

# Scientific metadata: Fundamentals of structured and standardized research data annotation

---

workshop @ Enabling reproducibility in data science - learn why it matters and how you can do it - 9 June 2022

---

Dear participants,

welcome to the HedgeDoc handout for our session about the “**Fundamentals of scientific metadata**”. This handout serves two purposes: first, you can bookmark the document and get back to its content at any time. Secondly, we will use it collaboratively to add and discuss results.

Best,  
Annika & Silke

## Tools

---




### HedgeDoc

We use this HedgeDoc handout to work on our notes collaboratively.



Here (<https://notes.desy.de/HslZpf2hT1uCv-6SbPc9uw#>) you can find a little HedgeDoc playground to get familiar with the tool & Markdown markup.

Some helpful things about HedgeDoc:

#### **You can change workspace view modes with the icons in the top left corner**

-  View: See only rendered HTML result.
-  Both: See Markdown text editor and rendered HTML result at the same time.
-  Edit: See Markdown editor only.

#### **Navigating the document**

While in -mode, you can see a table of contents on the top right corner of your screen. In -mode, you can open the table of contents via the burger menu in the bottom right corner. Use the headlines to navigate the document.

You can use emojis 😊

See full emoji list here (<https://www.webfx.com/tools/emoji-cheat-sheet/>).

We mark our tasks and take home messages like this:



This is a **task**



This is a **take home messages**

Get more information about HedgeDoc features here (<https://notes.desy.de/features?both>).

## Internet browser

For hands-on tasks, any web browser can be used.

## Session Schedule

---

**120 min**

Time	Content
10:30	Welcome & introduction
10:45	Data & metadata
11:15	<b>5 min break</b>
11:20	Structured metadata
11:40	Enabling technologies, standards
12:25	Wrap-up


## Data & Metadata

---

### TASK 1: Warm-up



Let's type a small JSON metadata record about ourselves and the cities we live in 😊.

Copy the example below, paste it to text field  **here** (<https://ahaslides.com/362YN>) and fill in your values.

#### Example:

```
{
  "firstName": "value",
  "ORCID": "value",
  "researchField": "value",
  "currentPosition": "value",
  "favoriteCake": "value",
  "hobbies": ["value", "value"]
  "city": {
    "name": "value",
    "url": "value"
  }
}
```

## Data – Information – Knowledge – Wisdom

The question “What are data?” seems trivial at first, but if we look at the definition, it is apparent that the question is not that easy to answer. In information science, we distinguish between Glyphs (or symbols), data, information, knowledge and wisdom. **GLYPHS** are the smallest unit of data representation. Glyphs represent the symbols of which data can be composed.

To cite the information scientist Jeffrey Pomerantz, “**DATA** is stuff. It is raw, unprocessed, possibly even untouched by human hands, unviewed by human eyes, un-thought-about by human minds”<sup>[1]</sup> (<https://notes.desy.de/s/EpeiR6oCS#fn1>). In other words, data is potential information, that requires processing and context to extract the information held within.

Accordingly, **INFORMATION** is processed, human-consumable data. If this information is internalized by a human being, it is called **KNOWLEDGE**. This knowledge can be applied in a broader context by the human being. Applied knowledge is called **WISDOM**.

The key to reaching wisdom from data is processing and contextualizing data to extract information. To achieve this goal we often need to add a description to the data: **metadata**.



Data is potential information and needs to be processed and contextualized to make it accessible for a human agent. Data can be understood as **potentially informative**.

# Metadata

Metadata are **(semi-)structured data** that provide information about characteristics of other (more complex) data objects (e.g. files or documents). Regarding research data, metadata gives the observer the necessary context to interpret the data and derive information from it. Although metadata is data itself, it can only exist in connection with a data object that is described by the metadata record (e.g. the meta-information in a book about said book). Metadata can be found inside of a data object (e.g. in a book, in a data record) or as a separate object (e.g. library catalogue, separate file).

## National Information Standards Organization

(<https://groups.niso.org/higherlogic/ws/public/download/17446/Understanding%20Metadata.pdf>) (NISO, 2004, from "Big Data, Little Data, No Data", Christine L. Borgman, 2015): "Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource".



Metadata is **a statement** about a potentially informative object.

## Types of metadata

**Descriptive metadata** provides information about the intellectual content of a (digital) object (e.g. title, author, date of publication, subject, description, unique identifier). <sup>[2]</sup>

(<https://notes.desy.de/s/EpeiR6oCS#fn2>)

**Administrative metadata** provides information to support the management of a resource (e.g. technical information regarding the file's creation and format, version, information about copyright, licence and intellectual property rights). <sup>[2:1]</sup> (<https://notes.desy.de/s/EpeiR6oCS#fn2>)

**Structural metadata** specifies the relationships between components of a (digital) object and between different (digital) objects (e.g. chapters in a book). <sup>[2:2]</sup>

(<https://notes.desy.de/s/EpeiR6oCS#fn2>)

## Structure & Schema

---

### Metadata records


#### Handwritten (lab) notes

Handwritten (lab) notes are still a common practice in many scientific disciplines. These notes are easy to take during data generation. The greatest disadvantage, however, is the physical separation from the data itself and the difficulty to find, store and share these information. Often, handwritten lab notes do not follow a predictable structure and, hence, are hard to interpret and sometimes even hard to read.

## Readme style text documents

Recording your metadata (additionally) in a digital readme style text document comes with one great advantage: the metadata can be associated and stored directly with the experimental data. Readme style metadata best practices include:<sup>[3]</sup>

(<https://notes.desy.de/s/EpeiR6oCS#fn3>)

- creating one readme file for each data file, whenever possible.
- naming the readme in a way that it is easily associated with the data file(s) it describes.
- writing the readme document as a plain text file for interoperability.
- formatting multiple readme files identically
- where possible follow established conventions for scientific vocabulary (i.e. from glossaries or resources such as the  IUPAC Gold book (<https://goldbook.iupac.org/>))

We strongly recommend to use  **this template** (<https://cornell.app.box.com/v/ReadmeTemplate>) for readme style metadata documents.

## Structured metadata records



Data exchange formats such as the **markup language XML** or **JSON** can be read and processed not only by humans but also by computers. Structured (meta)data is key to enable **machine-readability**.

## What is Markup?

**Markup** is not part of the natural text or content but tells something **about it...**<sup>[4]</sup>

(<https://notes.desy.de/s/EpeiR6oCS#fn4>) [5] (<https://notes.desy.de/s/EpeiR6oCS#fn5>)

**Punctuational markup** – for example – is placing periods or question marks at the end of sentences.

**Presentational markup** is mainly about style. Note, how the simple Markdown markup on the left tells HedgeDoc to display text **bold** on screen.

**Descriptive or declarative markup** declares what an element is; e. g. a member of a particular type or class like a:

## headline

---

If design rules for headlines change, the document structure remains intact and is still in line with the authors original intention.

**Referential markup** refers to entities external to the document and may be replaced by those entities during processing. The World Wide Web markup language HTML (**H**yper**T**ext **M**arkup **L**anguage) e.g. uses the anchor `<a>` tag for hypertext references (hyperlinks).

```
<a href="url">link text displayed to reader</a>
```



Rigorous markup can make text (character strings) more accessible for computer analysis.


**SGML** (**S**tandard **G**eneralized **M**arkup **L**anguage) was one of the first industry standards for electronic publishing – a meta-language for generalized, descriptive markup languages – first accepted as an ISO standard in 1986. Both, HTML (1989) and XML (1998) are based on SGML.<sup>[6]</sup> (<https://notes.desy.de/s/EpeiR6oCS#fn6>)

**HTML** (**H**yper**T**ext **M**arkup **L**anguage) is the standard markup language for web pages.

In contrast, the main purpose of **XML** (**eX**tensible **M**arkup **L**anguage) is the transfer and storage of arbitrary data on the World Wide Web. XML is software- and hardware-independent. It is considered human-readable and allows for hierarchical (tree-like) structures. Data elements are wrapped in start and end “tags” `<></>`. XML tags can be customized by the author of the document, its markup is extensible.<sup>[7]</sup>

(<https://notes.desy.de/s/EpeiR6oCS#fn7>)

```
<example>
  <title>This is the example title</title>
  <description>A simple XML example</description>
  <wordCount>1</wordCount>
</example>
```

**JSON** (**J**ava**S**cript **O**bject **N**otation) is not a markup language. It is a lightweight, human-readable, hierarchical format to store and transport data.<sup>[8]</sup> (<https://notes.desy.de/s/EpeiR6oCS#fn8>) 

JSON syntax (<https://www.json.org/json-en.html>) is inspired by JavaScript object notation.<sup>[9]</sup>

(<https://notes.desy.de/s/EpeiR6oCS#fn9>) Like XML, JSON is software- and hardware-independent.

- data elements are in key/value pairs
- keys are of data type *string* (in quotes)
- values must be of data type *string*, *number*, *boolean*, *array* or *object*
- elements are separated by commas
- curly braces hold objects
- square brackets hold arrays

- no comments supported

```
{
  "key": "value",
  "aString": "string",
  "anInteger": 5,
  "aFloat": 0.5,
  "aBoolean": true,
  "anArray": ["item1", "item2", "item3"],
  "anObject": {
    "key1": "value1",
    "key2": "value2",
    "key3": "value3"
  }
}
```

## Enabling Technology & Standards

---


### The Web is not the internet

In 1989 researchers Tim Berners-Lee and Robert Cailliau started their HyperText project called the WWW (**W**orld-**W**ide **W**eb, short Web) at the CERN research centre in Geneva, Switzerland. The Web was developed to “meet the demand for automated information-sharing between scientists in universities and institutes around the world”.<sup>[10]</sup>  
(<https://notes.desy.de/s/EpeiR6oCS#fn10>)

The WWW is a service on the application layer of the internet protocol stack TCP/IP (**T**ransmission **C**ontrol **P**rotocol/**I**nternet **P**rotocol) – invented by Vint Cerf and Robert (Bob) Elliot Kahn in the 1970s.

The World Wide Web main building blocks are:

- HTML markup language with “hyperlinks”
- HTTP (**H**yper**T**ext **T**ransfer **P**rotocol)
- URI (**U**niform **R**esource **I**dentifier)

HTML applies the ideas of HyperText and HyperMedia – terms coined by  Ted Nelson  
(<https://www.youtube.com/user/TheTedNelson/videos>) in the 1960s:


“a combination of natural language text with the computer’s capacity for interactive branching, or dynamic display ... ”

 Ted Nelson (<https://ieeexplore.ieee.org/document/1663693>)

HTTP is a simple protocol that sets communication rules for client and server software on the World Wide Web. For URI see chapter (Web) Location & Identifiers

(<https://notes.desy.de/s/EpeiR6oCS#web-location-identifiers>).

In 1992 Deutsches Elektronen-Synchrotron DESY in Hamburg connected a web server to the WWW. An even earlier adopter was the **arXiv preprint repository**. They switched from email to HTTP for manuscript dissemination in 1991.<sup>[11]</sup> (<https://notes.desy.de/s/EpeiR6oCS#fn11>)

So-called **web repositories** store and publish (scholarly) digital objects – like paper publications and research data – and their **metadata records**. This way, they aim to improve the persistent **findability and accessibility of research output**. Repositories in turn are indexed for findability in registry services like  <https://www.re3data.org/>


(<https://www.re3data.org/>) and  <https://v2.sherpa.ac.uk/openoar/> (<https://v2.sherpa.ac.uk/openoar/>).

## Metadata Schemas



**A metadata schema is a template** which exemplifies the metadata elements expected and how they should be structured.

**XML Schemas (.xsd)** are written in XML and used to describe & syntactically validate the structure of XML documents or (meta)data records.<sup>[12]</sup> (<https://notes.desy.de/s/EpeiR6oCS#fn12>)

The  **JSON Schema vocabulary** (<https://json-schema.org/>) is used to describe & syntactically validate the structure of JSON (meta)data records. We will focus on **JSON Schema** in our next hands-on task.<sup>[13]</sup> (<https://notes.desy.de/s/EpeiR6oCS#fn13>)

A simple JSON schema could look like the one below. It declares:

- JSON Schema version with `$schema`
- a list (an array) of required (i. e. mandatory) properties
- one required property (i.e. "superhero" )
- one optional property (i.e. "power" )
- data type constraints for record values (e.g. "type": "integer" )

There are also some `descriptions` added for the human reader.



```

{
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "description": "In real life you would add a meaningful description here.",
  "type": "object",
  "required": [
    "superhero"
  ],
  "properties": {
    "superhero": {
      "description": "A mandatory string property.",
      "type": "string"
    },
    "power": {
      "description": "An optional numeric property.",
      "type": "integer"
    }
  }
}

```

A JSON instance is syntactically valid, if it conforms to the definition described by the JSON schema. Note, that the JSON Schema `required` keyword holds a list of keys that must be present for a JSON object to be considered a valid instance of this schema.

```

{
  "superhero": "I am just a string"
}

```




The most challenging part of schema development can be to **have everyone agree on the same expectations.**

## Metadata Standards



A well **established metadata schema** can become a standard.

Researchers, librarians and web technologists drafted the **Dublin Core** – a set of 15 library-card-catalog-like metadata elements for the web – in 1995 at a meeting in Dublin, Ohio (USA).<sup>[14]</sup> (<https://notes.desy.de/s/EpeiR6oCS#fn14>)

Dublin Core and its extensions are widely used and referenced today. The  Dublin Core Metadata Initiative (DCMI) (<https://www.dublincore.org/about/>) states to work openly, with a paid-membership model.

The 15 generic Dublin Core metadata elements have been formally standardized for cross-domain resource description in e.g. **ISO 15836-1:2017**<sup>[15]</sup> (<https://notes.desy.de/s/EpeiR6oCS#fn15>)

**Creator**

**Contributor**

**Publisher**

**Title**

**Date**

**Language**

**Format**

**Subject**

**Description**

**Identifier**

**Relation**

**Source**

**Type**

**Coverage**

**Rights**

Many scholarly repositories expose a standardized **application programming interface (API)** for the harvesting of Dublin Core metadata as specified in [☞](#)

<http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm#dublincore>

(<http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm#dublincore>). **Try it yourself and check**

**oai\_dc XML records from [☞](#) Zenodo OAI-PMH endpoint** ([https://zenodo.org/oai2d?](https://zenodo.org/oai2d?verb=ListRecords&metadataPrefix=oai_dc)

**verb=ListRecords&metadataPrefix=oai\_dc**).

**TASK 2: Domain specific metadata terminologies & standards**



1. Open one of these metadata standard registries in your preferred browser:
  - 🔗 FAIRsharing.org (<https://fairsharing.org/search?fairsharingRegistry=Standard>)
  - 🔗 RDA Metadata Directory (<http://rd-alliance.github.io/metadata-directory/subjects/>)
  - 🔗 RDA Metadata Standards Catalog (<https://rdamsc.bath.ac.uk/>)
  - 🔗 RDA Metadata Directory (<http://rd-alliance.github.io/metadata-directory/subjects/>)
  - 🔗 DCC List of Metadata Standards (<https://www.dcc.ac.uk/guidance/standards/metadata/list>)
2. Search for a metadata schema, standard or vocabulary **relevant to your research domain**.
3. Inspect the **information provided**.
4. Take notes to **discuss your findings** with the group. Did you get any 404 (not found) responses clicking on links? Do you want to try a Google search in addition? Share some of your notes below. Consider discussing your findings with your colleagues.

#### Notes:

```
your research domain: link to metadata schema, standard or vocabulary found
your research domain: link to metadata schema, standard or vocabulary found
...
```



You did not find a standard, schema or recommendation for your research domain or **want to participate in metadata development? Get in touch with us at HMC.**

## (Web) Location & Identifiers

### The Web of today

Built for human understanding: even though web documents are made with computers, computers can NOT understand the content of these documents. **They can't read, see relationships or make decisions like human can.**

Most search engines are based on keywords:

- results are high recall and low precision
- results are highly sensitive to vocabulary
- results are single human-readable web pages
- **results do not support logical reasoning and query answering**

The World Wide Web is a hypermedia system. It contains:

### Resources

A web resource is any identifiable resource present on or connected to the World Wide Web. A resource can be anything that has identity.

### Links

(Web identifiers) between these resources.

## Web Identifiers

The **Uniform Resource Identifier (URI)** is a string of characters formulated to uniquely identify a resource, (most commonly on the Web) and enable interaction with it via common protocols such as HTTP.

A **Uniform Resource Name (URN)** is a type of URI. It is a standard, persistent and unique identifier for digital resources on the Internet. To link to the resource from the URN, a **resolver service is required.**

The **Uniform Resource Locator (URL)** is a string of characters used to access the information or resource by **using the address of the resource location** via communication protocols such as HTTP.



A URL does not ensure the link to a resource is maintained if it is moved within its repository. **URL links “rot” over time.**

A **Persistent Identifier (PID)** is a long-lasting reference to digital objects:


- articles
- datasets
- tables
- figures
- videos
- persons
- instruments
- organizations

**PIDs**<sup>[16]</sup> (<https://notes.desy.de/s/EpeiR6oCS#fn16>)

- are globally unique
- persistent over time
- ensure permanent identifiability, referencing and linking of digital objects

- **make researchers, their affiliations and their contributions more easily discoverable**
- are a key element in FAIR (<https://www.nature.com/articles/sdata201618>)

### Frequently used PID schemes

- Digital Object Identifiers (DOIs)
- Persistent Uniform Resource Locators (PURLs)
- International Standard Book Number (ISBN)
-  ORCIDs (<https://orcid.org/0000-0002-1825-0097>)
- Archival Resource Keys (ARKs)
- International Standard Name Identifier (ISNI)

## The Semantic Web

**The term “Semantic Web” refers to W3C’s** (<https://www.w3.org/standards/semanticweb/>) **vision of a Web of Linked Data.**

It provides a way for machines to be able to process and understand the data that they were only to display on the traditional Web. It is a vision for the future Web (a web of meaning — semantics); originally defined by Tim Berners-Lee. The Semantic Web is not a separate Web, but an extension of the current one. In the Semantic Web, metadata are invisible as people read the page, but they’re clearly visible to computers.



The goal of a Semantic Web is to **make computers perform more of the tedious work involved in finding, sharing and combining information on the Web efficiently.**

## Contact Information

---

### Helmholtz Metadata Collaboration (HMC)

#### Hub Information

Institute for Advanced Simulation - Materials Data Science and Informatics (IAS-9)  
Forschungszentrum Jülich  
Dennewartstraße 25  
52068 Aachen

[HMC@fz-juelich.de](mailto:HMC@fz-juelich.de) (<mailto:HMC@fz-juelich.de>)

[www.helmholtz-metadaten.de](http://www.helmholtz-metadaten.de) (<http://www.helmholtz-metadaten.de>)

 [@helmholtz\\_hmc](#)

## Instructors

Annika Strupp

☎ +49 241 927803-49

✉ [a.strupp@fz-juelich.de](mailto:a.strupp@fz-juelich.de) (<mailto:a.strupp@fz-juelich.de>)

🆔 0000-0002-0070-4337 (<https://orcid.org/0000-0002-0070-4337>)

Dr. Silke Christine Gerlich










☎ +49 241 927803-46




✉ [s.gerlich@fz-juelich.de](mailto:s.gerlich@fz-juelich.de) (<mailto:s.gerlich@fz-juelich.de>)

🆔 0000-0003-3043-5657 (<https://orcid.org/0000-0003-3043-5657>)

## Further Reading

---

1. Pomerantz, J. (2015). Metadata. Cambridge, MA: MIT Press.   
(<https://notes.desy.de/s/EpeiR6oCS#fnref1>)
2. Zhang, A. B., Gourley, D. (2008). "Metadata strategy" in Creating Digital Collections. Sawston, UK: Woodhead Publishing.  (<https://notes.desy.de/s/EpeiR6oCS#fnref2>)   
(<https://notes.desy.de/s/EpeiR6oCS#fnref2:1>)  (<https://notes.desy.de/s/EpeiR6oCS#fnref2:2>)
3. Chadwick, I. (2016). "Research Data Management: guide to writing "readme" type metadata." The Open University. <https://www.open.ac.uk/library-research-support/sites/www.open.ac.uk.library-research-support/files/files/RDM-Guidelines-for-creating-readme-style-metadata.pdf> (<https://www.open.ac.uk/library-research-support/sites/www.open.ac.uk.library-research-support/files/files/RDM-Guidelines-for-creating-readme-style-metadata.pdf>)  (<https://notes.desy.de/s/EpeiR6oCS#fnref3>)
4. James H. Coombs et al. (November 1987). Markup Systems and the Future of Scholarly Text Processing. Communications of the ACM 30.  
<http://xml.coverpages.org/coombs.html#Note1> (<http://xml.coverpages.org/coombs.html#Note1>)   
(<https://notes.desy.de/s/EpeiR6oCS#fnref4>)
5. Cynthia Zender (2005). Markup 101: Markup Basics. SAS Institute.  
<https://www.lexjansen.com/pharmasug/2005/Tutorials/tu12.pdf>  
(<https://www.lexjansen.com/pharmasug/2005/Tutorials/tu12.pdf>)  (<https://notes.desy.de/s/EpeiR6oCS#fnref5>)
6. <https://www.iso.org/standard/16387.html> (<https://www.iso.org/standard/16387.html>)   
(<https://notes.desy.de/s/EpeiR6oCS#fnref6>)
7. "XML Tutorial". © 1999-2022. Refsnes Data, W3Schools.  
<https://www.w3schools.com/xml/> (<https://www.w3schools.com/xml/>)   
(<https://notes.desy.de/s/EpeiR6oCS#fnref7>)

8. <https://www.ecma-international.org/publications-and-standards/standards/ecma-404/>  
(<https://www.ecma-international.org/publications-and-standards/standards/ecma-404/>)   
(<https://notes.desy.de/s/EpeiR6oCS#fnref8>)
9. "JSON Introduction". © 1999-2022. Refsnes Data, W3Schools.  
[https://www.w3schools.com/js/js\\_json\\_intro.asp](https://www.w3schools.com/js/js_json_intro.asp) ([https://www.w3schools.com/js/js\\_json\\_intro.asp](https://www.w3schools.com/js/js_json_intro.asp))  
 (<https://notes.desy.de/s/EpeiR6oCS#fnref9>)
10. <https://home.cern/science/computing/birth-web> (<https://home.cern/science/computing/birth-web>)  
 (<https://notes.desy.de/s/EpeiR6oCS#fnref10>)
11. <https://ar5iv.labs.arxiv.org/html/1709.07020> (<https://ar5iv.labs.arxiv.org/html/1709.07020>)   
(<https://notes.desy.de/s/EpeiR6oCS#fnref11>)
12. "XML Schema Tutorial". © 1999-2022. Refsnes Data, W3Schools. [https://www.w3schools.com/xml/schema\\_intro.asp](https://www.w3schools.com/xml/schema_intro.asp) ([https://www.w3schools.com/xml/schema\\_intro.asp](https://www.w3schools.com/xml/schema_intro.asp))   
(<https://notes.desy.de/s/EpeiR6oCS#fnref12>)
13. "Understanding JSON Schema. The basics", © Copyright 2013-2016 Michael Droettboom, Space Telescope Science Institute; Last updated on Feb 07, 2022.  
<https://json-schema.org/understanding-json-schema/basics.html> (<https://json-schema.org/understanding-json-schema/basics.html>)  (<https://notes.desy.de/s/EpeiR6oCS#fnref13>)
14. <https://www.dublincore.org/resources/metadata-basics/>  
(<https://www.dublincore.org/resources/metadata-basics/>)  (<https://notes.desy.de/s/EpeiR6oCS#fnref14>)
15. <https://www.iso.org/standard/71339.html> (<https://www.iso.org/standard/71339.html>)   
(<https://notes.desy.de/s/EpeiR6oCS#fnref15>)
16. John Kunze (2018). "Ten persistent myths about persistent identifiers".  
<https://escholarship.org/uc/item/73m910w8> (<https://escholarship.org/uc/item/73m910w8>)   
(<https://notes.desy.de/s/EpeiR6oCS#fnref16>)