Datendokumentation - Die Basis hoher Datenqualität

Vortragender: Roman Gerlach

Mittwoch, 24.02.2021

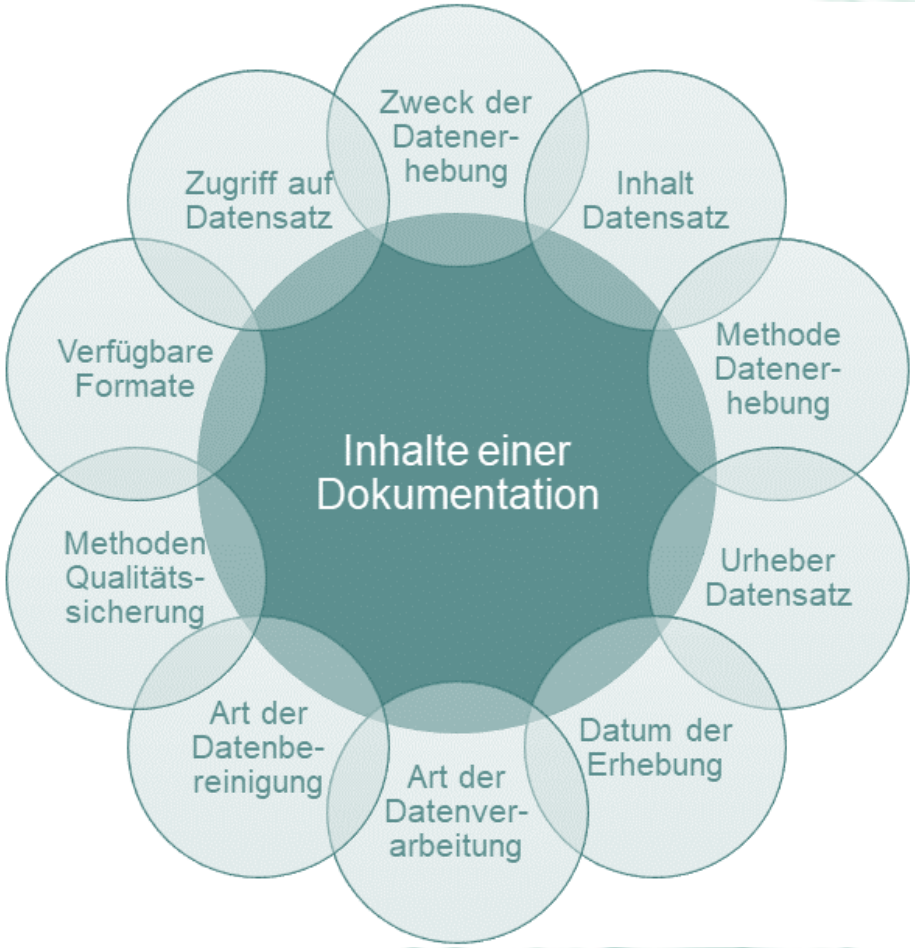
Warum ist Datendokumentation wichtig?

Gute wissenschaftliche Praxis

Leitlinie 12: Dokumentation

„Wissenschaftlerinnen und Wissenschaftler **dokumentieren** alle für das Zustandekommen eines Forschungsergebnisses relevanten Informationen so nachvollziehbar, wie dies im betroffenen Fachgebiet erforderlich und angemessen ist, um das Ergebnis überprüfen und bewerten zu können. […]“

DFG Kodex, 2019



Warum ist Datendokumentation wichtig?

FAIR Guiding Principles (2016)

Findable

F1. **(meta)data** are assigned a globally unique and eternally persistent identifier.
F2. **data are described with rich metadata**.
F3. **(meta)data** are registered or indexed in a searchable resource.
F4. **metadata** specify the data identifier.

Accessible

A1. **(meta)data** are retrievable by their identifier using a standardized communications protocol.
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

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

Interoperable

I1. **(meta)data** use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. **(meta)data** use vocabularies that follow FAIR principles.
I3. **(meta)data** include qualified references to other (meta)data.

Reusable

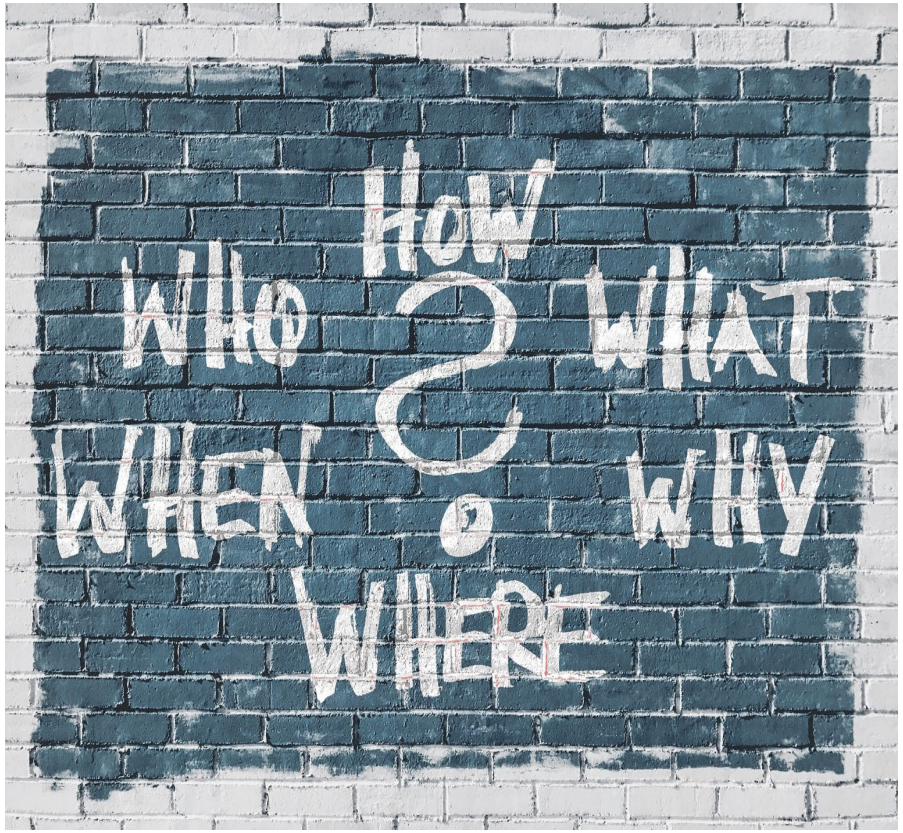
R1. **meta(data)** have a plurality of accurate and relevant attributes.
- R1.1. **(meta)data** are released with a clear and accessible data usage license.
- R1.2. **(meta)data** are associated with their provenance.
- R1.3. **(meta)data** meet domain-relevant community standards.

Warum ist Datendokumentation wichtig?

- Ohne Datenbeschreibung ist eine Nachnutzung von Daten kaum möglich
- Dokumentation ist langfristig die einzige Form der Kommunikation zwischen Datenerzeuger und -nutzer
- Suchmaschinen nutzen vor allem Metadaten, nicht die Daten

Inhalte einer Datendokumentation

- WER hat die Daten erzeugt?
- WIE wurden die Daten erzeugt und verarbeitet?
- WAS ist der Inhalt der Daten?
- WARUM wurden die Daten erzeugt?
- WANN wurden die Daten erzeugt?
- WO wurden die Daten erzeugt?

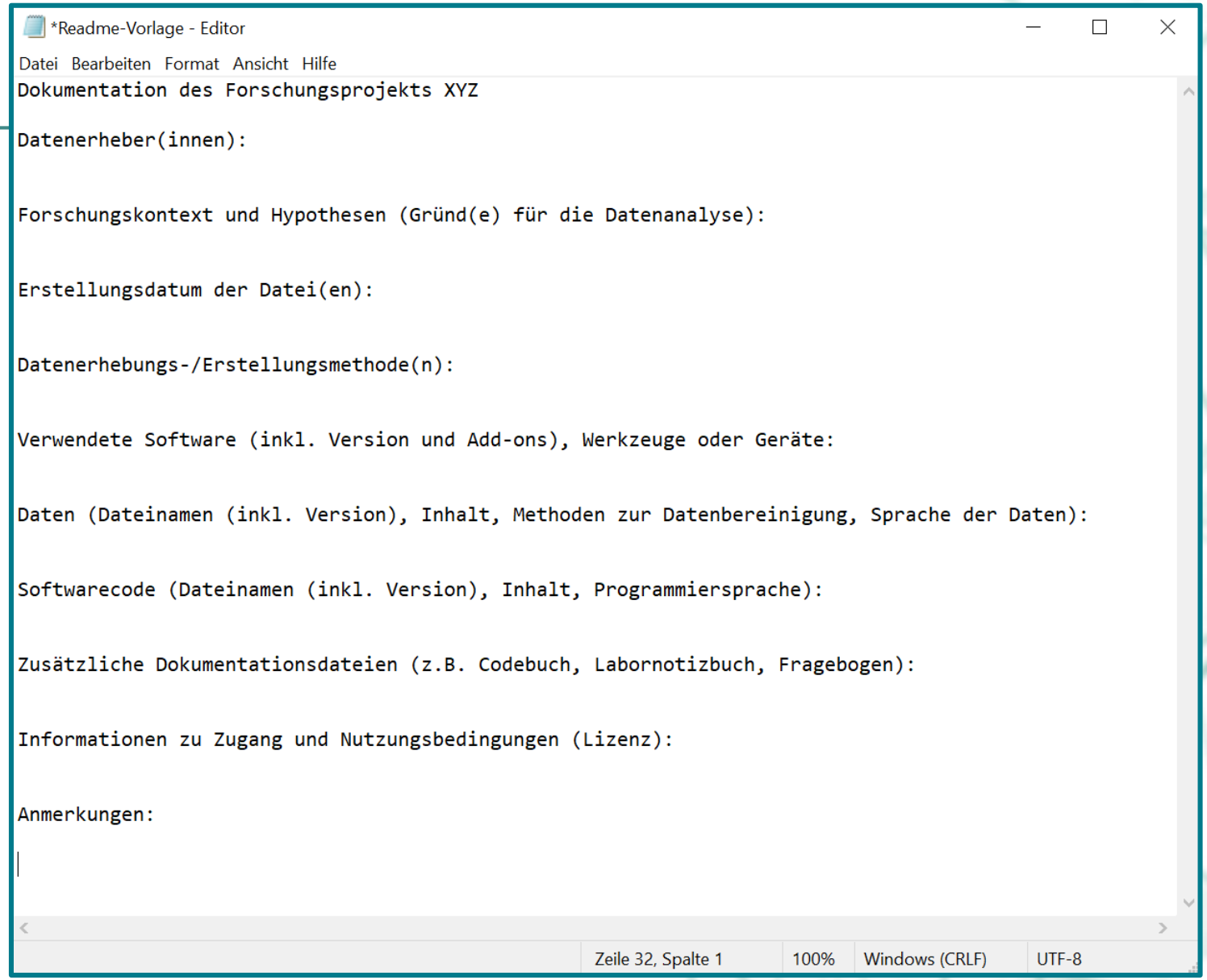


Dokumentationsformen

- README Files
- Tagging von Dateien
- Versionierung
- Data Dictionary
- Codebuch
- Feld- oder Laborbuch
- Artikel in einem Data Journal
- Tools

Readme Files

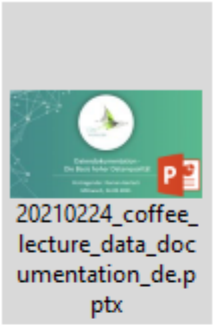
- menschenlesbar
- durchsuchbar
- Metadaten unstrukturiert, aber besser als keine Metadaten

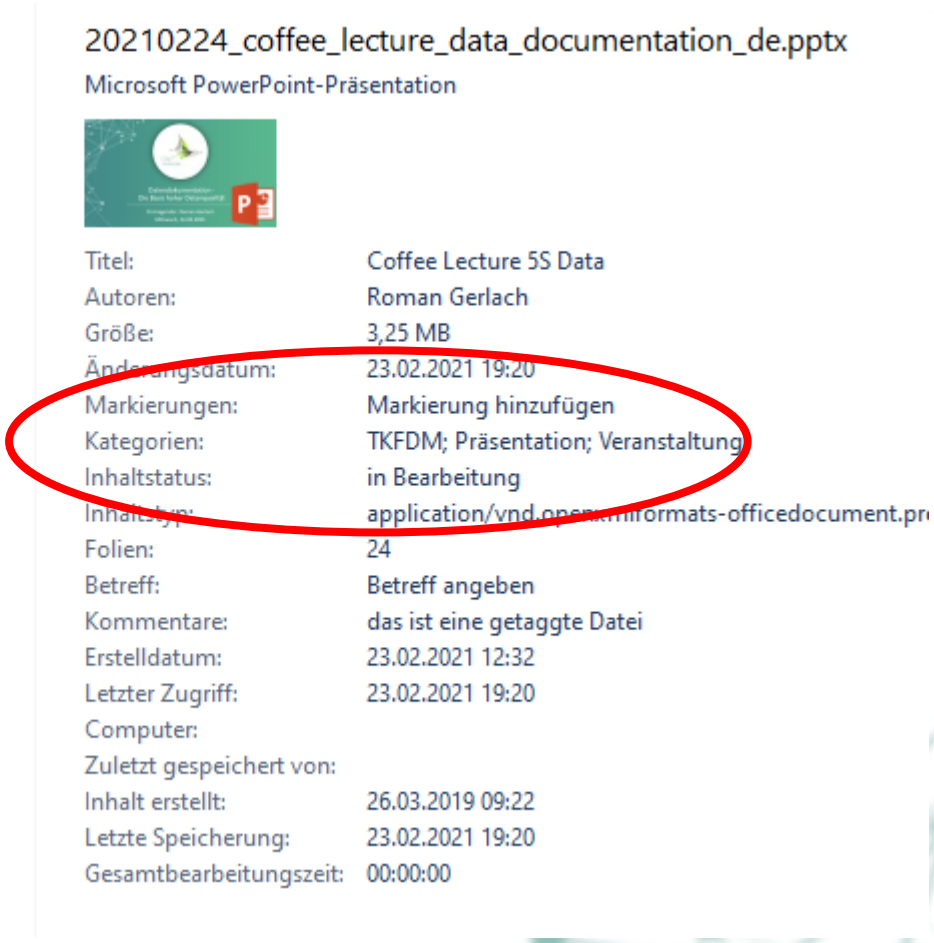


Tagging von Dateien

in Windows:

Schlagworte und Beschreibungen im Explorer hinzufügen





Versionierung

Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

Andreas Rauber, Ari Asmi, Dieter van Uytvanck and Stefan Pröll

Revision of October 20th 2015

I. Making Data Citable

These WGDC recommendations enable researchers and data centers to identify and cite data used in experiments and studies. Instead of providing static data exports or textual descriptions of data subsets, we support a dynamic, query centric view of data sets. The proposed solution enables precise identification of the very subset and version of data used, supporting reproducibility of processes, sharing and reuse of data.

Goals of this WG are to create identification mechanisms that:

- allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner

A. Preparing the Data and the Query Store

Prepare existing data sources and provide the required infrastructure, which is needed for implementing the query based approach.

- **R1 – Data Versioning**: Apply versioning to ensure earlier states of data sets can be retrieved.
- **R2 – Timestamping**: Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp.
- **R3 – Query Store Facilities**: Provide means for storing queries and the associated metadata in order to re-execute them in the future.

B. Persistently Identify Specific Data Sets

When a data set should be persisted, the following steps need to be applied:

- **R4 – Query Uniqueness**: Re-write the query to a

http://dx.doi.org/10.15497/RDA00016

Data Dictionary

- menschenlesbar
- durchsuchbar
- Metadaten strukturiert

The Data Dictionary

File name	Data Type	Method	Creator	Date	Description	Rights	Long-term availability	

Codebuch

- Verzeichnis von Codes, Abkürzungen, Variabeln
- Beschreibung der Struktur in tabellarischen Daten

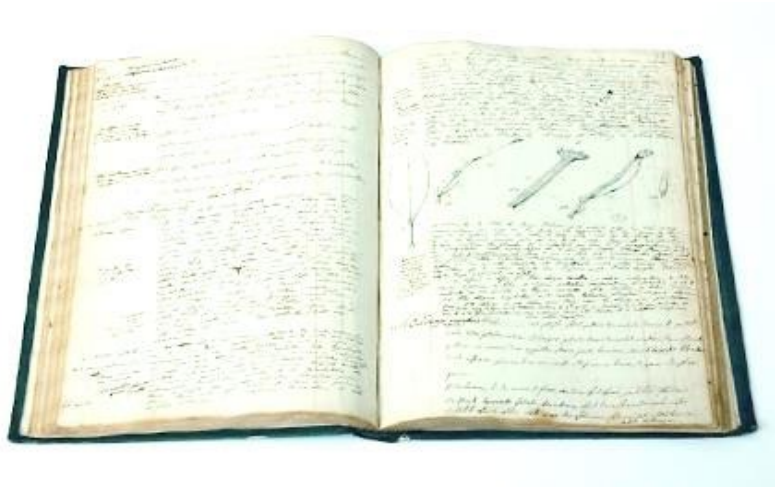
Show rows with cells including:

Variable	Variable name	Mesaurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential order
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender
Date of survey	SURVEY	mm/dd/yyyy	01/01/2015 – 01/01/2016	When the participant completed the survey
Self-reported consumer spending	SPEND	Numeric	0-100,000,000	Self-reported average yearly expenditure
Market sentiment	SENTIMENT	Numeric	1 = negative 2 = neutral 3 = positive	Sentiment towards US domestic economy
Actual GDP growth	GDP	Numeric	-5.0-5.0	Average US yearly GDP growth

Quelle: Bowman, S. How to Make a Data Dictionary. Online verfügbar: https://help.osf.io/hc/en-us/articles/360019739054-How-to-Make-a-Data-Dictionary (Public Domain).

Feld- oder Laborbuch vs. ELN

Feld- oder Laborbuch



nur für Menschen lesbar

Elektronisches Laborbuch

```
<?xml version="1.0"?>
<dwr:DarwinRecordSet
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://rs.tdwg.org/dwc/dwcrecord/  http://rs.tdwg.org/dwc/xsd/tdwg_dwc_classes.xsd"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:dwc="http://rs.tdwg.org/dwc/terms/"
    xmlns:dwr="http://rs.tdwg.org/dwc/dwcrecord/">
    <dcterms:Location>
        <dwc:locationID>http://guid.mvz.org/sites/arg/127</dwc:locationID>
        <dwc:country>Argentina</dwc:country>
        <dwc:countryCode>AR</dwc:countryCode>
        <dwc:stateProvince>Neuquén</dwc:stateProvince>
        <dwc:locality>25 km al NNE de Bariloche por Ruta 40 (=237)</dwc:locality>
    </dcterms:Location>
    <dwc:Occurrence>
        <dcterms:type>PhysicalObject</dcterms:type>
        <dcterms:modified>2009-02-12T12:43:31</dcterms:modified>
        <dcterms:rightsHolder>Museum of Vertebrate Zoology</dcterms:rightsHolder>
        <dcterms:rights>Creative Commons License</dcterms:rights>
```

menschen- und maschinenlesbar

+ durchsuchbar

+ strukturiert

+ standardisiert

Data Journals

- dokumentiert und beschreibt Forschungsdaten in einem Artikel
- informiert über die Datenerhebung, Charakteristiken, Funktionen und potenzielle Nachnutzungsmöglichkeiten

npg nature publishing group

SCIENTIFIC DATA

WILEY ONLINE LIBRARY

RMetS Royal Meteorological Society

Geoscience Data Journal

A peer-reviewed open-access journal

Biodiversity Data Journal

Making your data count!

ICSU CODATA

DATA SCIENCE Journal

Tools

- **Open Science Framework** (OSF) – free, open platform supporting collaboration and sharing of research
  https://osf.io/
- **CEDAR Workbench** – creating and submitting FAIR metadata
  https://metadatacenter.org/
- **REDCap** - Research Electronic Data Capture
  https://projectredcap.org/
- **git** - freie Software zur verteilten Versionsverwaltung
  https://git-scm.com/

Automatisierte Datenbeschreibung

- **Ziel:** automatisierte Erzeugung standardkonformer Metadaten
- Bsp. Fotografie

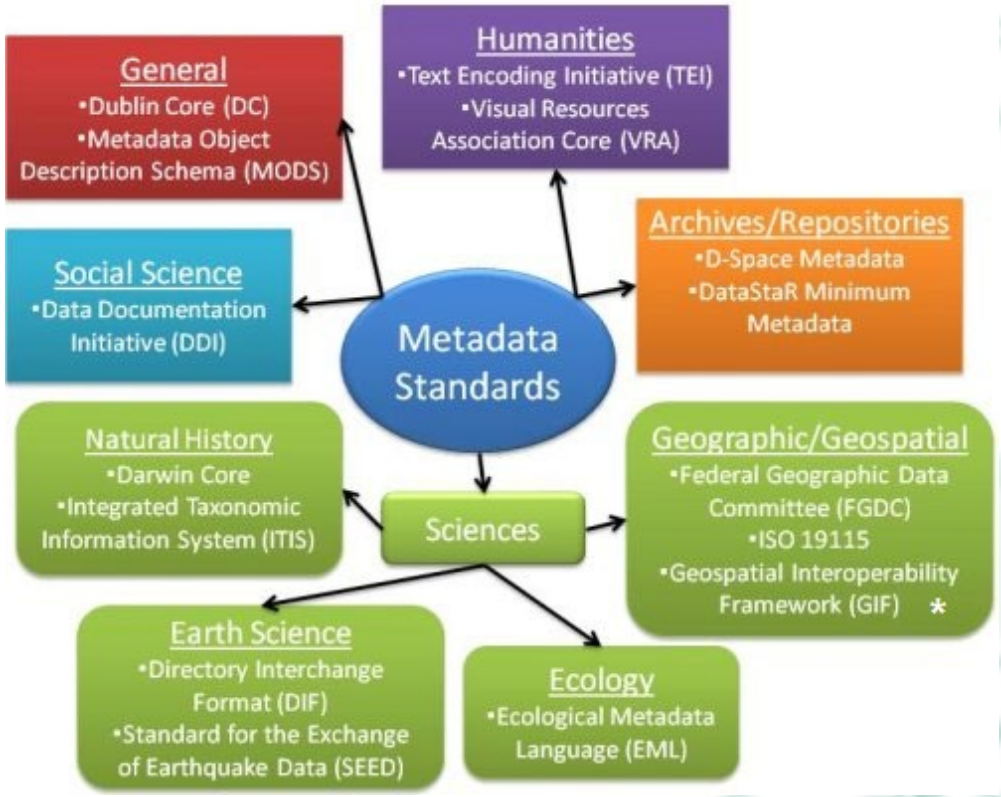
Exif Properties | Maker Data | Summary

Item	Details
Image Description	DCF 1.0
Make	Minolta Co., Ltd.
Model	DiMAGE S304
Orientation	Normal
XResolution	72.00
YResolution	72.00
Resolution Unit	Inch
Software	Adobe Photoshop CS Win
Date Time	2005:02:28 10:08:32
YCb Cr Positioning	Centered
Exposure Program	Normal
ISOSpeed Ratings	100
Exif Version	"0210"
Date Time Original	2001:01:02 15:43:30
Date Time Digitized	2001:01:02 15:43:30
Components Configuration	YCbCr
Shutter Speed Value	0.0039 sec (1/256)
Aperture Value	F6.0
Exposure Bias Value	0/10
Max Aperture Value	F3.7
Metering Mode	MultiSegment
Light Source	Unidentified
Flash	Off
Focal Length	11.81 mm
Flash Pix Version	"0100"

http://www.alexnolan.net/photodata

Metadatenstandards

- Metadata Standards Catalog
  https://rdamsc.bath.ac.uk/
- RDA Metadata Directory
  https://rd-alliance.github.io/metadata-directory/
- FAIRsharing.org
  https://fairsharing.org/

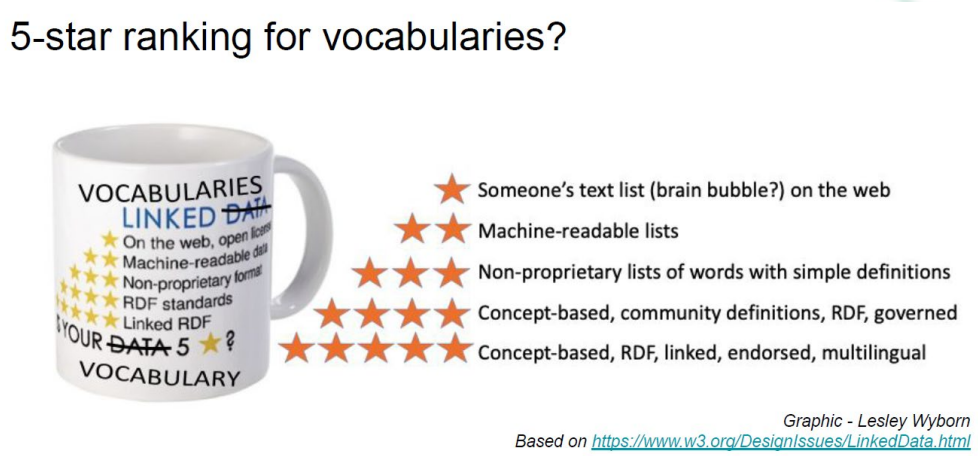


Metadata Concept Map by AmandaTarbet

Vokabularien / Thesauri / Ontologien

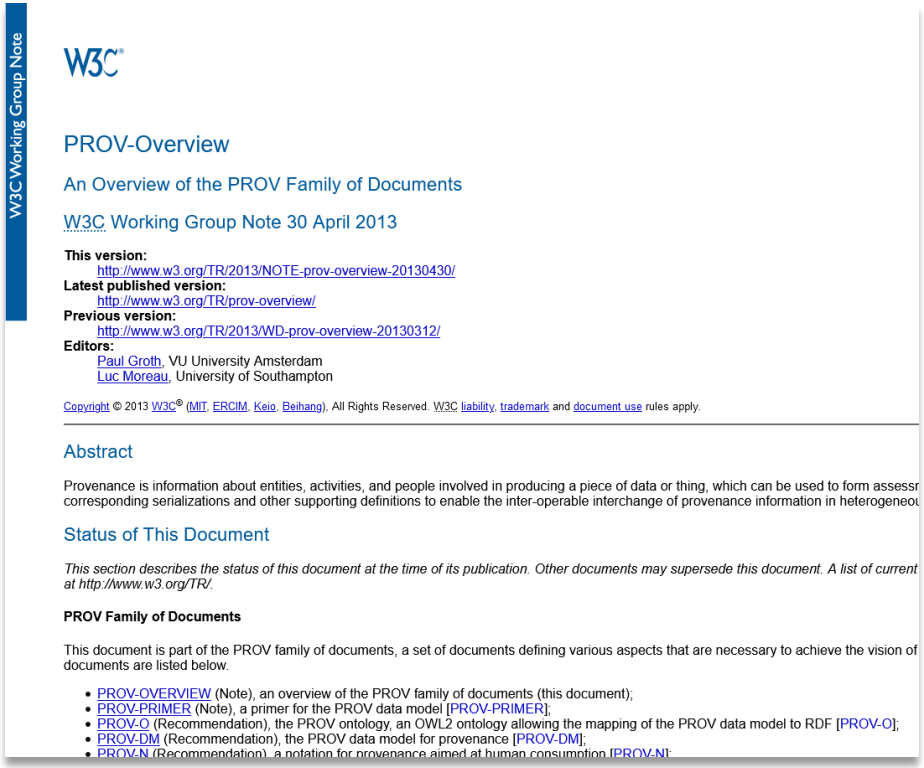
- fördern die Konsistenz und Genauigkeit
- helfen dabei Mehrdeutigkeiten aufzulösen
- ermöglichen es Beziehungen zwischen Begriffen und Konzepten darzustellen

- Basel Register of Thesauri, Ontologies & Classifications
  http://www.bartoc.org/
- GFBio Terminology Service
  https://terminologies.gfbio.org/



Provenance & Workflows

W3C PROV
https://www.w3.org/TR/prov-overview/



W3C Working Group Note

W3C

PROV-Overview

An Overview of the PROV Family of Documents

W3C Working Group Note 30 April 2013

This version:
http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
Latest published version:
http://www.w3.org/TR/prov-overview/
Previous version:
http://www.w3.org/TR/2013/WD-prov-overview-20130312/
Editors:
Paul Groth, VU University Amsterdam
Luc Moreau, University of Southampton

Abstract

Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessr
corresponding serializations and other supporting definitions to enable the inter-operable interchange of provenance information in heterogeneou

Status of This Document

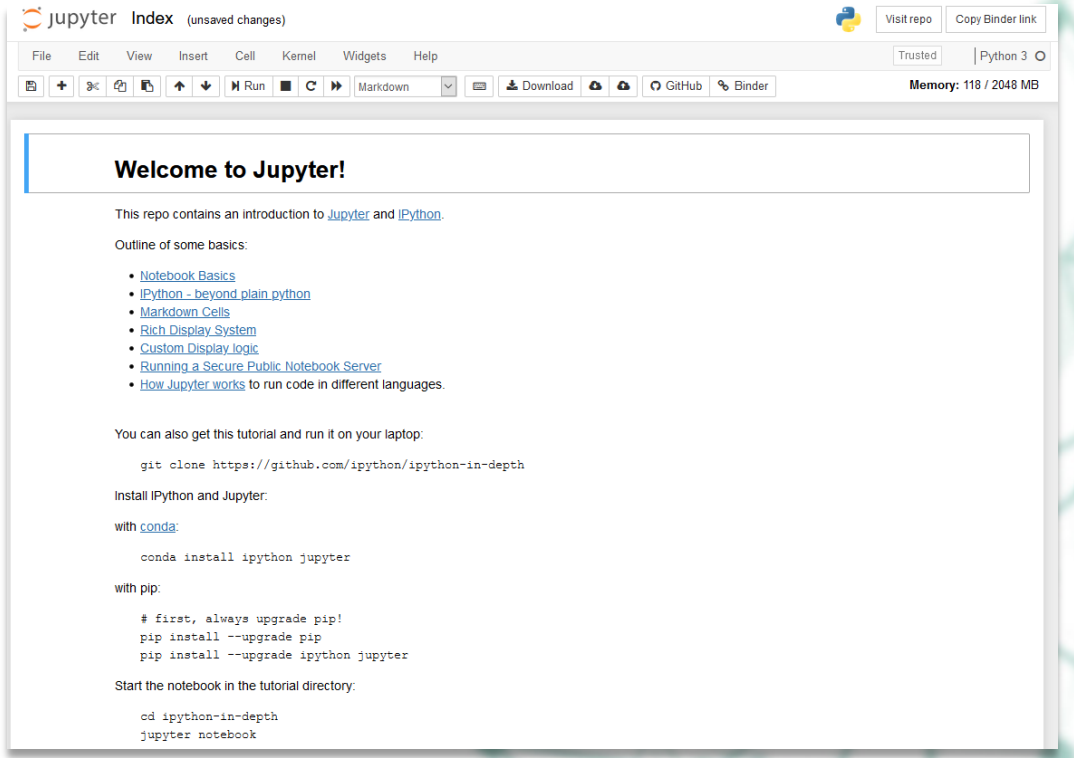
*This section describes the status of this document at the time of its publication. Other documents may supersede this document. A list of current
at http://www.w3.org/TR/.*

PROV Family of Documents

This document is part of the PROV family of documents, a set of documents defining various aspects that are necessary to achieve the vision of
documents are listed below.

- PROV-OVERVIEW (Note), an overview of the PROV family of documents (this document);
- PROV-PRIMER (Note), a primer for the PROV data model [PROV-PRIMER];
- PROV-O (Recommendation), the PROV ontology, an OWL2 ontology allowing the mapping of the PROV data model to RDF [PROV-O];
- PROV-DM (Recommendation), the PROV data model for provenance [PROV-DM];
- PROV-N (Recommendation), a notation for provenance aimed at human consumption [PROV-N];

Jupyter Notebook
https://jupyter.org/



jupyter Index (unsaved changes)

Visit repo | Copy Binder link

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Markdown | Download | GitHub | Binder

Memory: 118 / 2048 MB

Welcome to Jupyter!

This repo contains an introduction to Jupyter and IPython.

Outline of some basics:

- Notebook Basics
- IPython - beyond plain python
- Markdown Cells
- Rich Display System
- Custom Display logic
- Running a Secure Public Notebook Server
- How Jupyter works to run code in different languages.

You can also get this tutorial and run it on your laptop:

```
git clone https://github.com/ipython/ipython-in-depth
```

Install IPython and Jupyter:

with conda:

```
conda install ipython jupyter
```

with pip:

```
# first, always upgrade pip!
pip install --upgrade pip
pip install --upgrade ipython jupyter
```

Start the notebook in the tutorial directory:

```
cd ipython-in-depth
jupyter notebook
```

Weitere Quellen

- Forschungsdatenmanagement und das TKFDM:
  - Portal von TKFDM
  - TKFDM Community auf Zenodo
- Datendokumentation:
  - Katarzyna Biernacka, Kerstin Helbig, Matthias Senft, Ute Trautwein-Bruns, Datendokumentation leicht gemacht! Ein interaktiver Online-Workshop, DINI Nestor AG Schulungen/Fortbildungen, 2020
  - forschungsdaten.info „Datendokumentation." accessed 24.02.2021
  - Managing Qualitative Social Science Data - An interactive online course . accessed 24.02.2021
  - Guide for data documentation. accessed 24.02.2021
  - How to Tag Files in Windows for Easy Retrieval. accessed 24.02.2021

Vielen Dank für Ihre Aufmerksamkeit

Kontakt:
info@forschungsdaten-thueringen.de