



Introduction to scientific metadata

Love Data Week

February 15, 2024

Data Science and Machine Learning Group

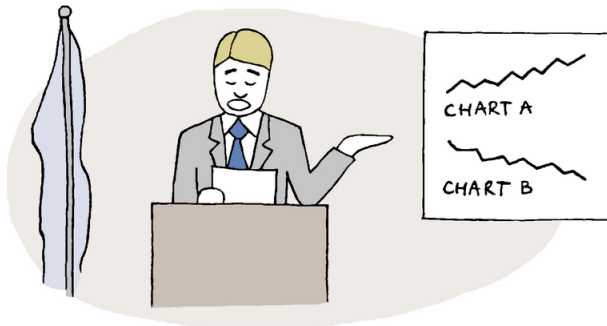
Julia Matela

julia.matela@hs-wismar.de



Introduction

Why do we need metadata?



AS YOU CAN SEE, OUR EXPORTS ARE GROWING
AND UNEMPLOYMENT IS FALLING. OR IT'S THE
OTHER WAY AROUND. I DON'T HAVE ANY METADATA...

 Dataedo /cartoon

Piotr@Dataedo

Fig. 1: Value of Metadata for the Public by Piotr Kononow¹

¹<https://dataedo.com/cartoon/tag/metadata>

Introduction to scientific metadata



Overview

1. Data and metadata

- What is data?
- What is metadata?
- history of metadata
- examples of metadata

2. Types of metadata

- Descriptive metadata
- Structural metadata
- Administrative metadata

3. Controlled vocabularies and ontologies

- Controlled vocabularies
- Ontologies

4. Standards and best practices

- Dublin Core
- METS
- PREMIS
- Best Practices
- Compliance with FAIR principle

Data

What is data?



- raw, unprocessed facts or figures
- lack context and meaning
- various forms, including numbers, text, images, sound

Data

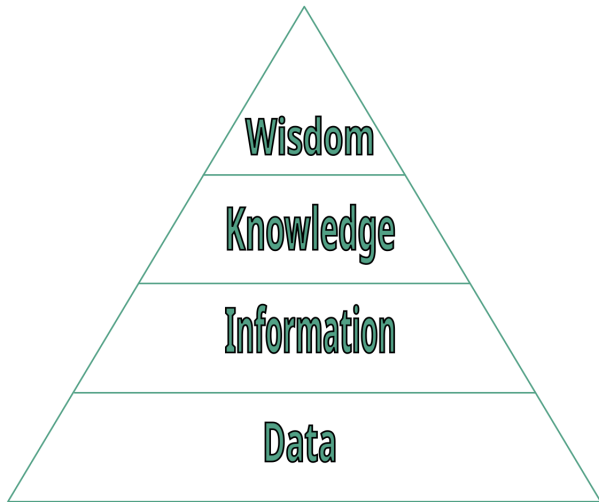
Data are measurements or observations that are collected as a source of information. There are a variety of different types of data, and different ways to represent data.²

Australian Bureau of Statistics

²<https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/data>

Data

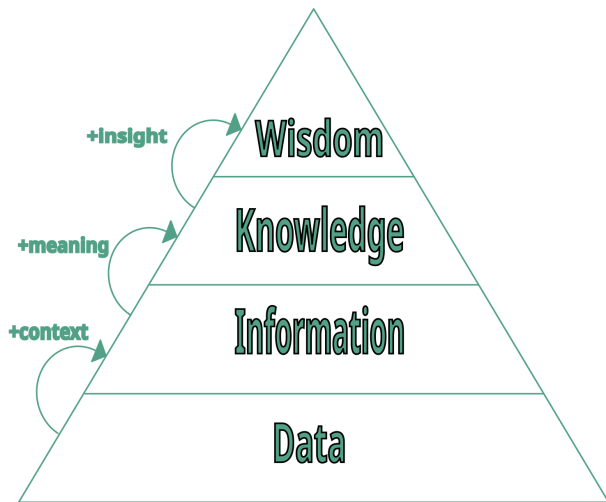
Data-Information-Knowledge-Wisdom pyramid



- **Wisdom** is the ability to make well-informed decisions and take effective action based on understanding of the underlying knowledge.
- **Knowledge** is the result of analyzing and interpreting information to uncover patterns, trends, and relationships. It provides an understanding of how and why certain phenomena occur.
- **Information** is organized, structured, and contextualized data. Information is useful for answering basic questions like who, what, where, and when.
- **Data** refers to raw, unprocessed facts and figures without context. It is the foundation for all subsequent layers but holds limited value in isolation.

Data

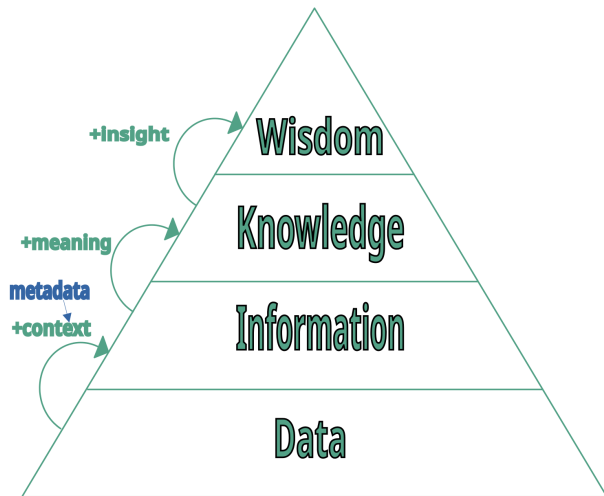
Data-Information-Knowledge-Wisdom pyramid



- **Data** refers to raw, unprocessed facts and figures without context. It is the foundation for all subsequent layers but holds limited value in isolation.
- **Example:**
 - sequence of numbers 15022024

Data

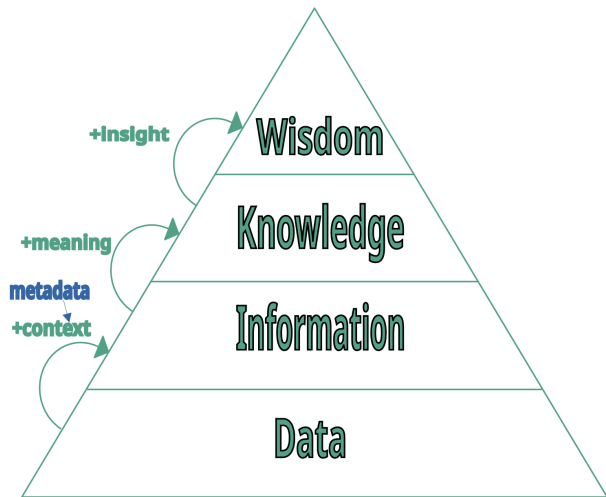
Data-Information-Knowledge-Wisdom pyramid



- **Information** is organized, structured, and contextualized data. Information is useful for answering basic questions like who, what, where, and when.
- **Example:**
 - sequence of numbers 15022024
 - a date
 - 15th February 2024

Data

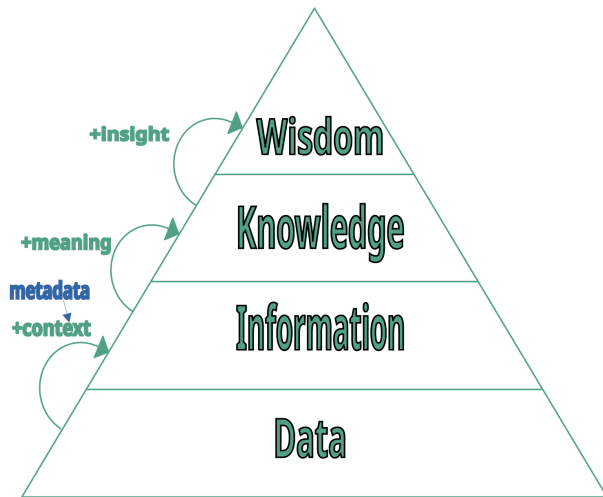
Data-Information-Knowledge-Wisdom pyramid



- **Knowledge** is the result of analyzing and interpreting information to uncover patterns, trends, and relationships. It provides an understanding of how and why certain phenomena occur.
- **Example:**
 - sequence of numbers 15022024
 - a date
 - 15th February 2024
 - today's date
 - a day of presentation about scientific metadata

Metadata

Data-Information-Knowledge-Wisdom pyramid



- **Wisdom** is the ability to make well-informed decisions and take effective action based on understanding of the underlying knowledge.
- **Example:**
 - sequence of numbers 15022024
 - a date
 - 15th February 2024
 - today's date
 - a day of presentation about scientific metadata
 - participating in today's presentation

Metadata

What is metadata?



Metadata

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource.

National Information Standards Organization, NISO (2004)

- data that provides information about other data
- describes various aspects of a dataset, document, or resource
- makes it easier to understand, manage, and discover
- includes details such as the title, author, date created, file format, and keywords

DATA



METADATA



 Dataedo /cartoon

Piotr@Dataedo

Fig. 2: Data vs Metadata (8) by Piotr Kononow³

- **bottle label** turns the bottle from just a container with something in it into an object with information about properties, functions and potential uses
- **metadata** turns unstructured, noisy data into meaningful digital objects, which makes their exchange and management easier.

³<https://dataedo.com/cartoon/tag/metadata>

DATA



METADATA



 Dataedo /cartoon

Piotr@Dataedo

Fig. 3: Data vs Metadata (2) by Piotr Kononow⁴

- a **map** guides us to the hidden treasure
- **metadata** guides us to relevant Data

⁴<https://dataedo.com/cartoon/tag/metadata>

Metadata

history of metadata



Many of the concepts and techniques of metadata creation, management and use originated with the development of **library catalogues**. Books are repositories of information and a catalogue contains data about that information and can therefore be regarded as metadata.

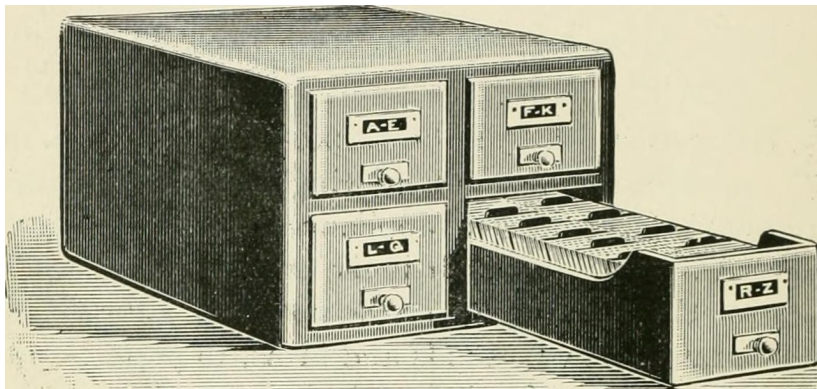


Fig. 4: Illustration from Manual of library classification and shelf arrangement, 1898⁵

⁵<https://www.flickr.com/photos/internetarchivebookimages/14781001892/>

Metadata

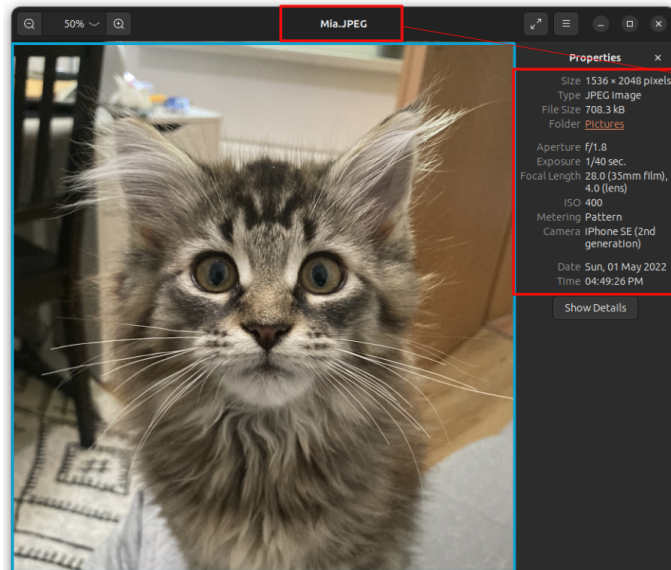
history of metadata



- The idea of cataloguing information has been around at least since the Alexandria Library in ancient Egypt (**ca. 3 century B.C.**)
- the term 'metadata' is fairly new. It became established in **1970s**
- **mid 1990s** – Dublin Core Metadata Initiative (DCMI) – established a standard for describing web content. It was one of the earliest attempts to standardize metadata elements for online resources, focusing on simplicity and broad applicability.
- In the **early 21st century**, the concept of Linked Data and the Semantic Web gained prominence. These ideas emphasized connecting and enriching data across the web through standardized, machine-readable metadata

Metadata – examples

A photo

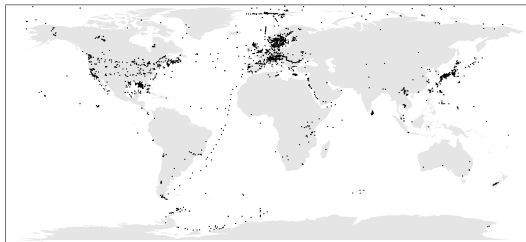
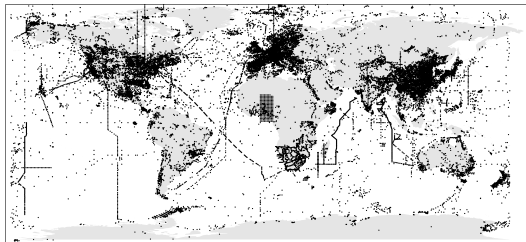


Metadata

Data

Metadata – examples

European Nucleotide Archive (ENA)



Available:

- ~2.4 million data sets
- ~32 000 studies

Easily re-usable:

- ~3.7% of data sets [1]
- ~2.7% of data volume [Mbp]

Fig. 5: Global distribution of metabarcoding data sets from presumably environmental samples (source: ENA; figure: Christiane Hassenrück)

Types of metadata

3 types of metadata

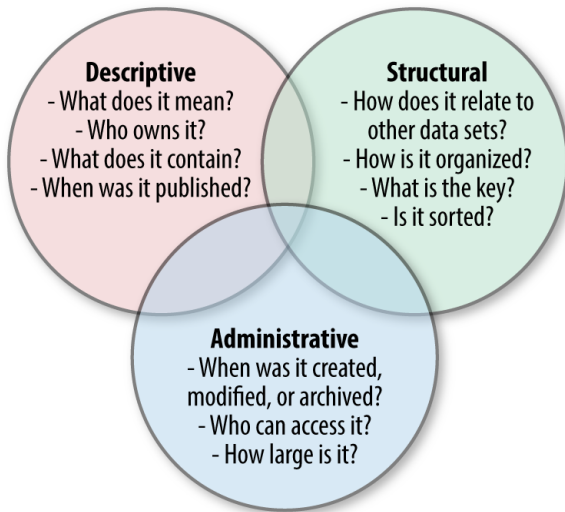


Fig. 6: Classes of metadata [3]

Types of metadata

Descriptive metadata



- focuses on providing information about the content of a resource
- helps to understand what the data is, who created it, when it was created, and what its subject or topic is
- is like the cover of a book, offering a snapshot of what's inside
- the most important purpose of descriptive metadata is resource discovery

Examples of descriptive metadata include:

- In a photography archive, descriptive metadata, including details like title, photographer, and description, offers a comprehensive overview of each image's content and context.
- Within a library catalog entry, descriptive metadata such as title, author, and summary provides a concise summary of a book's key attributes, aiding potential readers in their selection.

Types of metadata

Structural metadata



- outlines the organization and relationships between different parts of a resource
- It's like the table of contents or index of a book, and what its subject or topic is
- guides us on how individual pieces come together to form a cohesive whole
- The primary purpose of structural metadata is to facilitate the navigation, interpretation, and proper usage of complex information structures.

Examples of structural metadata include:

- In a database, structural metadata might define the relationships between tables, indicating how data is organized and linked.
- For multimedia files, structural metadata could specify the timing and sequencing of audio and video components.

Types of metadata

Administrative metadata



- handles the management and maintenance aspects of data
- includes information about the rights, permissions, and version history of a resource

Examples of administrative metadata include:

- User permissions on a shared cloud document. It determines who can view, edit, or share the document, ensuring that sensitive information is accessible only to authorized individuals.
- A collaborative project where you and your team are working on a shared document. Administrative metadata tracks changes, showing who made edits and when, allowing everyone to stay on the same page and revert to earlier versions if needed.

Controlled vocabularies and ontologies

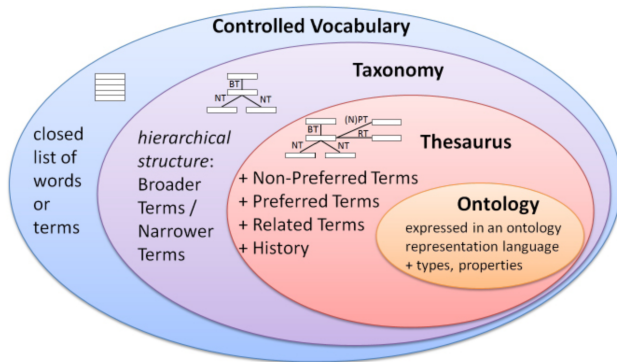


Fig. 7: Categories of classification [2]

Controlled vocabularies and ontologies

controlled vocabularies and ontologies stand as invaluable tools in the realm of scientific metadata, providing a structured and semantically rich foundation for effective communication, collaboration, and data representation. Their continued refinement and application promise to elevate the quality and impact of scientific research across disciplines.

Controlled vocabularies and ontologies

Controlled vocabularies



Controlled vocabularies

Controlled vocabularies are curated sets of standardized terms used to categorize and describe scientific concepts. Unlike free-form or open-ended vocabularies, controlled ones provide a structured framework, minimizing ambiguity and ensuring uniformity in the interpretation of data.

The benefits of controlled vocabularies:

- **Search and Retrieval:** Standardized terms enhance the discoverability of scientific datasets, making it easier for researchers to locate and utilize relevant information.
- **Interdisciplinary Collaboration:** By adopting controlled vocabularies, interdisciplinary collaboration becomes more seamless, as researchers can bridge gaps in terminology and better understand each other's work.

Controlled vocabularies and ontologies

Controlled vocabularies – Example



Medical Coding Systems In healthcare

In healthcare, coding systems such as ICD-10 (International Classification of Diseases, 10th Edition) employ controlled language. Specific codes denote diseases, symptoms, and procedures, providing a standardized and universally recognized language for healthcare professionals globally.

▼ ICD-10 Version:2019



▼ I Certain infectious and parasitic diseases

▼ A00-A09 Intestinal infectious diseases

▼ A00 Cholera

A00.0 Cholera due to *Vibrio cholerae* 01, biovar cholerae

A00.1 Cholera due to *Vibrio cholerae* 01, biovar eltor

A00.9 Cholera, unspecified

▶ A01 Typhoid and paratyphoid fevers

▶ A02 Other salmonella infections

▶ A03 Shigellosis

▶ A04 Other bacterial intestinal infections

▶ A05 Other bacterial foodborne intoxications, not elsewhere classified

▶ A06 Amoebiasis

▶ A07 Other protozoal intestinal diseases

Fig. 8: A snip from ICD-10 Version:2019⁶

⁶<https://icd.who.int/browse10/2019/en>

Controlled vocabularies and ontologies

Ontologies



Ontologies

Ontologies take controlled vocabularies a step further by incorporating semantic relationships and hierarchical structures. They define not only terms but also the relationships between them, providing a richer context for scientific concepts.

The benefits of ontologies:

- **Semantic Precision:** Ontologies allow researchers to express not just what a term is but also how it relates to other terms, adding semantic precision to data representation.
- **Machine Understanding:** Ontologies enable machines to understand the meaning of terms, fostering improved data integration and automated reasoning.

Controlled vocabularies and ontologies

Ontologies – Example



Gene Ontology (GO)

The Gene Ontology is a widely used example in genomics. It categorizes genes into biological processes, cellular components, and molecular functions. For instance, a gene might be classified under “cell division” as a biological process, “nucleus” as a cellular component, and “kinase activity” as a molecular function.

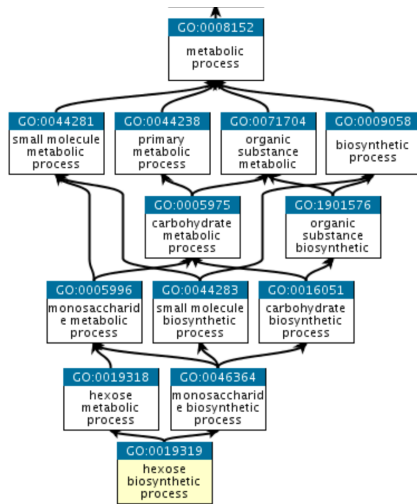


Fig. 9: The GO graph⁷

⁷<https://geneontology.org/docs/ontology-documentation/>

Standards and Best Practices in Metadata

Overview of popular metadata standards



Why do we need Standards?

Imagine a library where every book follows a different cataloging system. Chaos would ensue when trying to find specific books. Metadata standards are formal documents that establish uniform criteria, methods, processes and practices. They act as the common language, ensuring consistency and order. They enable seamless collaboration, interoperability, and data exchange across different systems and disciplines.

There are numerous schemas/ standards available. Every metadata schema has the susceptibility to be or to become a standard, so the expression metadata schema is interchangeable with metadata standard. An overview of existing standards is provided among others by:

- *the RDA Metadata Standards Directory Working Group*
- *the RDA Metadata Standards Catalog*

Standards and Best Practices in Metadata

Dublin Core



Fig. 10: Dublin Core Metadata Initiative (DCMI) Logo⁸

- **Dublin Core** serves as the foundational framework for metadata, acting as a universal language to describe a wide array of resources. Its simplicity and versatility make it highly accessible, and it has become a standard entry point for individuals new to metadata concepts.
- **Dublin Core** offers a set of straightforward elements that cover essential aspects of resource description. These elements include titles, authors, dates, and more. The simplicity of these elements facilitates easy adoption and integration across diverse datasets and projects.
- One of **Dublin Core**'s notable strengths lies in its widespread adoption across various domains. Its universal applicability makes it a go-to choice for describing resources in libraries, archives, museums, and on the web. This broad acceptance ensures that Dublin Core remains a pivotal element in the metadata landscape.

⁸<https://www.dublincore.org/>

Standards and Best Practices in Metadata

Dublin Core



Dublin Core Metadata for Resource Discovery [4]

Content

- Title
- Subject
- Description
- Type
- Source
- Relation
- Coverage

Intellectual Property

- Creator
- Publisher
- Contributor
- Rights

Instantiation

- Date
- Format
- Identifier
- Language

Standards and Best Practices in Metadata

METS (Metadata Encoding and Transmission Standard)



Fig. 11: METS (Metadata Encoding and Transmission Standard) Logo⁹

- **Metadata Encoding and Transmission Standard (METS)** focuses on the structural and encoding aspects of complex digital objects. It provides a standardized way to organize and link together various files or components within a digital resource. Think of METS as the architectural blueprint that ensures seamless integration of all the pieces within a digital collection.
- **METS** excels in addressing the structural intricacies of digital objects. It goes beyond basic descriptive elements, offering a comprehensive framework for expressing the relationships and organization of different components within a digital resource. This structural emphasis makes METS particularly valuable for projects dealing with complex digital materials.

⁹<https://www.loc.gov/standards/mets/mets-schemadocs.html>

Standards and Best Practices in Metadata

METS (Metadata Encoding and Transmission Standard)



Characteristics of METS

- An XML-Schema
- An open standard
- Developed by the library community
- Relatively simple
- Extensive
- Modular

Sections of METS document

- METS header
- Descriptive Metadata
- Administrative Metadata
- File Section
- Structural Map
- Structural Links
- Behavioral



```

-<mets:mets OBJID="8291" xsi:schemaLocation="http://www.loc.gov/standards/
premis/ http://www.loc.gov/standards/premis/v2/premis-v2-0.xsd http://www.loc.gov/
mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-7.xsd http://www.loc.gov/
METS/ http://www.loc.gov/standards/mets/mets.xsd http://www.loc.gov/standards/mix/
http://www.loc.gov/standards/mix/mix.xsd">
-<mets:metsHdr CREATEDATE="2023-12-09T15:28:31Z">
-<mets:agent OTHERTYPE="SOFTWARE" ROLE="CREATOR"
  TYPE="OTHER">
  <mets:name>Goobi - 447dc - 2022-10-17T06:34:30Z</mets:name>
  <mets:note>Goobi</mets:note>
</mets:agent>
</mets:metsHdr>
-<mets:dmdSec ID="DMDLOG_0000">
-<mets:mdWrap MDTYPE="MODS">
-<mets:xmlData>
-<mods:mods>
-<mods:titleInfo>
-<mods:title>
  Handschriftenkatalog der Universitätsbibliothek Greifswald
</mods:title>
</mods:titleInfo>
-<mods:name type="personal">
-<mods:role>
  <mods:roleTerm authority="marcrelator" type="code">aut</
  mods:roleTerm>
</mods:role>
<mods:namePart type="family">Pertz</mods:namePart>
<mods:namePart type="given">Karl August Friedrich</
  mods:namePart>
<mods:displayForm>Pertz, Karl August Friedrich</
  mods:displayForm>
</mods:name>
-<mods:name type="personal">
-<mods:role>
  <mods:roleTerm authority="marcrelator" type="code">aut</
  mods:roleTerm>
</mods:role>
<mods:namePart type="family">Muldener</mods:namePart>
<mods:namePart type="given">Wilhelm</mods:namePart>
<mods:displayForm>Muldener, Wilhelm</mods:displayForm>
</mods:name>
-<mods:originInfo>
-<mods:place>
  <mods:placeTerm type="text">Greifswald</mods:placeTerm>
</mods:place>
<mods:dateIssued encoding="w3cdtf" keyDate="yes">1872</
  
```

Dublin Core Metadata (oai_dc)

Title	Handschriftenkatalog der Universitätsbibliothek Greifswald
Author or Creator	Pertz, Karl August Friedrich
Author or Creator	Muldener, Wilhelm
Subject and Keywords	Allgemeines
Subject and Keywords	Handschriften
Subject and Keywords	Universitätsbibliothek Greifswald
Date	1872
Resource Type	Multivolume work
Resource Type	text
Format	image/jpeg
Format	application/pdf
Resource Identifier	https://www.digitale-bibliothek-mv.de/viewer/resolver?urn=urn:nbn:de:gbv:9-g-4879751
Source	Pertz, Karl August Friedrich, Muldener, Wilhelm: Handschriftenkatalog der Universitätsbibliothek Greifswald, Greifswald: - 1872.
Rights Management	info:eu-repo/semantics/openAccess

Fig. 12: MV – Digitale Bibliothek – METS and Dublin Core – Example: Pertz, Karl August Friedrich, and Wilhelm Muldener. Handschriftenkatalog Der Universitätsbibliothek Greifswald. 1872.¹⁰

¹⁰https://www.digitale-bibliothek-mv.de/viewer/toc/PPN772242178/1/LOG_0000/

Standards and Best Practices in Metadata

PREMIS (Preservation Metadata Implementation Strategies)



Fig. 13: Premis Logo¹¹

- **Preservation Metadata Implementation Strategies (PREMIS)** focuses specifically on supporting the preservation of digital objects and ensuring their long-term usability.
- It provides a comprehensive data dictionary that outlines metadata elements crucial for preserving the integrity, authenticity, and accessibility of digital content over time.
- **PREMIS** has gained recognition as an international standard for preservation metadata. Its guidelines and data dictionary offer a systematic approach to capturing and managing metadata essential for the ongoing preservation efforts of digital materials.

¹¹<https://www.dpconline.org/events/digital-preservation-awards/dpa2022-premis>

Standards and Best Practices in Metadata

PREMIS (Preservation Metadata Implementation Strategies)



The **PREMIS** Data Model consists of four main Entities:

- Object
 - Intellectual Entity
 - Representation
 - File
 - Bitstream
- Event
- Agent
- Rights

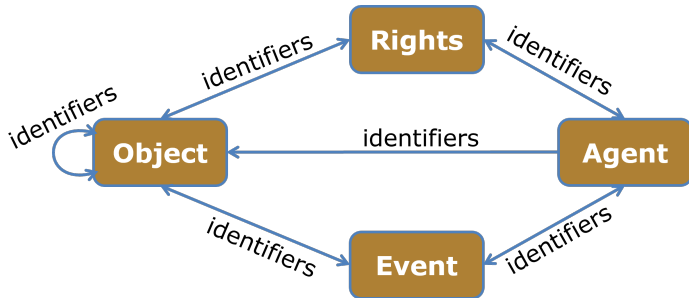


Fig. 14: The PREMIS Data Model¹²

¹²<https://www.dpconline.org/events/digital-preservation-awards/dpa2022-premis>

Standards and Best Practices in Metadata

Best practices for creating standardized and interoperable metadata



1. Adopt standardized metadata schemas
2. Identify core metadata elements
3. Use consistent terminology and controlled vocabulary
4. Provide detailed descriptions
5. Ensure data accessibility and use restrictions

Standards and Best Practices in Metadata

Compliance with FAIR principles (Findable, Accessible, Interoperable, Reusable)



1. Findable

- Clear Identification
- Searchable Keywords

2. Accessible

- Clear Usage Licenses
- Accessible Formats

3. Interoperable

- Standardized Metadata
- Crosswalks and Mapping

4. Reusable

- Well-Documented Metadata
- Community Engagement

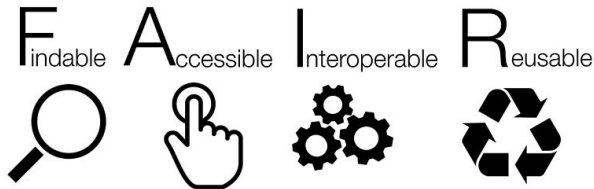


Fig. 15: FAIR data principles¹³

¹³https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg



- [1] Christiane Hassenrück et al. *FAIR enough? A perspective on the status of nucleotide sequence data and metadata on public archives*. Sept. 2021. DOI: 10.1101/2021.09.23.461561.
- [2] Sándor Kopácsi, Rastislav Hudak, and Raman Ganguly. "Implementation of a Classification Server to Support Metadata Organization for Long Term Preservation Systems". In: *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare* 70 (Sept. 2017), p. 225. DOI: 10.31263/voebm.v70i2.1897.
- [3] Ashish Thusoo and Joydeep Sen Sarma. *Creating a Data-Driven Enterprise with DataOps*. O'Reilly Media, Inc., Mar. 2017. ISBN: 9781491977835.
- [4] Misha Wolf et al. *Dublin Core Metadata for Resource Discovery*. RFC 2413. Sept. 1998. DOI: 10.17487/RFC2413. URL: <https://www.rfc-editor.org/info/rfc2413>.