FIELD OF ACTION RESEARCH

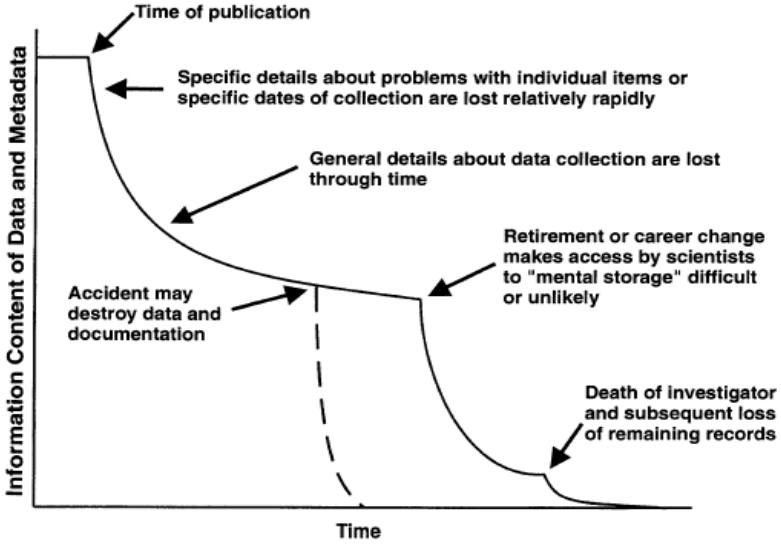
Metadata
RDM Team TU Graz

Content

- Introduction
  - Why is documentation important?
  - What is metadata?
  - Metadata categories
- Metadata in practice
  - Documentation planning
  - Documentation of research processes
  - Save metadata
  - Automatically captured metadata
  - Metadata extraction
  - Data management software
- Metadata schema and standards

Why is documentation important?

- most of the data is **useless without a description**
- over time, information is forgotten - documentation enables **long-term understanding of** data no archiving without metadata
- **Overview of** versions and different formats **is lost**
- Documentation is the essential element of good data management and **facilitates data exchange**
- structured metadata enable **machine processing of** data (search, automation)
- Despite initial extra work, **future work** is **made easier**



MICHENER, William K., et al. Nongeospatial metadata for the ecological sciences. Ecological Applications, 1997, 7. Jg., Nr. 1, S. 330-342.

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

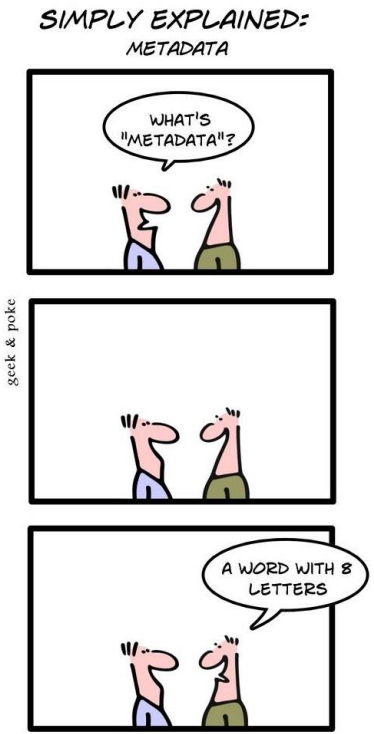
What is metadata?

Metadata are

- "Data about data"
- serve the documentation
- contain descriptive information about the context of data

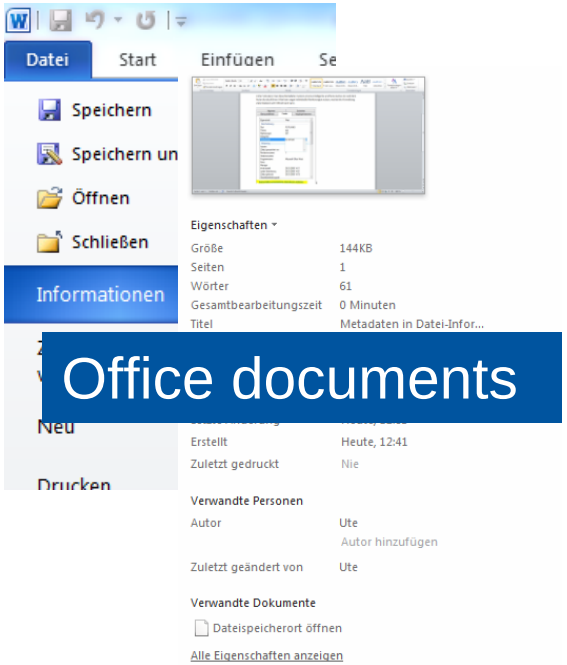
Context is e.g.

- ... Technology used for data generation (hardware/software)
- ...administrative details (project, participants, institution, etc.)
- ...relations to publications, other data, persons, projects etc.

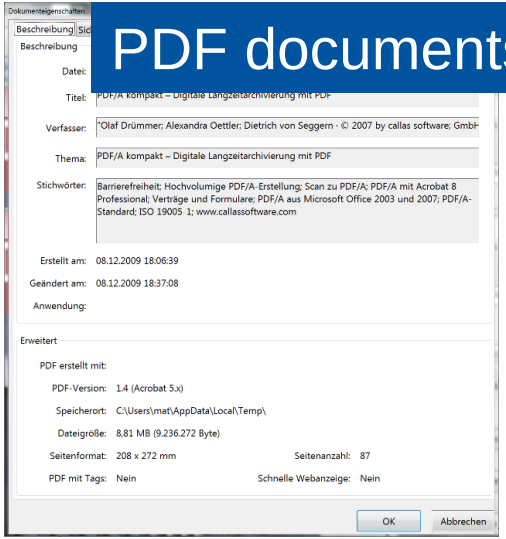


Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Examples Metadata in File Information

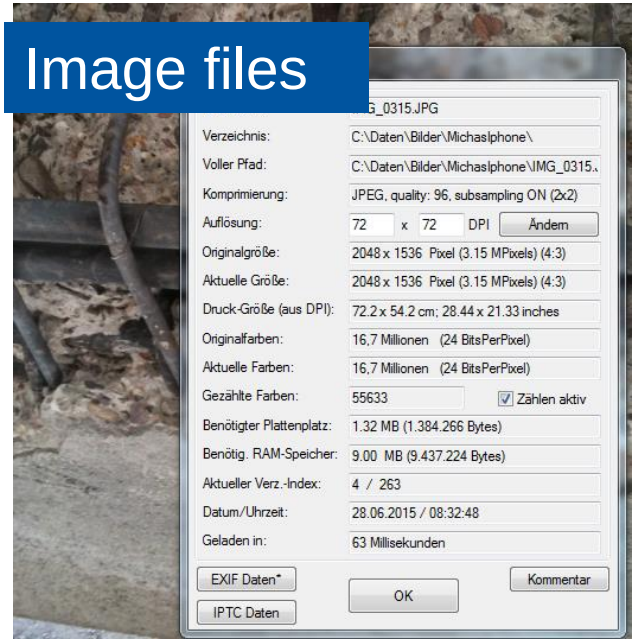


Office documents



PDF documents

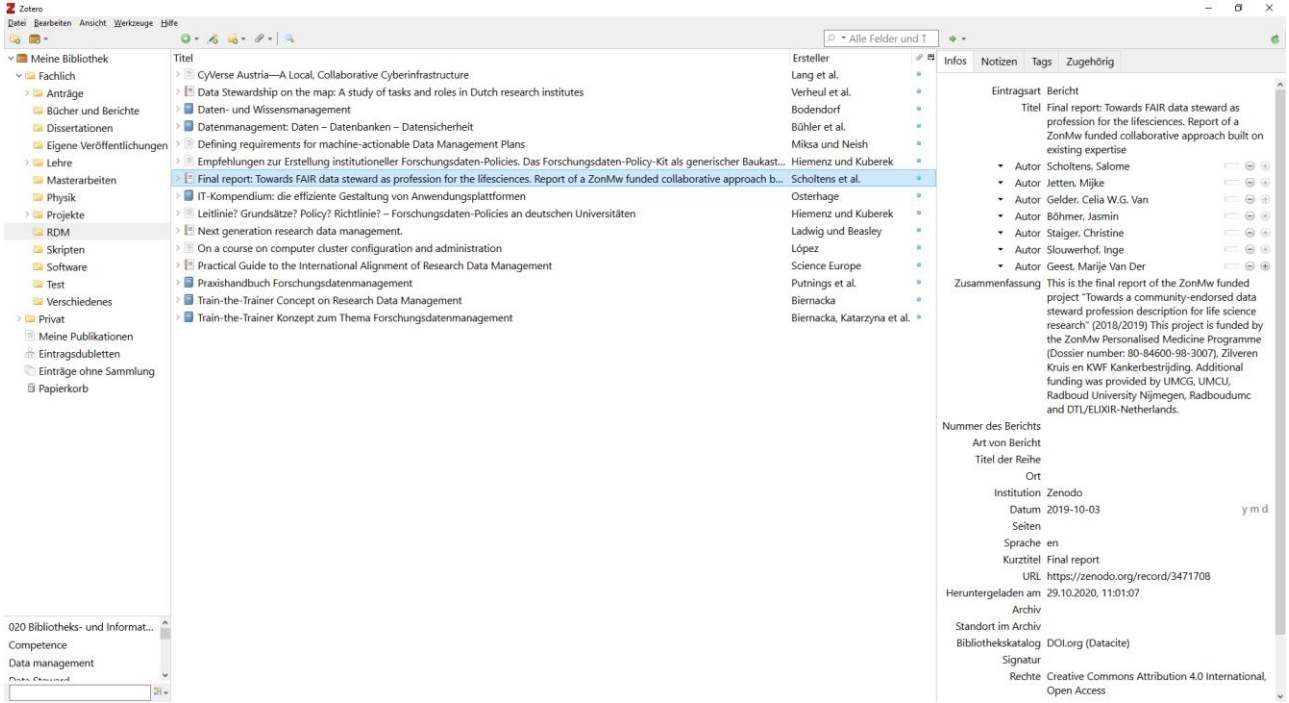
Image files



Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Examples of literature management

- EndNote
- Zotero
- Citavi
- Mendeley
- JabRef



Examples of metadata for research data

References

Test design

Topic

Identifier

File size

System requirements

Metadata
-data about data-

Editor

File format

Survey method

Evaluation strategy

Subject

File type

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Metadata categories

Administrative metadata

- Technical specifications and parameters, legal information
- Examples: File format, file size, date and time of creation, licences, access rights.

Structuring metadata

- Structure of the data and linkage with other data
- Examples : Table description, chapters and pages in a book

Descriptive metadata

- Describe the content of the data
- Examples: Description of the experimental set-up, vocabulary (predefined wording).

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Documentation planning

Data Management Plan (DMP)

- What information do you want to map?
- What standards are there for this?
- Are there regulations in the working group/project/institute for the description of data?
- Where do you want to reuse the metadata? Do you want/need to publish/share the data later? Should it be available to others in the working group?
- What can facilitate documentation?
  - Automation scripts
  - Forms for frequently used data sets
  - Metadata tools
  - Databases, interfaces

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

What do you need to ...

... to find data

- General search criteria: Link to publication, author, year, project
- General search criteria for the research question: variables collected/simulated/observed, controlled variables, boundary conditions and parameters
- Subject-specific search criteria: Parameters of the system under consideration (e.g. force fields in thermodynamics), parameters of the survey method (e.g. temporal and spatial resolution, geographical classification).

... understand data

- Mapping the research process: steps, methods, software, hardware, parameters
- Variables collected/observed/simulated
- Controlled variables

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Documentation of research processes

A process is a sequence of activities that have a temporal beginning and an end

- Process thinking fits well with the scientific way of working
- Flowcharts serve to visualise processes and are suitable for documenting research processes in a comprehensible way.
- Visualisation supports structured work!
- The simplest representation of a process is the black box as a symbol for "input-processing-output" (EVA principle).
- Modularisation of complex contexts

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

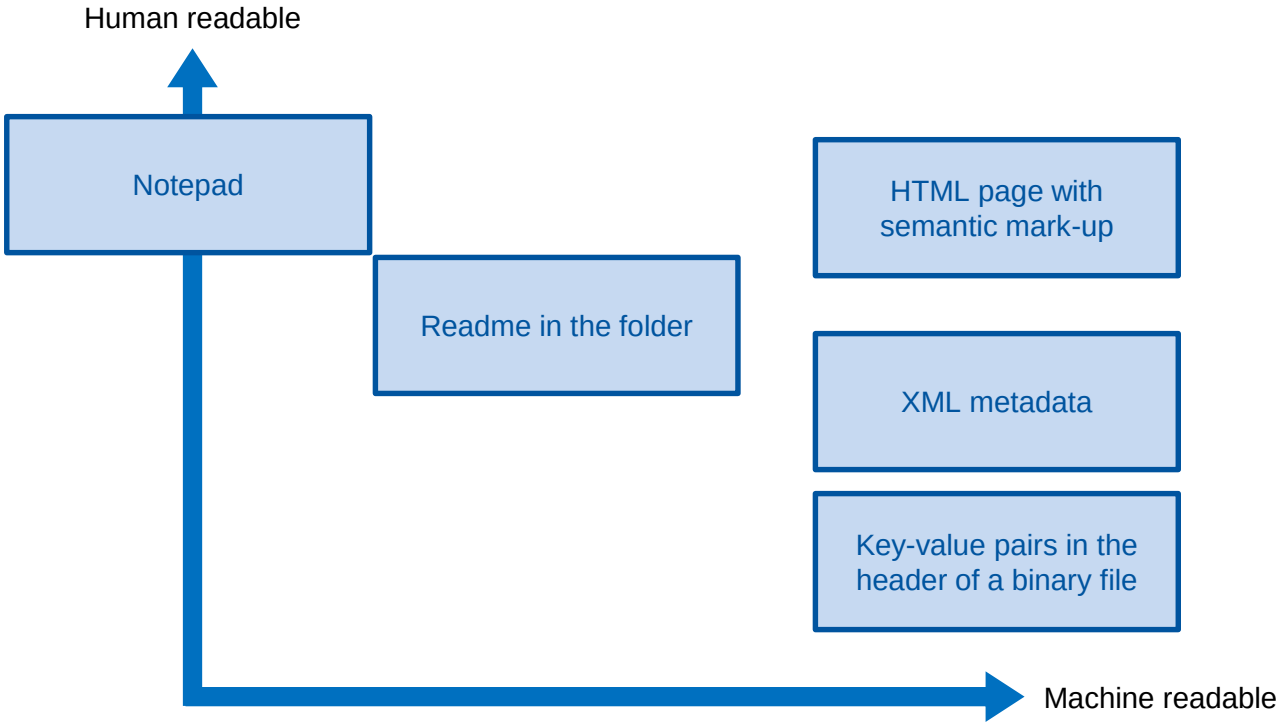
Structured metadata preferred

- Both human- and machine-readable
- Supplementary short textual description possible
  - Full text indexing and search possible
  - But: Search results less precise (Remember the usefulness of many Google search results).
  - In the case of data publication: add a link to the paper or report (description, metadata, use, sources....).
- Metadata schema (e.g. ICAD)
  - Use a standard or offer a mapping

Unstructured	Structured
"The experiment was conducted in Dresden on 12th May 2016"	"Date: 20160512; Place: Dresden..."
Easy to read for humans, but difficult to process by machine	Can be read by both humans and machines.

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Structuring metadata



Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Save metadata

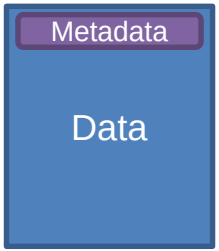
- Within the file (there are standards for many file formats)
- In a README file or other text file, table, XML file...
- Database
- Data management system (e.g. ICAT, MASI-Metadata Management for Applied Sciences)...

Challenges

- Scalability with growing data volume
- Linking data and metadata (using links and PIDs)

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

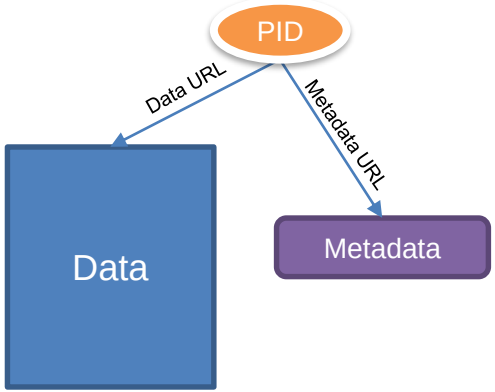
Linking data and metadata

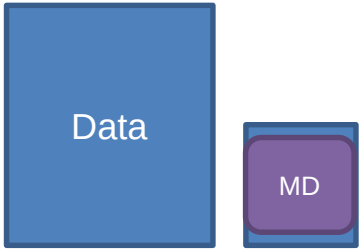


Metadata in the data (in the header of data formats, e.g. HDF5)

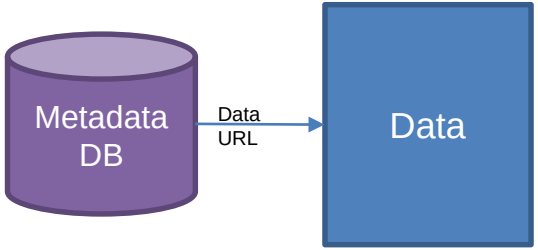
Metadata on the data
(e.g. object storage, file and folder names)

Persistent identifiers
with linking of metadata and data (e.g. DOI, EPIC-PID, URN





Metadata with the data
(e.g. readme-file, metadata)



Metadata DB
with link to the data
(e.g. search index, repository)

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Table documentation

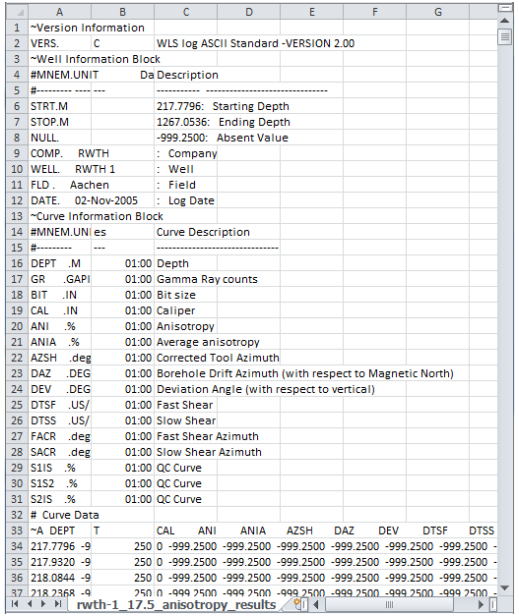
Recommendation: Information in file itself (e.g. first spreadsheet)

Metadata	Description
Description of table/worksheets	What is the purpose of the table/ worksheets?
Worksheets name	Listing of the names of the worksheets.
Column heading	Each column of a table must have a name.
Column description	Description and listing of format specifications, abbreviations, codes, value lists, input conventions, specialized vocabularies, characters for empty cells or measurement units used in the respective column.
Number of columns/rows/worksheets	How many columns/rows/sheets does the table/spreadsheet contain?
Relations/Formulas/Macros	What relations/formulas/macros exist in the spreadsheet

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Automatically captured metadata - Measurements

- In some cases, devices/software record relevant metadata.
  - Camera automatically writes metadata to the generated image file
  - Measuring instruments write metadata in the header to the measurement data
  - Units generate separate configuration files/ metadata files in addition to the measurement data
- In some cases, metadata collection can be configured in the unit software.
- In some cases, the export of metadata must be deliberately initiated.
- Electronic lab books (e.g. eLabFTW, https://www.elabftw.net/; or https://elabftw.vice.cyverse.tugraz.at/login.php)



	A	B	C	D	E	F	G
1	~Version Information						
2	VERS.	C	WLS log ASCII Standard -VERSION 2.00				
3	~Well Information Block						
4	#MNEM.UNIT	Da	Description				
5	#--------- ----	---	---------- ----------				
6	STRT.M		217.7796: Starting Depth				
7	STOP.M		1267.0536: Ending Depth				
8	NULL.		-999.2500: Absent Value				
9	COMP. RWTH		: Company				
10	WELL. RWTH 1		: Well				
11	FLD . Aachen		: Field				
12	DATE. 02-Nov-2005		: Log Date				
13	~Curve Information Block						
14	#MNEM.UNI	es	Curve Description				
15	#---------	---	----------				
16	DEPT .M	01:00	Depth				
17	GR .GAPI	01:00	Gamma Ray counts				
18	BIT .IN	01:00	Bit size				
19	CAL .IN	01:00	Caliper				
20	ANI .%	01:00	Anisotropy				
21	ANIA .%	01:00	Average anisotropy				
22	AZSH .deg	01:00	Corrected Tool Azimuth				
23	DAZ .DEG	01:00	Borehole Drift Azimuth (with respect to Magnetic North)				
24	DEV .DEG	01:00	Deviation Angle (with respect to vertical)				
25	DTSF .US/	01:00	Fast Shear				
26	DTSS .US/	01:00	Slow Shear				
27	FACR .deg	01:00	Fast Shear Azimuth				
28	SACR .deg	01:00	Slow Shear Azimuth				
29	S1IS .%	01:00	QC Curve				
30	S1S2 .%	01:00	QC Curve				
31	S2IS .%	01:00	QC Curve				
32	# Curve Data						
33	~A DEPT T		CAL ANI	ANIA	AZSH	DAZ	DEV DTSF DTSS
34	217.7796 -9	250	0 -999.2500 -999.2500	-999.2500	-999.2500	-999.2500	-999.2500 -
35	217.9320 -9	250	0 -999.2500 -999.2500	-999.2500	-999.2500	-999.2500	-999.2500 -
36	218.0844 -9	250	0 -999.2500 -999.2500	-999.2500	-999.2500	-999.2500	-999.2500 -
37	218.2368 -9	250	0 -999.2500 -999.2500	-999.2500	-999.2500	-999.2500	-999.2500 -

rwth-1_17.5_anisotropy_results

Example: Metadata in the header to the data log of an acoustic borehole logging probe

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version; adapted

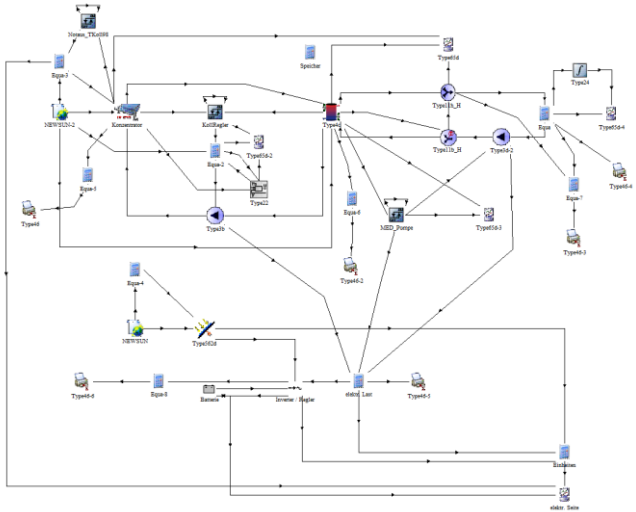
Automatically captured metadata - Simulation

- In part, simulation programmes output metadata
  - Date and time
  - Start, stop and step size of the simulation
  - Convergence tolerances
- In automatic variant studies (e.g. with Excel), parameters and designations are also written down
- Choose names of output files and column labels sensibly
- UNITS!
- Electronic lab books (e.g. eLabFTW, https://www.elabftw.net/; or https://elabftw.vice.cyverse.tugraz.at/login.php)

```
VERSION 17
*******************************************************************************
*** TRNSYS input file (deck) generated by TrnsysStudio
*** on Samstag, April 04, 2020 at 16:29
*** from TrnsysStudio project: D:\227_NEWSUN\05_AP2_Glob_num_SysOpt
(4.5PM)\30_Szenariensimulation01a\Szenariensim01.tpf
***
*** If you edit this file, use the File/Import TRNSYS Input File function in
*** TrnsysStudio to update the project.
***
*** If you have problems, questions or su
*** TRNSYS distributor or mailto:software
***
*****************************************

*****************************************
*** Units
*****************************************

*****************************************
*** Control cards
*****************************************
* START, STOP and STEP
CONSTANTS 3
START=0
STOP=8760
STEP=0.033333332
SIMULATION  START STOP STEP     ! Start t
TOLERANCES 0.001 0.001     ! Integration
LIMITS 30 500 50     ! Max iterationsMax
DFQ 1     ! TRNSYS numerical integration
WIDTH 80     ! TRNSYS output file width,
LIST     ! NOLIST statement
     ! MAP statement
SOLVER 0 1 1     ! Solver statementMinimu
NAN_CHECK 0     ! Nan DEBUG statement
OVERWRITE_CHECK 0     ! Overwrite DEBUG s
TIME_REPORT 0     ! disable time report
EQSOLVER 0     ! EQUATION SOLVER statemen
```



Example: Metadata in the header of the input file for a TRNSYS simulation (www.trnsys.com)

Metadata extraction

Automatic extraction of metadata - no more manual work!

Supports the collection of metadata for uniform data sets!

Collect metadata in your own working environment!

- Where is the metadata hidden?
- Can they be extracted automatically? With a script?
- Extraction for literature management programmes (e.g. Zotero)
- **Cyverse also extracts metadata**

Metadata extraction software - example: Apache Tika

- Based on Java → Windows, Linux, Mac
- Extracts from many file types
- Output of metadata in different formats
- Graphical user interface and command line tool or server

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Data management software

Stores both metadata and data

Required functions

- Search function (find data via metadata)
- Support for metadata extraction, creation, editing ....
- Integration into the user environment (e.g. browser-based)
- Interfaces for accessing the data to be integrated into the analysis environment (e.g. POSIX, http, REST-API)

Attention: Prevent dependence on software - Can data/metadata be exported from the system?

Examples: ICAT, MASI, KIT Datamanager

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Metadata schema and standards

- Metadata schema
  - Defines the structure and content of the metadata fields
- Metadata standards
  - are standardised metadata formats
  - ensure comparability
  - facilitate the exchange of metadata (interoperability)
  - enable machine readability
- Quasi-standards
  - No official standard adopted by a standards body (such as ISO or W3C)
  - but widespread and accepted

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Example metadata schema DataCite

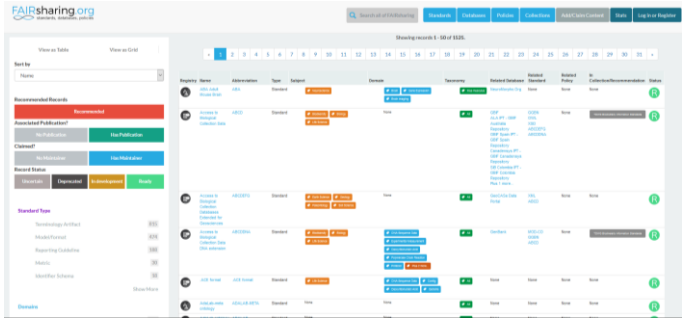
- http://schema.datacite.org/
- Metadata scheme for publication and citation of research data
- Includes metadata core elements and recommendations for use
- used for DOI registration
- Metadata generator: https://dhvlab.gwi.uni-muenchen.de/datacite-generator/
- Used at **invenioRDM** (institutional repository at TU Graz)

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version, adapted

Search for metadata standards

- https://rd-alliance.github.io/metadata-directory/

- https://fairsharing.org/



Advantages and disadvantages of standards

Advantages

- Clearly defined meaning of fields
- Partially controlled vocabularies
- Machine readability
- Easy interchangeability

Disadvantages

- Low flexibility
- Sometimes high complexity

Source: TU9 German Universities of Technology, http://doi.org/10.5281/zenodo.2660187; german version

Further information

- Interactive online course on metadata

  https://mantra.edina.ac.uk/documentation_metadata_citation/
- Train-the-trainer concept on research data management version 3.1, unit 8: documentation and metadata (page 88); good summary with further literature

  https://doi.org/10.5281/zenodo.4322849
- Videos (2 - 8 min):
  - https://www.youtube.com/watch?v=4HJENeUY4Uc
  - https://www.youtube.com/watch?v=JDueeDrQdLU
  - https://www.youtube.com/watch?v=y7XulIpa6gk