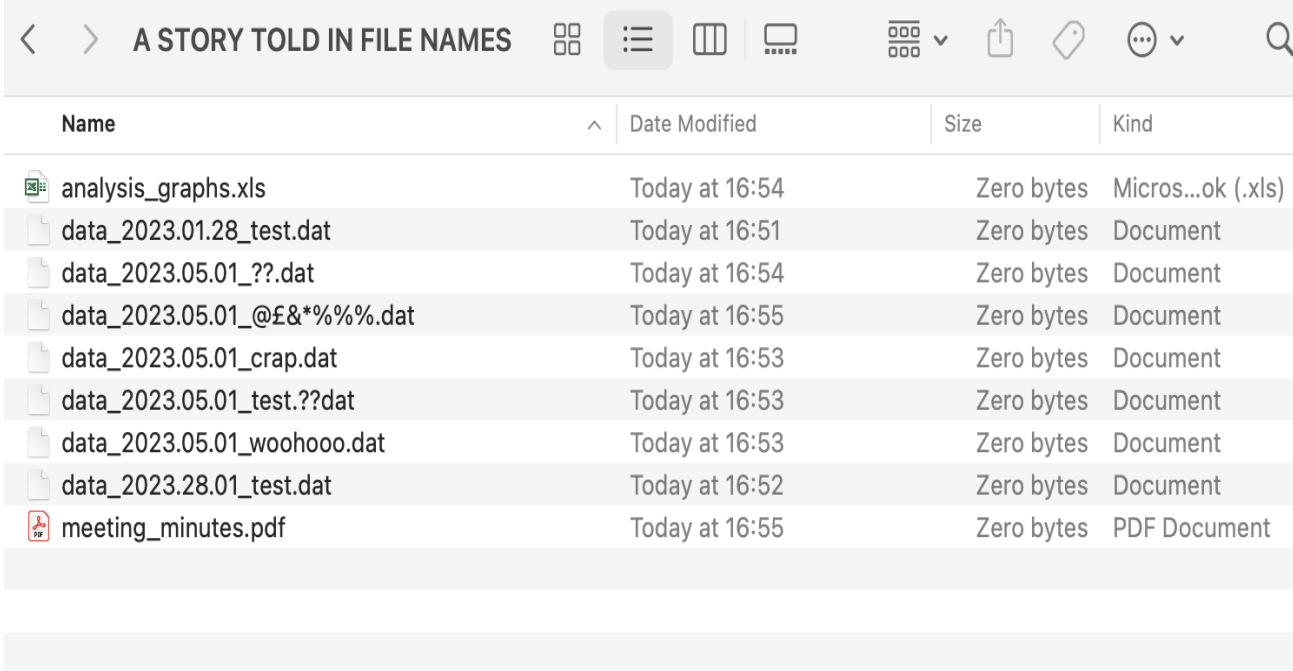
Payment terms
Period allowed to a buyer
to pay off the amount
due.



Data documentation – File naming

→ File naming is a very universal and basic mean to provide metadata

! Be careful in naming your files



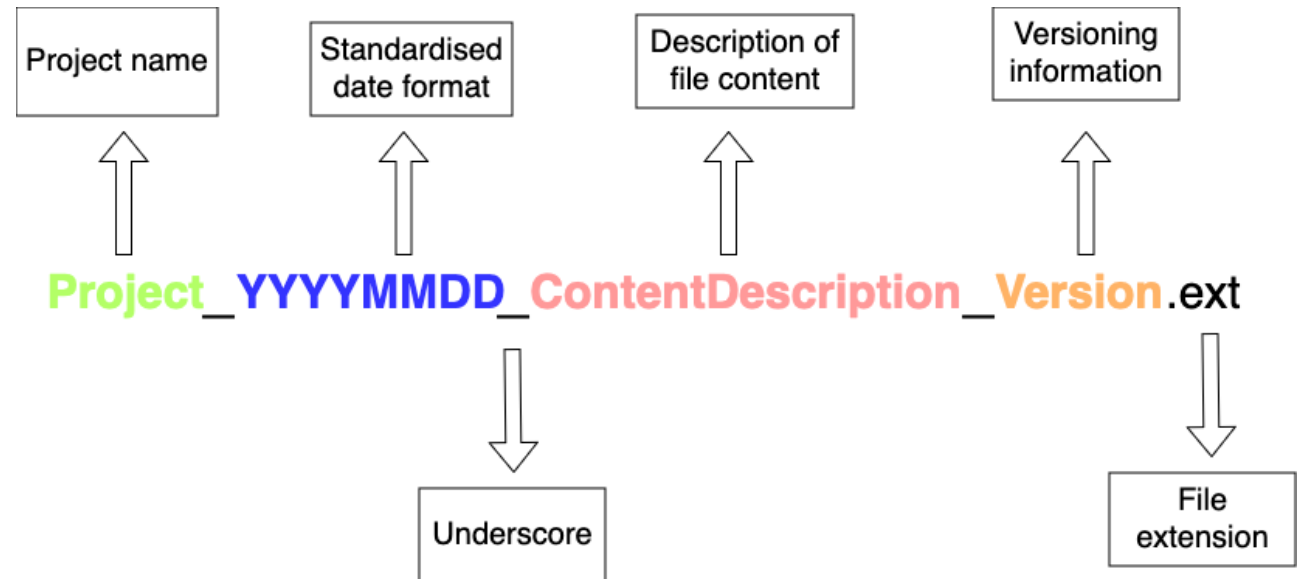
A screenshot of a file explorer window titled "A STORY TOLD IN FILE NAMES". The window displays a list of files with columns for Name, Date Modified, Size, and Kind. The files listed are:

Name	Date Modified	Size	Kind
analysis_graphs.xls	Today at 16:54	Zero bytes	Micros...ok (.xls)
data_2023.01.28_test.dat	Today at 16:51	Zero bytes	Document
data_2023.05.01_???.dat	Today at 16:54	Zero bytes	Document
data_2023.05.01_@£&*%%.dat	Today at 16:55	Zero bytes	Document
data_2023.05.01_crap.dat	Today at 16:53	Zero bytes	Document
data_2023.05.01_test.???.dat	Today at 16:53	Zero bytes	Document
data_2023.05.01_woohooo.dat	Today at 16:53	Zero bytes	Document
data_2023.28.01_test.dat	Today at 16:52	Zero bytes	Document
meeting_minutes.pdf	Today at 16:55	Zero bytes	PDF Document



An example of the file naming

- Guidelines to use files names in a principled way
- Write date following ISO 8601 standard (YYYY-MM-DD)
- Meaningful names

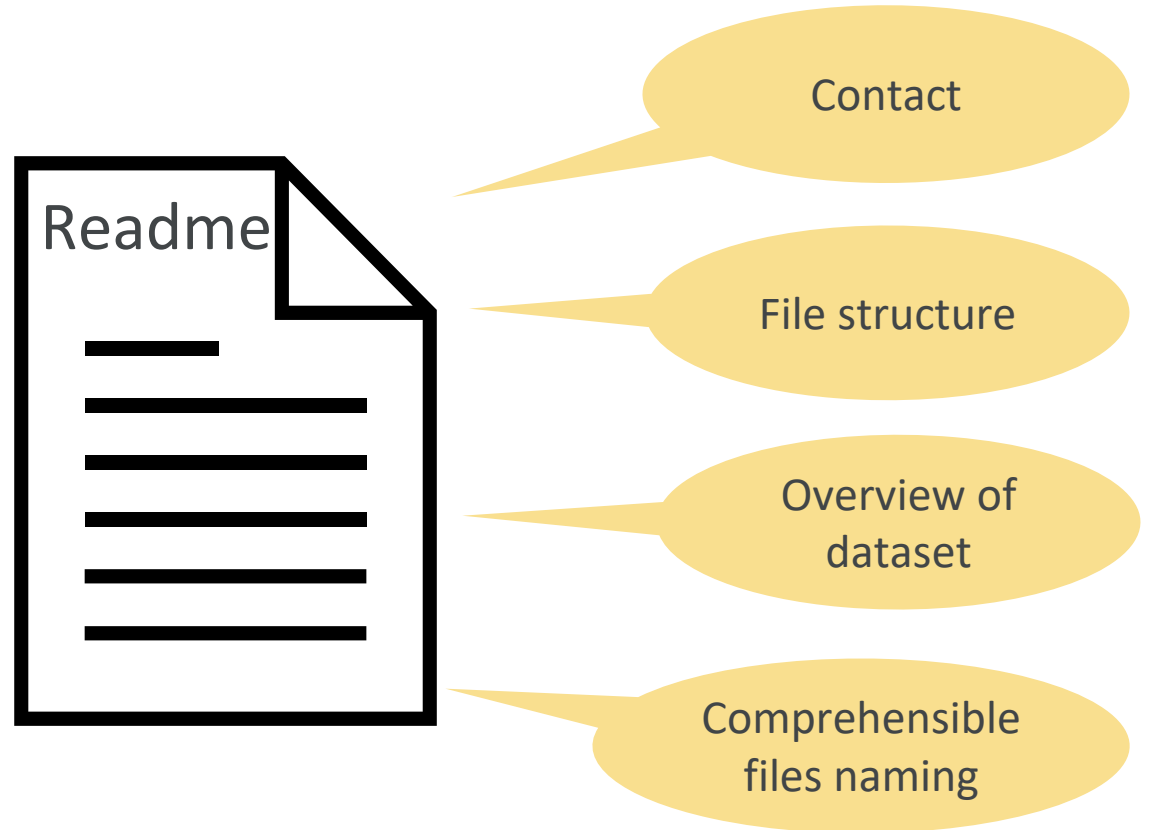


Data documentation - Readme file

Definition and content

- Provides a human-readable description of dataset big picture
 - Commonly meant to support rich metadata

What should be included in the README file?

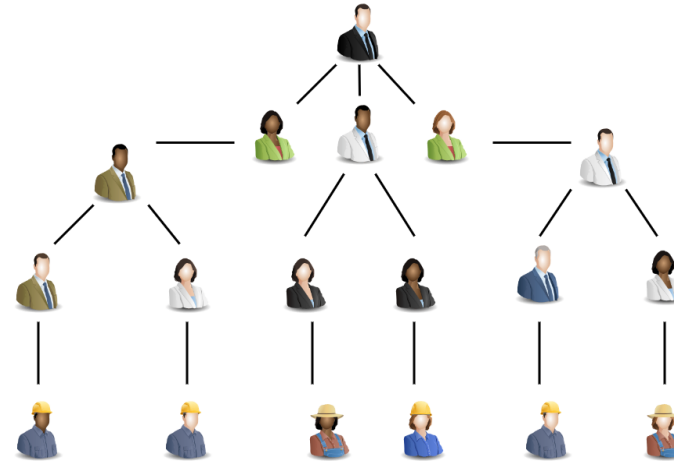


Data structure

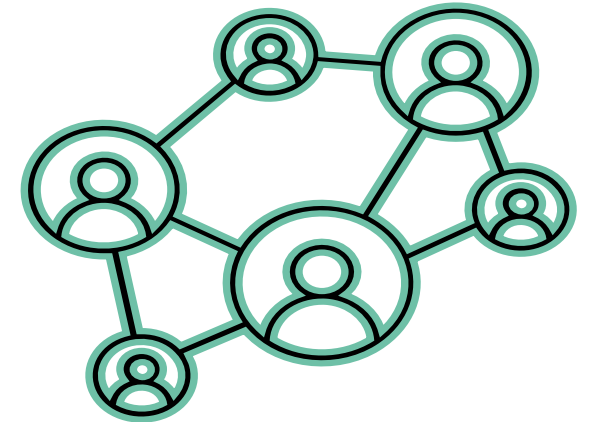
Common Data Structure

Customer ID	Product name	Quantity
101	Iphone 12	2
102	Samsung Galaxy S20	1
106	Macboock Pro	1
103	Ipad	3
104	Iphone 12	1

Tabular Data
: e.g. Customer purchase data



Hierarchical Data : e.g.
Company organisational chart



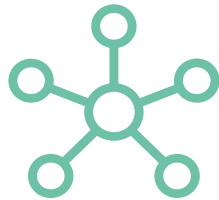
Network Data : e.g.
Social network

Data format

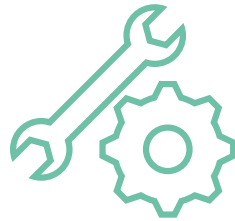
Data format definition and standards

→ The way in which the data is structured and made available for humans and machines

→ Criteria for choosing data format:



Structure



processing



Easiest to share

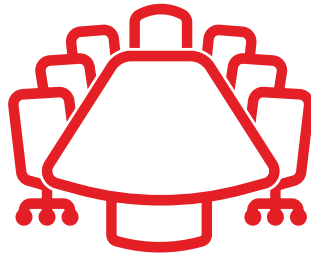
Data Format

Data standards

→ Non-proprietary or open formats are more interoperable (e.g. CSV).



Usability



Management



Access



Data format

Which format for which types of data

Type of data	Recommended formats
Text	Extensible Markup Language (.xml), Hypertext Markup Language (.html)
Tables, spreadsheets, and databases	Comma-separated values (.csv)
Image files	TIFF (.tiff or .tif), JPEG (.jpg or .jp2), Portable Network Graphics (.png), Scalable Vector Graphics (.svg)
Sound files	WAVE (.wav) , MPEG-3 (.mp3)
Web data	Javascript Object Notation (.json), Extensible Markup Language (.xml), Hypertext Markup Language (.html)
Geospatial data	Geo-Referenced TIFF (.tiff)

Data store

Definition

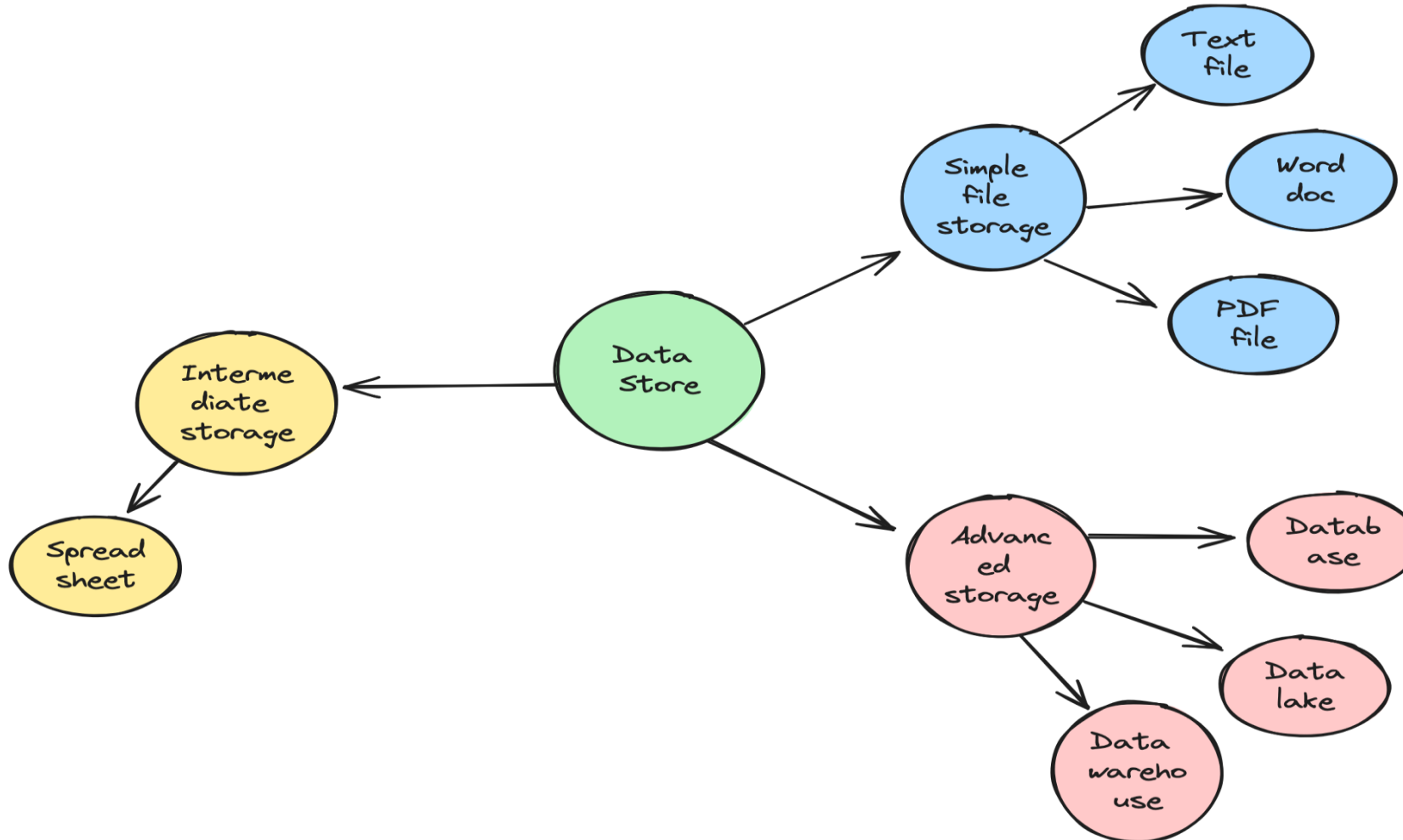
- Hierarchical organisation of data: bit, field, record, file, database
- A **data store** is "a repository for storing data that allows for data management, processing, and analysis - "Technopedia"
- Can be file, database, warehouse, data lake, or other forms of storage systems

Datastore



Data store

Hierarchical data store





Data classification

Data classification

- Data classification is the process of separating and organising data into relevant groups
- Necessary when data is identified as a first-class citizen, requiring specific attention and management
- Organisations typically design their own data classification models

Data can share characteristics such as their "level of sensitivity"





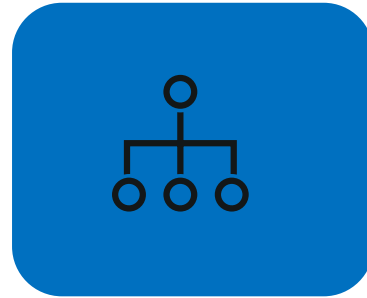
Example of data classification based on sensitivity level



Public

Data that may be freely disclosed to the public

Vaccination schedule,
Contact information



Internal

Internal data not meant for public disclosure

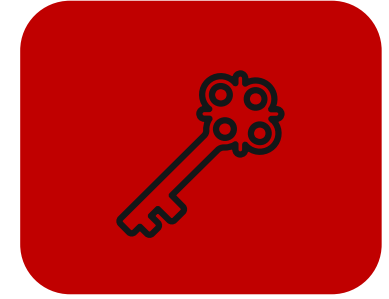
Organisational chart,
emails



Confidential

Sensitive data that if compromised could negatively affect operations

Patient medical record,
employee reviews



Secret

Highly sensitive data that if compromised could put the organisation in financial or legal risk

Biometric data,
genetic data

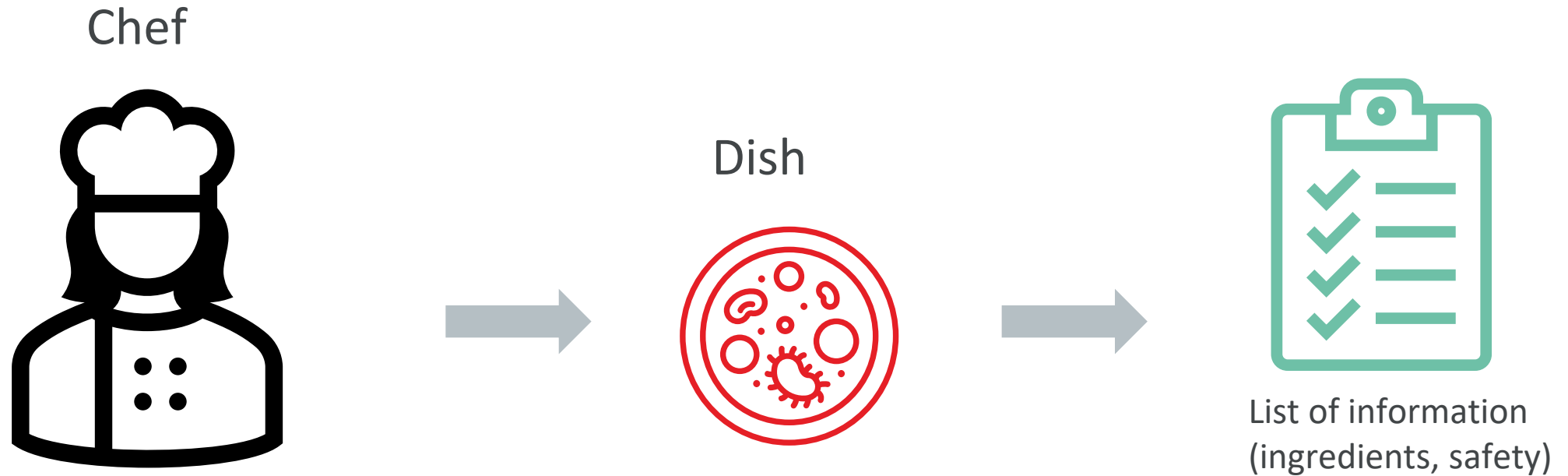


Data Management (DM)

→ The mindful and active data handling throughout project or analysis lifecycle



Data Stewardship (DS)

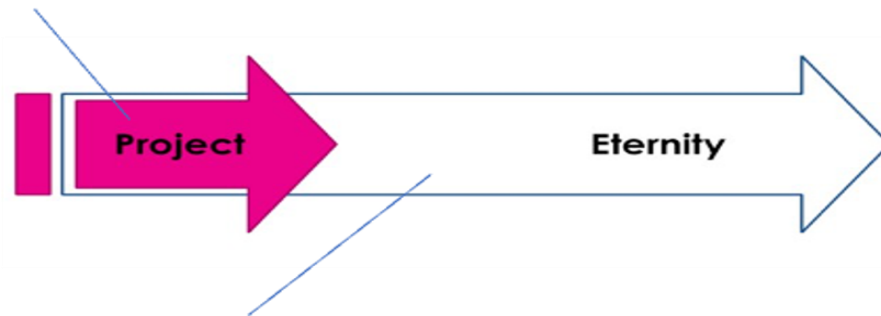


Data stewardship involves implementing and monitoring the processes and rules that a data organisation has in place regarding the management of its data



Data management vs stewardship

- Data management (DM) is **operational, data-related activities** in any phase of the **data lifecycle** including data's creation, collection, storage, quality control, sharing



- Data stewardship (DS) includes the notion of '**long-term care**' of valuable digital assets, with the goal that they should be **discovered and re-used for downstream analyses**, either alone or in combination with newly generated data. Data stewardship includes the assignment of responsibilities in, and planning of, data management.



Why do we need Data Management and Data Stewardship?

- The need of implementing data mandates
- AI is becoming mainstream
- Ensuring accountability with data exchange
- Reproducibility purposes
- **FAIR principles**



Findable

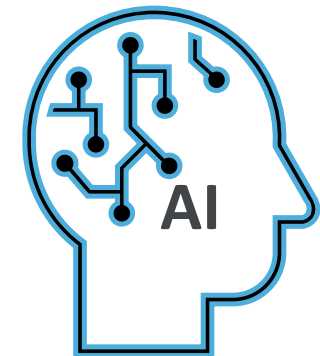
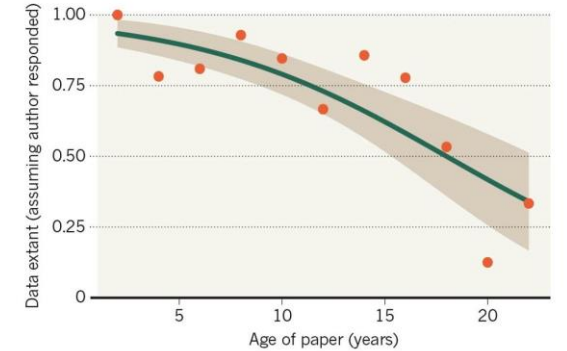
Accessible

Interoperable

Reusable

MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.



FAIR principles

FAIR Data

Findable

Metadata and data should be findable for both humans and computers

Interoperable

Data needs to work with applications or workflows for analysis, storage and processing

F

A

I

R

Accessible

Once found, users need to know how the data can be accessed

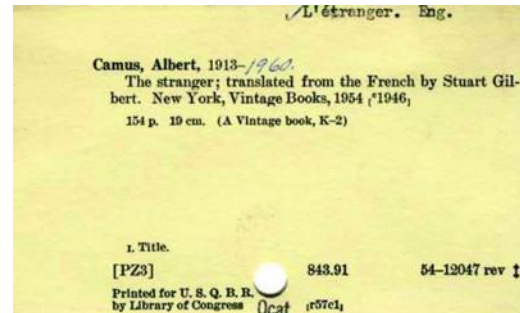
Reusable

The goal of FAIR is to optimise data reuse via comprehensive well-described metadata

FAIR principles

Findable

- (Meta)data
- Unique and persistent identifiers for (meta)data
- Indexed in a searchable resource
- Metadata contains data identifier



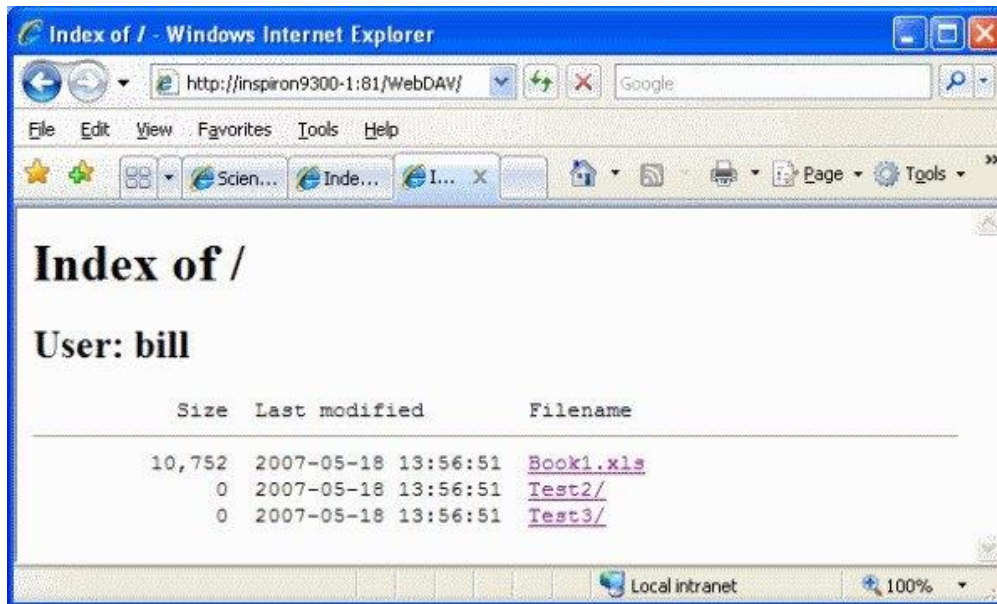
Hibsh, D., Schori, H., Efroni, S. & Shefi, O. *Figshare*
<http://dx.doi.org/10.6084/m9.figshare.1289242> (2015).

NCBI Sequence Read Archive [SRP059260](https://www.ncbi.nlm.nih.gov/sra/SRP059260) (2015).

FAIR principles

Not (so) Findable

- (Meta)data
- Identifiers for (meta) data
- Indexed in a searchable resource



Name	Date Modified	Size	Kind
analysis_graphs.xls	Today at 16:54	Zero bytes	Micros...ok (.xls)
data_2023.01.28_test.dat	Today at 16:51	Zero bytes	Document
data_2023.05.01_???.dat	Today at 16:54	Zero bytes	Document
data_2023.05.01_@£&*%?.dat	Today at 16:55	Zero bytes	Document
data_2023.05.01_crap.dat	Today at 16:53	Zero bytes	Document
data_2023.05.01_test.???.dat	Today at 16:53	Zero bytes	Document
data_2023.05.01_woohooo.dat	Today at 16:53	Zero bytes	Document
data_2023.28.01_test.dat	Today at 16:52	Zero bytes	Document
meeting_minutes.pdf	Today at 16:55	Zero bytes	PDF Document

Pinar Alper. (2021, June 17). Introduction to FAIR principles. Zenodo. <https://doi.org/10.5281/zenodo.5078286>

FAIR principles

Accessible

- (Meta)data are retrievable by a protocol
- Open, free, universally implementable
- Authentication/Authorization
- Metadata available even when data is not



Hibsh, D., Schori, H., Efroni, S. & Shefi, O. *Figshare*
<http://dx.doi.org/10.6084/m9.figshare.1289242> (2015).



HOME | HANDBOOK | FACTSHEETS | FAQs | RESOURCES | USERS | NEWS | MEMBERS AREA

Resolve a DOI Name

doi:

Go

A DOI is a unique persistent identifier for a published digital object

Pinar Alper. (2021, June 17). Introduction to FAIR principles. Zenodo. <https://doi.org/10.5281/zenodo.5078286>

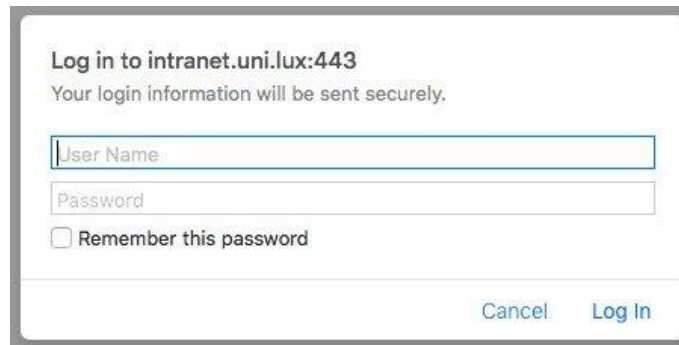


FAIR principles

Not (so) Accessible

- (Meta)data are retrievable by a protocol
- Open, free, universally implementable
- Authentication/Authorization
- Metadata available even when data is not

Data are available on request due to privacy or other restrictions



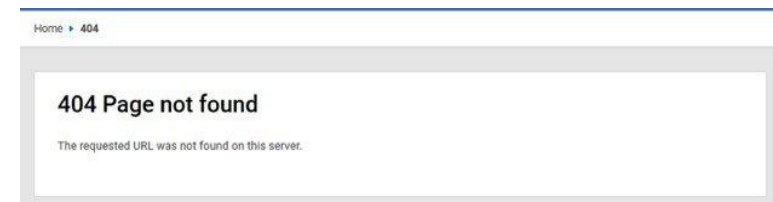
Log in to intranet.uni.lux:443
Your login information will be sent securely.

User Name

Password

Remember this password

Cancel Log In



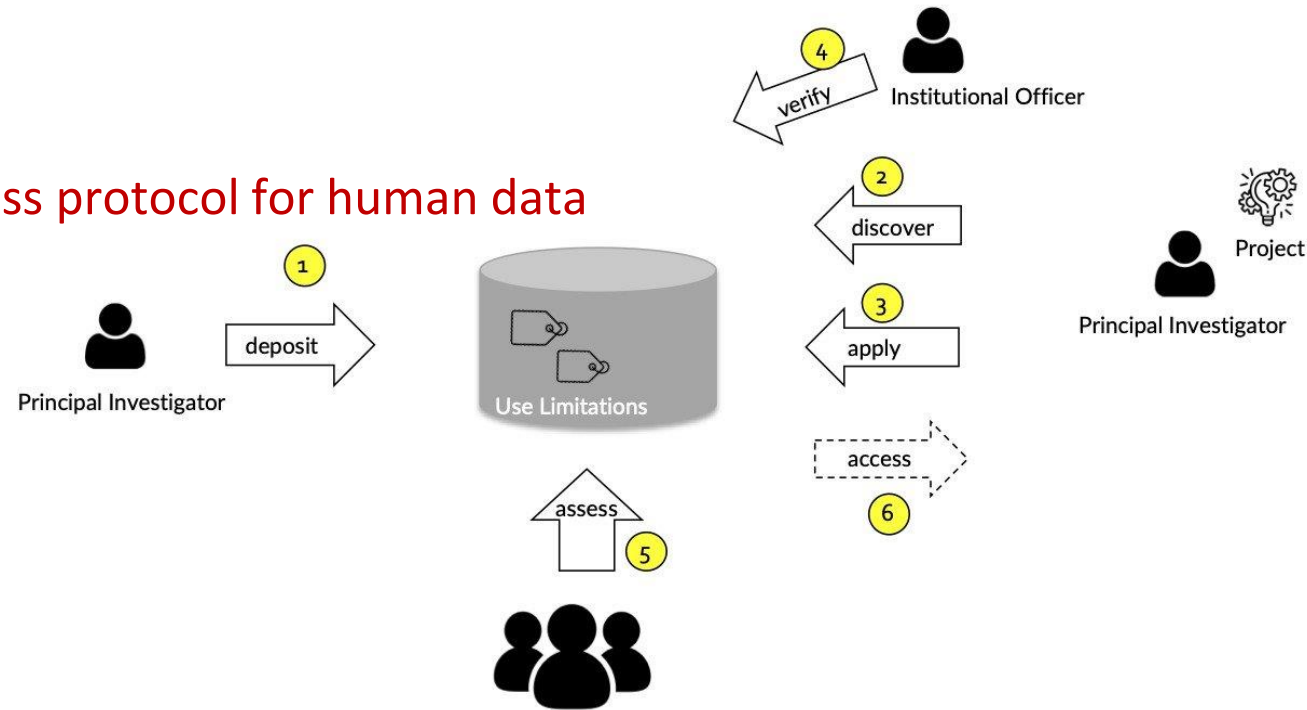
FAIR principles

Accessible

→ Accessible ≠ Unrestricted for all

→ Accessible ≠

Access protocol for human data



FAIR principles

Interoperable

- (Meta)data represented in formal , shared language
- Machine actionable
- Controlled vocabulary Tumour ≠ Tumor
- Community formats and standards



FAIR principles

Not (so) Interoperable

Customer purchase data

Customer ID	Product name	Quantity
101	Iphone 12	2
102	Samsung Galaxy S20	1
106	Macboock Pro	1
103	Ipad	3
104	Iphone 12	1

Product inventory data

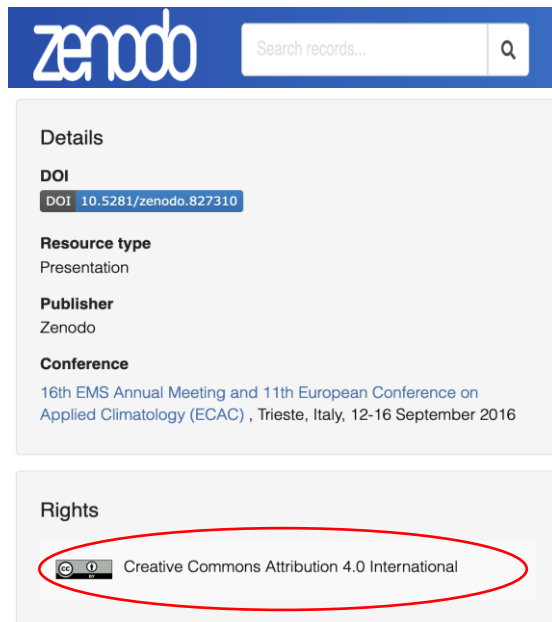
Product ID	Product Name	Stock quantity	Supplier
1001	Iphone twelve	20	Supplier X
1002	Samsung Galaxy S20	30	Supplier Y
1003	MacBook Pro	20	Supplier Z
1004	ipad	25	Supplier X
1005	Airpods	40	Supplier Y



FAIR principles

Reusable

- Descriptive metadata, following community guidelines
- Provenance of data
- Clear and accessible data use licence



zenodo Search records... Q

Details

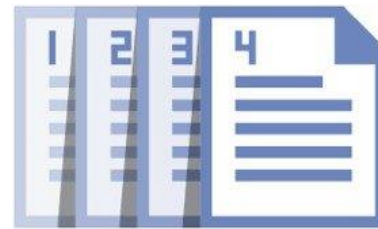
DOI
DOI 10.5281/zenodo.827310

Resource type
Presentation

Publisher
Zenodo

Conference
16th EMS Annual Meeting and 11th European Conference on Applied Climatology (ECAC) , Trieste, Italy, 12-16 September 2016

Rights
Creative Commons Attribution 4.0 International



Data versions



Access to data

Request data access

Export Metadata in

RDF

TTL

JSON-LD

Contact Point



Administration de l'Environnement (AEV)



Created on 01.12.2023



Updated on 01.12.2023

FAIRification is an expertise

The screenshot shows the FAIR Cookbook interface. At the top, there is a navigation bar with a back arrow, the title 'FAIRCOOKBOOK', a GitHub icon, and a search bar labeled 'Search Wizard...'. On the left, a sidebar lists various sections under 'FOREWORD' and 'RECIPES AT A GLANCE'. The main content area features a recipe card for '1. Unique, persistent identifiers'. The card includes a 'Recipe Overview' section with icons for 'Reading Time 30 minutes', 'Executable Code No', and 'Difficulty' (represented by four water droplets). The main body of the card is titled 'Introducing unique, persistent identifiers' and contains details about the 'Recipe Type' (Background information), 'Audience' (Principal Investigator, Data Manager, Data Scientist), and 'Maturity Level & Indicator' (DSM-1-C0). A 'Cite me with FCB006' button is located at the bottom right of the card. Below the card, the text states: 'The FAIR principles, under the Findability and the Accessibility chapters respectively, state that:'. Two principles are listed: 'F1. (Meta)data are assigned a globally unique and persistent identifier' and 'A1. (Meta)data are retrievable by their identifier using a standardised communications protocol'. The section '1.1. Main Objectives' is highlighted, and the text explains: 'The main goals of this recipe are therefore: To understand the purpose of a globally unique and persistent identifier and how they can be used to retrieve the associated (meta)data using a standardized communication protocol. To provide explanations on how to generate globally unique identifiers, explain what IRIs are and how they can be generated, retrieved and resolved.'



Data Management part of bigger landscape

- Data reuse underscores the necessity of Data Management (DM) and Data Stewardship (DS).
- DM and DS are parts of a bigger landscape.
- Data go into workflow and some automation is needed.
- It is crucial to consider additional data artefacts such as coding and analysis.



Idea



To Do



Doing



Done



Coding and analysis

- One of the most important things when dealing with data is code workflow and analysis.
- People are only familiar with spreadsheets.
- Familiarise yourself with analysis languages such as Python, and R.
- Combine spreadsheet and code analysis.

Spreadsheet	Coding
Manual and repetitive tasks	Automation of repetitive tasks
Visually intuitive	Text based and linear
Obscure the computational process	Computational process is explicit
Memory limitation	Analysis of much larger datasets

Solving issues when dealing with data

A beginner's guide

- Understand your problem
- Break down problems into small steps
- Identify parts of the problem you can solve
- Formulate effective searches
- Always read the official docs
- Don't ignore error messages
- Follow best practices when asking for help



Takeaways

- Data plays a critical role in our society.
- It is important to differentiate between data which is raw, information which is data that have been put into a context and knowledge which is actionable information.
- Choosing the right format for your data can help you organise your data efficiently. Data standards help improve data quality, consistency and interoperability.
- Different types of data stores including files, databases, data warehouses, etc exists and serve of various purposes.
- Each Organisation typically design their own data classification models. Classifying data based on the sensitivity levels helps with compliance and enhances data security.
- The FAIR principles (Findable, Accessible, Interoperable, Reusable) guide best practices in data management, ensuring data optimal reuse of data.



LNDS

LUXEMBOURG NATIONAL DATA SERVICE

