

Metadata

Why we use it all the time but seldomly think about it.

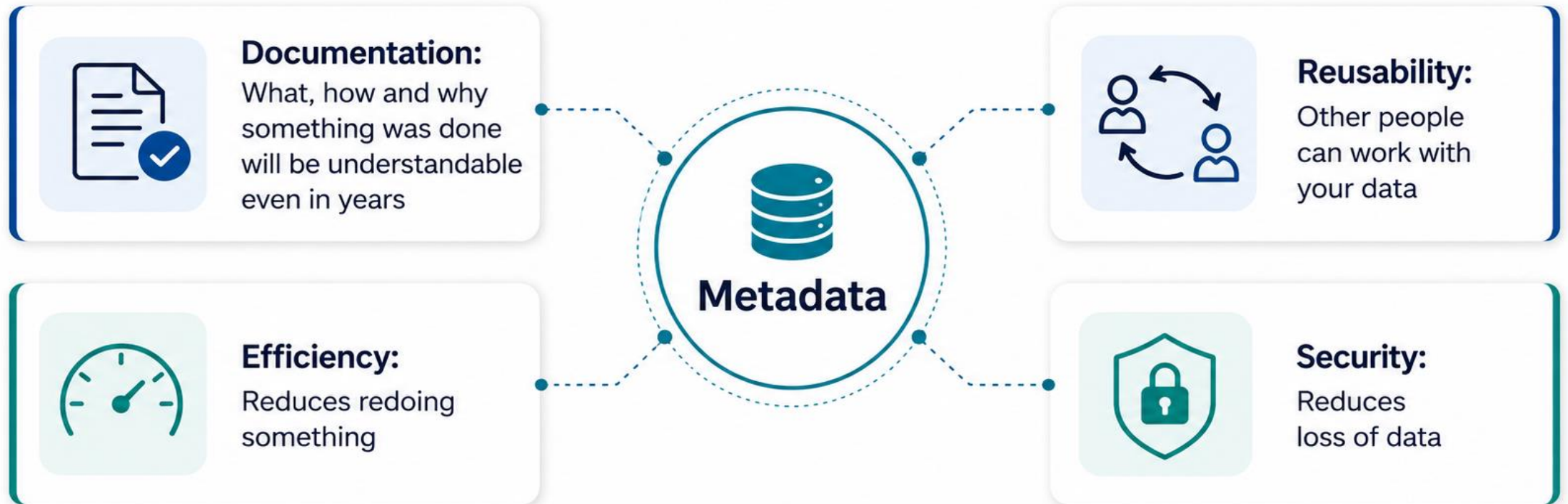


LEAVE WITH THESE HABITS

Five habits worth keeping.

- 01 Context before interpretation.
- 02 Make the implicit explicit.
- 03 Document decisions while they are still obvious.
- 04 Use shared language instead of personal shorthand.
- 05 Describe data so it can outlive its creator and its project.

Why is it relevant?



What is it?
What's in it?



Descriptive

Information about intellectual content

Administrative

Information to support resource management

Structural

Relationships between parts of a digital object

Technical

Information about the creation



Zutaten: Erbsen, Wasser, Speisesalz.

Durchschnittliche Nährwerte (abgetropft)	pro 100 g	%RM* pro 100 g
Brennwert	300 kJ / 72 kcal	4 %
Fett	0,5 g	< 1 %
– davon gesättigte Fettsäuren	0,1 g	< 1 %
Kohlenhydrate	8,0 g	3 %
– davon Zucker	2,0 g	2 %
Ballaststoffe	6,5 g	–
Eiweiß	5,5 g	11 %
Salz	0,25 g	4 %

* Referenzmenge für einen durchschnittlichen Erwachsenen (8400 kJ/2000 kcal)

Hergestellt für
Netto Marken-Discount Stiftung & Co. KG,
Industriepark Ponholz 1,
93142 Maxhütte-Haidhof

222022 (K3)

Beste Ernte

JUNGE ERBSEN

extra fein

Erbsen, extra fein

Füllmenge: **200 g e** **212**

Abtropfgewicht: **140 g**

Dose: 40 FE Etikett: 22 PAP

Trocken lagern.

Mindestens haltbar bis:
siehe Boden

Descriptive metadata

Administrative metadata

Technical metadata

Structural metadata

1. Ingredients

Zutaten: Erbsen, Wasser, Speisesalz.

2. Nutritional content

Nutrition table with values per 100 g.

3. Brand and product identity

"Beste Ernte" and "JUNGE ERBSEN".

4. Product descriptor

"extra fein" and "Erbsen, extra fein".

5. Quantity information

Füllmenge: 200 g
Abtropfgewicht: 140 g.

6. Nutri-Score label

Nutri-Score rating front-of-pack label.

1. Producer / responsible organisation information

Producer name and address.

2. Internal / lot-like code

Internal code or batch/lot identifier.

Zutaten: Erbsen, Wasser, Speisesalz.

Durchschnittliche Nährwerte (abgetropft)	pro 100 g	%RM* pro 100 g
Brennwert	300 kJ/72 kcal	4 %
Fett	0,5 g	< 1 %
– davon gesättigte Fettsäuren	0,1 g	< 1 %
Kohlenhydrate	8,0 g	3 %
– davon Zucker	2,0 g	2 %
Ballaststoffe	6,5 g	–
Eiweiß	5,5 g	11 %
Salz	0,25 g	4 %

* Referenzmenge für einen durchschnittlichen Erwachsenen (8400 kJ/2000 kcal)

Hergestellt für
Netto Marken-Discount Stiftung & Co. KG,
Industriepark Ponholz 1,
93142 Maxhütte-Haidhof

222022 (K3)

Serviervorschlag



Beste Ernte

JUNGE ERBSEN
extra fein

Erbsen, extra fein

Füllmenge:
200 g e

Abtropfgewicht:
140 g

Dose: 40 FE

Etikett: 22 PAP

Trocken lagern.

Mindestens haltbar bis:
siehe Boden

212



1. Numeric can-marking code

Can-marking code "212".

2. Packaging material / recycling codes

Material and recycling codes for can and label.

3. Barcode / EAN code

EAN-13 barcode for retail scanning.

1. Information panels (label structure)

The label is organised into three information panels: left, centre, right.
This structure groups related content and guides reading.

2. Nutrition table structure

The nutrition table is organised into columns and rows with headings (e.g., "pro 100 g", "%RM* pro 100 g") and values for each nutrient.

3. Section headings

Headings (e.g., "Zutaten", "Füllmenge", "Dose", "Etikett") and visual grouping organise information into meaningful sections.

3. Storage instruction

Instruction for proper storage.

4. Best-before information

Best-before date reference.

Metadata: data about data

Five lenses for describing a dataset so others can understand and reuse it.

Aa

Descriptive

Information about intellectual content

- Title
- Description
- Authors and contact
- Keywords



Technical

Information about the creation

- Used tools
- Coordinates or setup
- Settings
- Units and sampling rate

Metadata

context
for trust

ID

Administrative

Information to support resource management

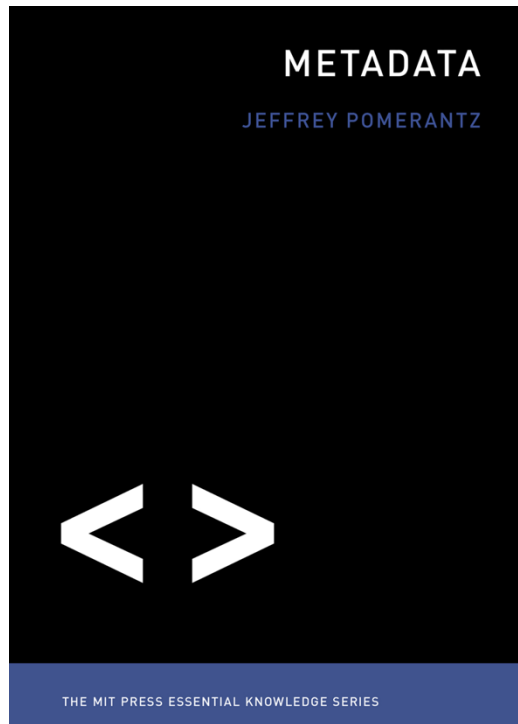
- File format
- Version
- Creator / responsibility
- Date



Structural

Relationships between parts of a digital object

- Links between files
- Raw → checked → analysis
- Notebook, plots, README



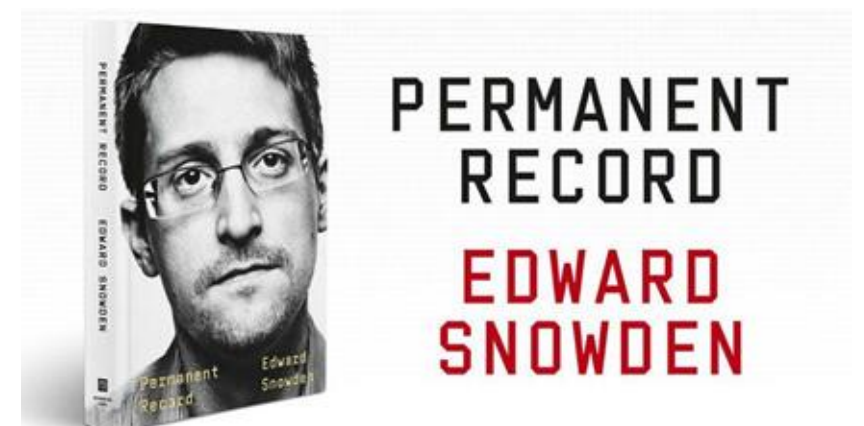
2015

Metadata is a map.
Metadata is a means by
which the complexity of
an object is represented
in a simpler form.

<https://direct.mit.edu/books/book/3111/Metadata>

Most people, even today, tend to think of mass surveillance in terms of content—the actual words they use on a phone call or an email. The unfortunate truth is that the unwritten, unspoken information **reveals the additional context** and our patterns of behavior.

The NSA calls this information **metadata**. The prefix *meta* usually means *above* or *beyond*, but here it means *about*. In this sense metadata are *data about data*—or, more precisely, *data generated by data*: a **bundle of markers and clues that makes data truly usable**.



<https://www.romania-insider.com/book-review-permanent-record-edward-snowden>

Use metadata: what actions leave behind

Downloads, calls, clicks, views, edits, purchases, access logs.

Use metadata can describe behavior and relationships — often without revealing content.

Who?

user / device / account

When?

timestamp / duration

Where?

location / IP / service

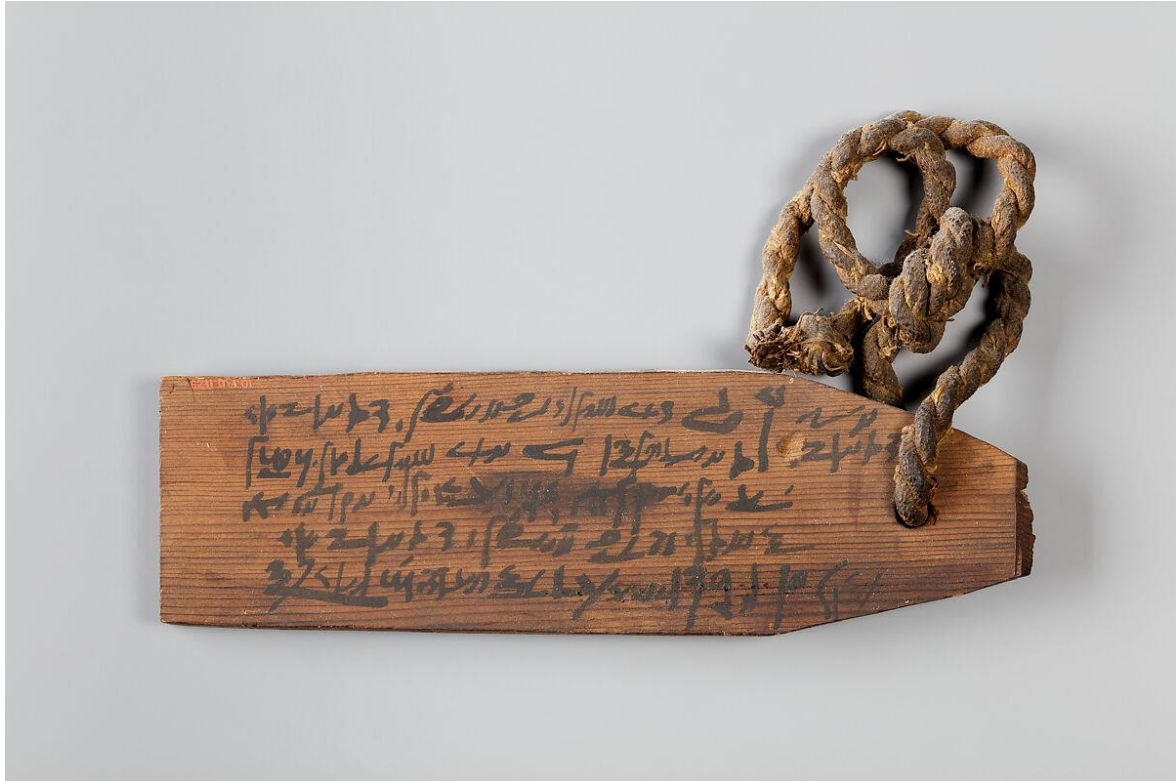
What?

resource used / action

With whom?

network relation

**Ethics note: the same metadata that makes systems useful can also make people legible.
Treat personal metadata as sensitive.**



<https://www.metmuseum.org/art/collection/search/557908>

Mummy label of Tsenpetese daughter of Panahib

Roman Period

A.D. 3rd century ?

Petempto the Younger

Age: **79 years**

Father: Petemin

Mother: Kolluthe

A typical Rhodian amphora with stamps on the upper parts of the handles, giving the maker's name and the date of manufacture.



Terracotta transport amphora



Greek, Rhodian



late 3rd–mid-2nd century
BCE



On view at The Met Fifth
Avenue in Gallery 171



Stamped inscriptions indicate the month
(Panamos) the amphora was made, the name of
the magistrate (Pythogenes) in office at the time,
and the name of the amphora's maker
(Hephaistion).

Left handle: **ἐ[πι Πυ]θογενεὺς Δευτέρου
Πανάμου**
Right handle: **Ἡφαιστίωνος**



AI-assisted synthesized

Standards reduce ambiguity

Everyone agrees standards are good — the hard part is using the same ones.

Schema

Defines what statements are allowed: required fields, element names, structure, and sometimes rules.

Syntax encoding

Defines how values are written: dates, identifiers, file formats, language codes.

Controlled vocabulary

Defines which terms are permitted: subjects, names, categories, places.

Without shared rules, metadata becomes private shorthand. With shared rules, it becomes interoperable.

HOW STANDARDS PROLIFERATE: (SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)



Paper full of layers



EXAMPLE HEALTH CENTRE
123 Wellness Way, Healthtown 2000
Tel: (02) 5555 1234 | www.examplehealth.edu


Reg No: EHC12345

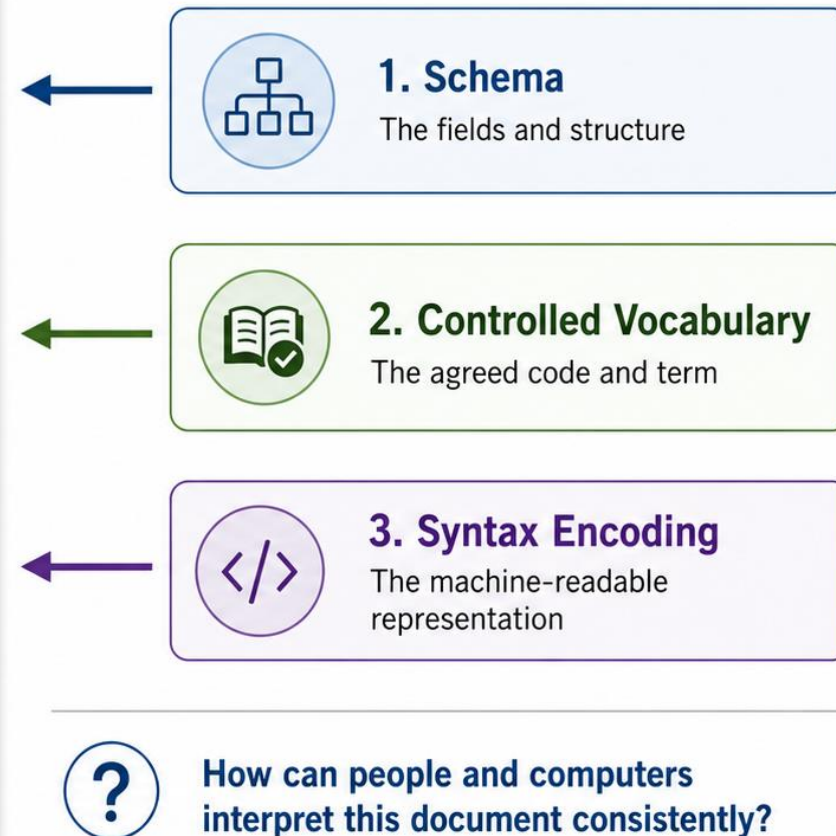
Certificate of Incapacity for Work

Patient:	Alex Morgan
Certificate ID:	MC-2026-0042
Unable to work:	20 July 2026 – 22 July 2026
Diagnosis:	Acute nasopharyngitis
Common term:	Common cold
ICD-11 code:	CA00
Diagnosis status:	Confirmed
Issued by:	Dr Maya Chen
Organisation:	Example Health Centre
Date issued:	20 July 2026

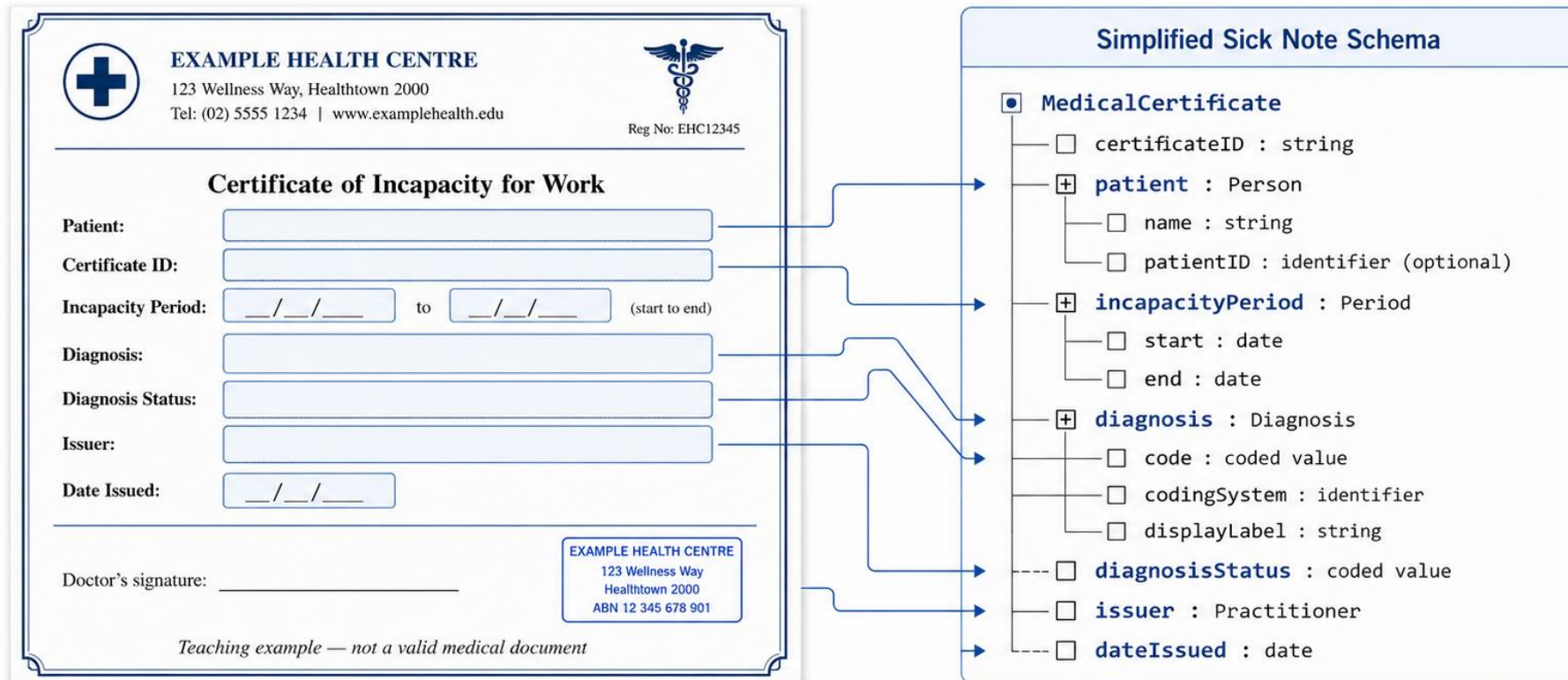
Doctor's signature: 

EXAMPLE HEALTH CENTRE
123 Wellness Way
Healthtown 2000
ABN 12 345 678 901

Teaching example — not a valid medical document



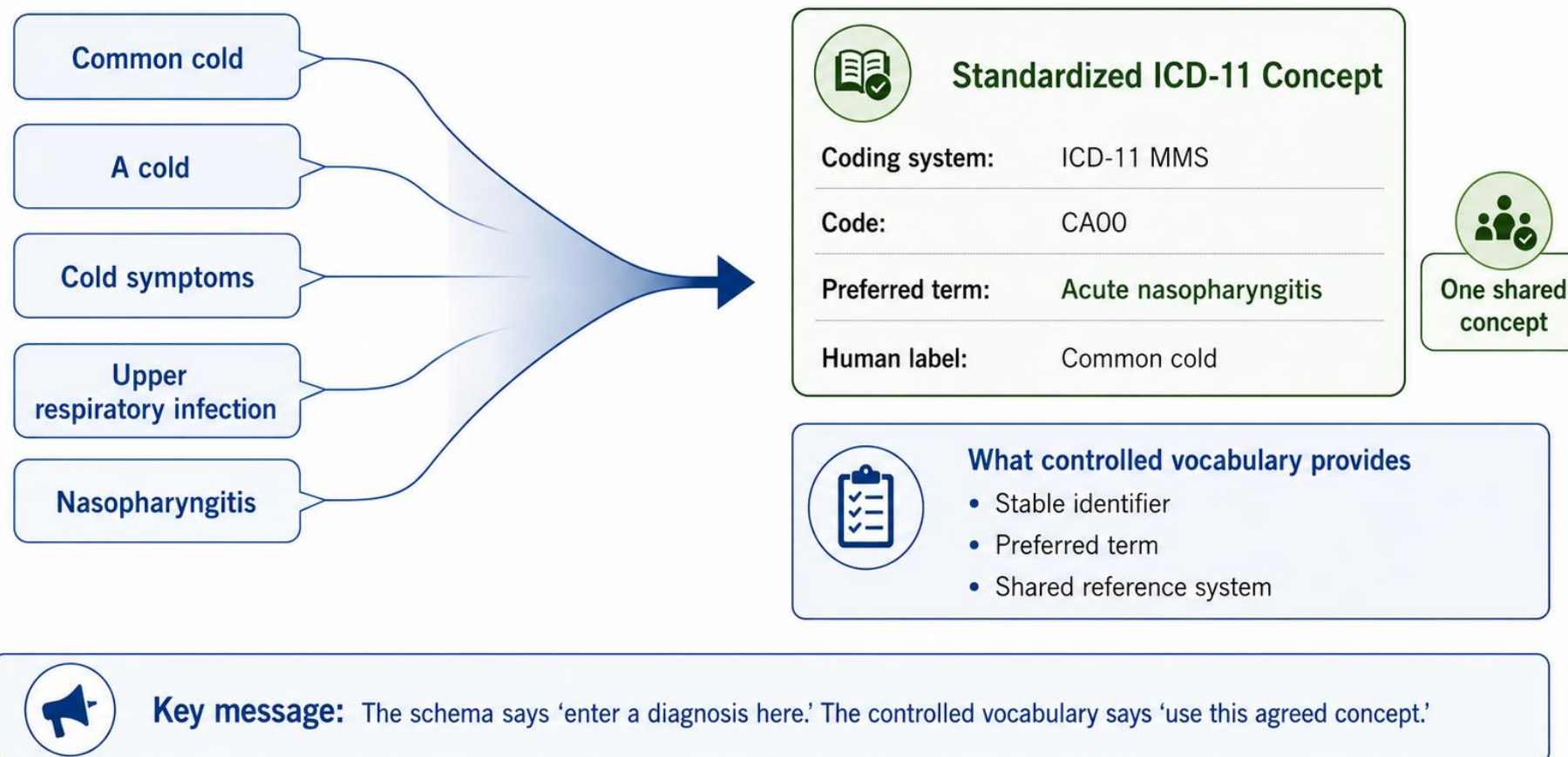
Layer 1: Schema



Key message: The schema defines where information belongs — not which diagnosis should be entered.

Layer 2: Controlled Vocabulary

International Classification of Diseases, 11th Revision — Mortality and Morbidity Statistics



Layer 3: Syntax Encoding

**EXAMPLE HEALTH CENTRE**
123 Wellness Way, Healthtown 2000
Tel: (02) 5555 1234 | www.examplehealth.edu


Reg No: EHC12345

Sick Leave Certificate (Excerpt)

Certificate ID:	MC-2026-0042
Patient:	Alex Morgan
Start:	2026-07-20
End:	2026-07-22
Diagnosis:	Acute nasopharyngitis
Coding system:	ICD-11 MMS
Code:	CA00
Diagnosis status:	confirmed
Issuer:	Dr Maya Chen
Organisation:	Example Health Centre
Date issued:	2026-07-20



Human-readable document

```
1 {
2   "certificateID": "MC-2026-0042",
3   "patient": {
4     "name": "Alex Morgan"
5   },
6   "incapacityPeriod": {
7     "start": "2026-07-20",
8     "end": "2026-07-22"
9   },
10  "diagnosis": {
11    "codingSystem": "ICD-11 MMS",
12    "code": "CA00",
13    "displayLabel": "Acute nasopharyngitis"
14  },
15  "diagnosisStatus": "confirmed",
16  "issuer": {
17    "name": "Dr Maya Chen",
18    "organisation": "Example Health Centre"
19  }
20  "dateIssued": "2026-07-20"
21 }
```



Machine-readable JSON

Alternative encodings

 XML

 CSV

 RDF



Key message: Syntax makes the information machine-readable. It does not define the medical meaning by itself.



1. REAL-WORLD INFORMATION

Alex cannot work because of a common cold



2. SCHEMA

A diagnosis must contain code + coding system + display label



3. CONTROLLED VOCABULARY

ICD-11 MMS:
CA00 = Acute nasopharyngitis



4. SYNTAX ENCODING

```
"diagnosis": {  
  "codingSystem": "ICD-11 MMS",  
  "code": "CA00",  
  "displayLabel":  
    "Acute nasopharyngitis"  
}
```



Schema

What information can be recorded, and where?



Controlled Vocabulary

Which agreed concepts and values can be used?



Syntax Encoding

How is the information written and exchanged?



Schema provides structure. Controlled vocabulary provides shared meaning. Syntax encoding provides machine-readable form.

Where metadata lives

Embedded metadata is authoritative; external metadata is flexible.

Embedded / internal

Examples: HTML meta tags, EXIF inside an image file, metadata on a copyright page.

Strength: travels with the object.

Risk: hard to update.

External / separate

Examples: library catalog record, repository landing page, API record, database row.

Strength: searchable and easy to revise.

Risk: authority must be clear.

The practical rule: make the relationship between resource and record explicit.

Metadata

Now, let's get it working.



Recap

01

WHY SHOULD WE
HAVE METADATA?

02

WHAT SHOULD BE
DESCRIBED WITH
METADATA?

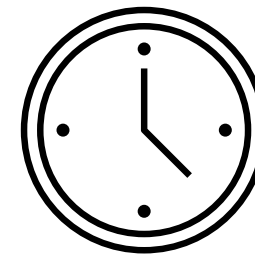
03

HOW SHOULD WE
DOCUMENT
METADATA?

Task

- Make group of 2-3 people
- Look at the dataset
<https://rwth-aachen.sciebo.de/s/WFi9XSJxWBr6rY6>
- Write down all your questions. What do you miss to understand the dataset?

```
experiment/  
├── analysis.ipynb  
├── analysis_final.ipynb  
├── data.csv  
├── data_new.csv  
├── final.csv  
├── notes.txt  
├── plot.png  
├── plot_final2.png  
├── temp/  
│   └── test.csv  
└── old/  
    └── backup.csv
```



6 min

Schema „EngMeta“



The closer the schema is to your dataset the more you make it reusable.

- Esp. for Engineering Sciences:
<https://rdamsc.bath.ac.uk/msc/m100>



Need other schema?
<https://rdamsc.bath.ac.uk/>

1 • First look at the technical schema

EngMeta is not only a list of questions: it is an XML Schema Definition (XSD) with element names, order and cardinalities.

EngMeta_v1.1.xsd • shortened excerpt

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
  xmlns:dtc="http://datacite.org/schema/kernel-4"
  xmlns:pm="http://www.loc.gov/premis/v3"
  xmlns:ex="urn:exptml:schema:draft:0.5">
  <xs:import namespace="http://datacite.org/schema/kernel-4" .../>
  <xs:import namespace="http://www.loc.gov/premis/v3" .../>
  <xs:import namespace="urn:exptml:schema:draft:0.5" .../>

  <xs:element name="dataset">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="contact" minOccurs="1" maxOccurs="unbounded"/>
        <xs:element name="creator" minOccurs="1" maxOccurs="unbounded"/>
        <xs:element name="title" minOccurs="1" maxOccurs="unbounded"/>
        <xs:element name="dates" minOccurs="0"/>
        <xs:element name="version" minOccurs="1" maxOccurs="1"/>
        <xs:element name="mode" minOccurs="0" maxOccurs="3"/>
        <xs:element name="measuredVariable" type="variable" maxOccurs="unbounded"/>
        <xs:element name="system" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element name="provenance" minOccurs="0" maxOccurs="1"/>
        <xs:element name="size" type="xs:long" minOccurs="1"/>
        <xs:element name="storage" type="pm:storageComplexType" minOccurs="1"/>
        <xs:element name="format" type="pm:formatComplexType" minOccurs="1"/>
        <xs:element name="checksum" type="checksumType" minOccurs="1"/>
        <xs:element name="rightsStatement" type="pm:rightsStatementComplexType"
minOccurs="1"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

dataset

description
contact • creator • title

event time
dates/date

engineering
mode • measuredVariable

observed system
system/temporalResolution

provenance
provenance/step/tool

preservation
size • format • checksum • rights

For 10 minutes we do not validate a full XML file.
We actively use this schema slice to decide
where evidence belongs.

2 • Your working XML fragment

Fill selected EngMeta elements with evidence from the spreadsheet. Mark each inserted value: E explicit • D derived • I inferred • M missing.

starter XML fragment for students

```
<dataset>
  <title>____</title> <!-- tag: ____ -->
  <resourceType resourceTypeGeneral="____">____</resourceType>
  <dates>
    <date dateType="Collected">____ / ____</date> <!-- tag: ____ -->
  </dates>
  <version>____</version>
  <mode>____</mode> <!-- tag: ____ -->

  <measuredVariable>
    <name>____</name><symbol>____</symbol><encoding>____</encoding>
  </measuredVariable>

  <system id="S1">
    <temporalResolution>
      <numberOfTimesteps>____</numberOfTimesteps>
      <interval>____</interval><unit>____</unit>
    </temporalResolution>
  </system>

  <provenance>
    <step order="1">
      <type>generation</type>
      <tool><name>____</name><softwareVersion>____</softwareVersion></tool>
    </step>
  </provenance>

  <instrument>
    <name>____</name><description>____</description></instrument>
  </step>
</dataset>

<rightsStatement>____</rightsStatement>
</dataset>
```

The schema is used actively when you decide:

Where does the date go?

dates/date
@dateType="Collected"

Is “Absolute” measured?

No: derived from x/y/z →
provenance note

Where does sampling rate
go?

system/temporalResolution

Where does app/device info
go?

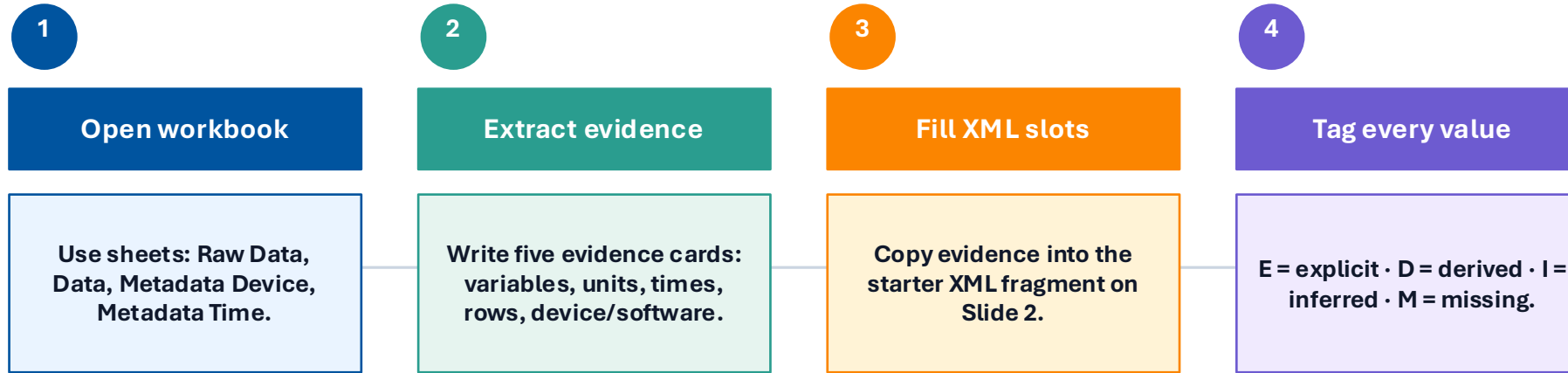
provenance/step/tool +
instrument

What cannot be filled?

creator, system/object,
orientation, license

**Do not “pretty describe” the file.
Place evidence into XML elements.**

3 · Ten-minute step-by-step instruction



Evidence you should find			
variables	Gyroscope x/y/z	unit	rad/s
timesteps	25,207	device/software	Apple · iPhone17,1 · app 1.2.0
date range	START → PAUSE		

Template: <https://pad.tools.coscine.dev/IA2026>

Recommended time: 10 minutes student work + 3 minutes comparison. No full validation in this activity.

4 · Student worksheet: map workbook evidence to EngMeta XML paths

Students complete the middle column and the tag column. The XML path column is fixed by the schema.

EngMeta XML path	Value from the spreadsheet	E/D/I/M	Evidence hint
dataset/title		—	filename + time
dataset/resourceType @resourceTypeGeneral		—	schema choice
dataset/dates/date @dateType="Collected"		—	Metadata Time
dataset/mode		—	schema enum
dataset/measuredVariable /name symbol encoding		—	Raw Data
dataset/system /temporalResolution /numberOfTimesteps		—	row count
dataset/system /temporalResolution /interval unit		—	Data sheet
dataset/provenance /step/tool		—	Device sheet
dataset/provenance /step/instrument		—	Device sheet
dataset/rightsStatement		—	not present?

Tag key

E

explicit

D

derived

I

inferred

M

missing

5 • Solution slide: compact EngMeta mapping

A good solution separates facts, calculations, interpretations and gaps instead of pretending everything is known.

EngMeta XML path	Suggested value	Tag	Why
dataset/title	Gyroscope rotation-rate measurement, 2026-07-28	I	from filename + times
dataset/resourceType @resourceTypeGeneral	Dataset	I	schema category
dataset/dates/date @dateType="Collected"	2026-07-28T14:22:58.699+02:00 / 2026-07-28T14:27:10.709+02:00	E	Metadata Time
dataset/mode	Experiment	I	schema enum
dataset/measuredVariable	angular velocity x/y/z; symbols wx, wy, wz; encoding rad/s	E	column names + units
dataset/system /temporalResolution /numberOfTimesteps	25,207	D	count rows
dataset/system /temporalResolution /interval unit	≈ 0.009997 s	D	time delta / median
dataset/provenance /step/tool	export app/software: version 1.2.0; file format 1.19	E/I	device sheet; app name may need verification
dataset/provenance /step/instrument	smartphone gyroscope; Apple iPhone 17,1	E/I	device explicit, sensor type inferred
dataset/rightsStatement	missing: no license / rights statement in workbook	M	not present

Debrief prompts

“Absolute (rad/s)” is derived, not a fourth sensor channel.

The observed system is unknown: what exactly moved?

Schema use reveals gaps, not just values.

Good final student questions: How was the phone mounted/oriented? What was rotated? Was the sensor calibrated? What license applies?

5 • Solution slide: compact EngMeta mapping

```
<dataset>
  <title>Gyroscope rotation-rate measurement – first 100
observations</title>
  <!-- tag: DERIVED -->

  <resourceType resourceTypeGeneral="Dataset">
    Gyroscope time-series subset
  </resourceType>

  <dates>
    <date dateType="Collected">
      2026-07-28T14:22:58.699+02:00 / 2026-07-28T14:22:59.692+02:00
    </date>
    <!-- tag: EXPLICIT + DERIVED -->
  </dates>

  <version>1.0</version>

  <mode>Experiment</mode>
  <!-- tag: INFERRED -->

  <measuredVariable>
    <name>Angular velocity around the x axis</name>
    <symbol> $\omega_x$ </symbol>
    <encoding>rad/s</encoding>
  </measuredVariable>

  <system id="S1">
    <temporalResolution>
      <numberOfTimesteps>100</numberOfTimesteps>
      <interval>0.009997</interval>
      <unit>s</unit>
    </temporalResolution>
  </system>
```

```
<provenance>
  <step order="1">
    <type>generation</type>

    <tool>
      <name>phyphox</name>
      <softwareVersion>1.2.0</softwareVersion>
    </tool>

    <instrument>
      <name>Embedded three-axis gyroscope</name>
      <description>
        Gyroscope integrated into an Apple device with model
        identifier iPhone17,1. The exact sensor model, calibration,
        mounting and orientation are not documented.
      </description>
    </instrument>
  </step>
</provenance>

  <rightsStatement>
    Unknown – no rights or licence information is provided in the source
    file.
  </rightsStatement>
</dataset>
```

Task



Write a README for your dataset of the experiment in JupyterHub.



Think about not only the content but also *how* you write the content.



README-files can be done for each folder (if necessary).



Need some inspiration?
<https://readme.so/>

<> Code

1,231 Commits

7 months ago

9 months ago

11 months ago

5 years ago

3 years ago

5 years ago

3 years ago

5 years ago

11 months ago

About

The Citation File Format lets you provide citation metadata for software or datasets in plaintext files that are easy to read by both humans and machines.

[citation-file-format.github.io](#)

attribution

format

citation

credit

research-software-engineering

citation-files

software-sustainability

wssspe

Readme

CC-BY-4.0 license

Code of conduct

Contributing

Cite this repository

Cite this repository

If you use this software in your work, please cite it using the following metadata. [Learn more about CITATION files.](#)

APA

BibTeX

Druskat, S., Spaaks, J. H., Chue Hong, N., I

View citation file

+ 2 releases

Deployments 71

github-pages 8 years ago

One level up: Metadata about the project

- Using a CFF-file is an additional way of writing metadata

<https://citation-file-format.github.io/>

```
cff-version: 1.2.0
message: "If you use CFF in your research, please cite it using these metadata."
abstract: CITATION.cff files are plain text files with human- and machine-readable
citation information for software. Code developers can include them in their
repositories to let others know how to correctly cite their software. This is the
specification for the Citation File Format.
authors:
  - family-names: Druskat
    given-names: Stephan
    orcid: https://orcid.org/0000-0003-4925-7248
title: Citation File Format
version: 1.2.0
identifiers:
  - type: doi
    value: 10.5281/zenodo.5171937
    description: The versioned DOI for the version 1.2.0 of the Citation File
Format.
date-released: "2021-08-09"
keywords:
  - citation file format
license: "CC-BY-4.0"
doi: 10.5281/zenodo.5171937
```

Task

- Create a CITATION.cff about your dataset



cff-files support your workflow when publishing your data (e.g. on Zenodo).

Need some inspiration?

<https://citation-file-format.github.io/cff-initializer-javascript/#/>

