

Making data better by adding some meta(data)

Alex Ball

2016-06-15

University of Bath

Metadata

What is metadata?

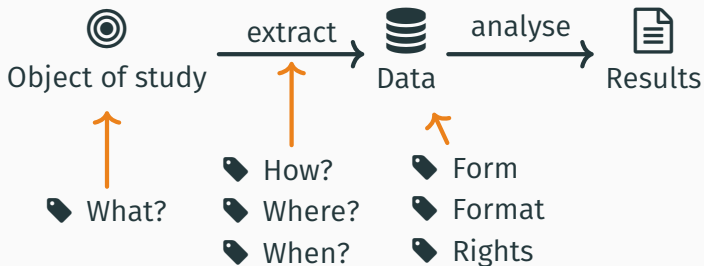
- Literally 'data about data'
- Information that helps you work with other information



- Context determines whether something is **data** or **metadata**

What is metadata?

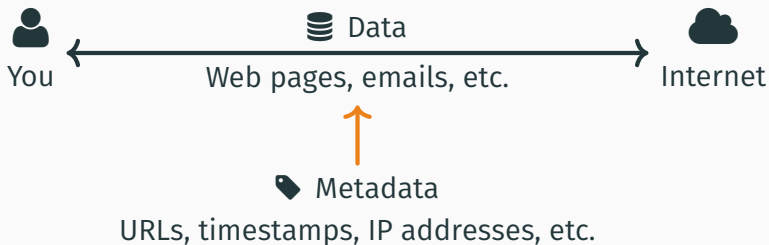
- Literally 'data about data'
- Information that helps you work with other information



- Context determines whether something is **data** or **metadata**

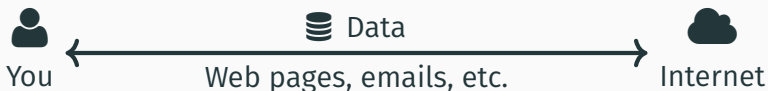
Example: Internet traffic

Your perspective:



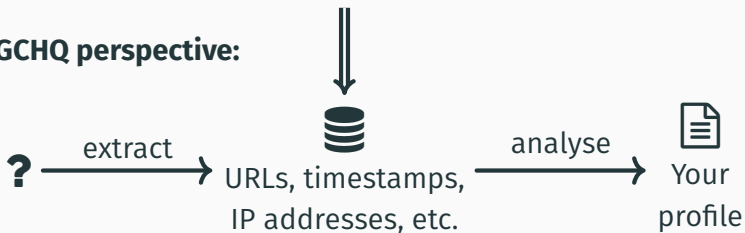
Example: Internet traffic

Your perspective:



URLs, timestamps, IP addresses, etc.

GCHQ perspective:



Types of metadata

Metadata is defined by what you are using it to achieve:

- Reference** Identifying, citing, searching for a known resource
- Descriptive** Speculative searching
- Provenance** Assessing authenticity or trustworthiness
- Contextual** Understanding a resource (externally)
 - Rights** Securing data against unauthorized/illegal actions
- Packaging** Arranging components of a resource
 - Fixity** Checking integrity
- Structural** Loading/opening a file
- Semantic** Understanding a resource (internally)

Types of metadata

As a researcher, you are mostly concerned with

- **Discovery metadata** – help other researchers find the data, give you credit, increase your impact
- **Contextual metadata** – keep your institution and funder happy
- **Discipline-specific metadata** – ensure you yourself (later), your colleagues and your peers can understand and use/reuse the data

Data Discovery Services

About the DCS

Search for data

Results

You are searching for...

everything in the catalogue.
2130 results returned in **0.16** seconds.

Shortcut search

Click to search records for

[Atmosphere](#)[Oceans](#)[Biota](#)[Geo-scientific Information](#)[Inland waters](#)[Environment](#)

Bookmark with:

Search for data

Please use one or more of the options below to search the catalogue. Alternatively, try our [advanced text search](#).

Free text

Please enter your search terms below or leave blank to match everything ([help](#))

Only return records with an online resource

Date range

Match data overlapping the period ([help](#))

From

To



Dates should be in the format YYYY-MM-DD or YYYY

Geographic search

The map display lets you draw a box to define an area of interest and only returns records that intersect the box. We also offer predefined areas below ([help](#))

Define the search box coordinates - input order is most westerly, southerly, easterly, northerly

Welcome to the INSPIRE geoportal

The INSPIRE Directive requires the Commission to establish a community geo-portal and the Member States shall provide access to their infrastructures through the geo-portal as well as through any access points they themselves decide to operate.

[More...](#)

Discovery / Viewer

Search, discover and access geographic information provided by European governmental, commercial, and non-commercial organizations.

[More ...](#)



Validator

The purpose of the INSPIRE Metadata Validator is to test the compliancy of INSPIRE metadata with the INSPIRE Metadata Regulation.

[More ...](#)

```
Invalid Element
number of inci
(2.2.5) Unique
(1/*:identi
(2005/gmd)
(2005/gmd)
(2.4) For dat
/www.isoct
```

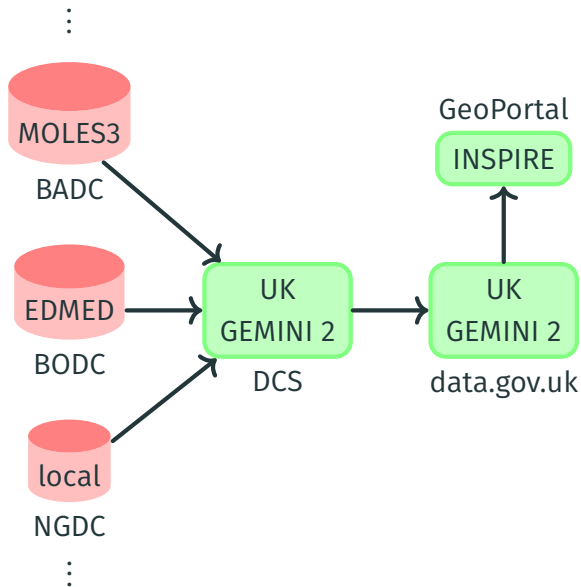
Metadata Editor

Create metadata according to the INSPIRE implementing rules.





[More ...](#)



Geospatial data: ISO 19115, ISO 19139



Search

-  CESSDA Catalogue
-  Browse by Topic
-  Browse by Keyword
-  Browse by Data Publisher

The CESSDA Catalogue

- Provides a seamless interface to datasets from social science data archives across Europe
- May be searched via a free text search from the top search box
- Browsed via the options in the left hand side menu
- Can be viewed in any of nine languages. The default language is dependant on the regional setting of the computer user
- The language can be switched at any time by selecting the relevant language from the drop down list above. **Please note:** changing the language will take the user back to this CESSDA Catalogue home page
- **In the free text search:**
 1. * truncation is supported
 2. '?' wildcards are supported

CESSDA Data Publishers

- (9140 studies available)
- [APIS](#) (27 studies)
- [ADPSS-Sociodata](#) (59 studies)
- [CSDA](#) (1009 studies)
- [DANS](#) (124 studies)
- [ADP](#) (1087 studies)
- [FORS](#) (754 studies)
- [FSD](#) (2010 studies)
- [NSD](#) (1034 studies)
- [ISSDA](#) (97 studies)
- [LIDA](#) (423 studies)
- [NSD Metadata](#) (1034 studies)
- [SND](#) (196 studies)
- [ESSDA](#) (15 studies)
- [UKDA](#) (642 studies)
- [GESIS ZACAT](#) (629 studies)

DDI Specification

For the most up-to-date news on the development process of DDI ([summary](#)), please visit [the DDI Collaboration Wiki](#).

Latest version

DDI Lifecycle 3.2

DDI-Lifecycle is designed to document and manage data across the entire life cycle, from conceptualization to data publication and analysis and beyond. It encompasses all of the DDI-Codebook specification and extends it. Based on XML Schemas, DDI-Lifecycle is modular and extensible.

Read more about DDI 3.2...

[Online field level documentation](#)[XSD Schema entry point](#)[Download documentation](#)

DDI Codebook 2.5

DDI-Codebook is a more light-weight version of the standard, intended primarily to document simple survey data. Originally DTD-based, DDI-C is now available as an XML Schema.

Read more about DDI 2.5...

[Online field level documentation](#)[XSD Schema entry point](#)[Download documentation](#)

<http://www.ddialliance.org/Specification/>

Find data for research

Find, access, and re-use data for research - from over one hundred Australian research organisations, government agencies, and cultural institutions

All Fields ▾

Search for Data

🔍 Search



Publicly accessible online

Advanced Search

Map Search

Explore



Themed Collections

Explore selected resources by theme



Services and Tools

Access data-related services and tools



Open Data

Find open data that is reusable



Grants and Projects

Search for research grants and projects



PARTICIPATE
DEPOSIT, JOIN

SEARCH
PUBLICATIONS, DATA, PROJECTS

STATISTICS
OA, PROJECTS, TOPICS

SUPPORT
FAQ, HELPDESK, GUIDES

OPEN ACCESS
IN EUROPE

Home Search Find

All	Publications	Data	Projects	People	Organizations	Datasources
-----	--------------	------	----------	--------	---------------	-------------

At a glance

Dataset Type

Unknown (611)

Dataset Language

English (101)
German (2)

Funder

European
Commission (4)

Funding Stream

Capacities (4)

Scientific Area

INFRA (4)

Publication Year

2013 (1)

Access Mode

UNKNOWN (608)
Open Access (3)

Datasource

Datacite (597)
Zenodo for Research
D... (14)

Filter

- allocator
- datacentre
- prefix
- resourceType
- contributor
- creator
- publicationYear
- publisher
- language

No active filters. Use the sidebar to filter search results.

10 documents found in 177ms

Page 1 of 1 

[Die hundertjährige Translationsfeier der beiden Wettinger Katakombenheiligen Marianus und Getulius](#) # 1

Getulius

 [doi:10.5169/SEALS-115655](#)

Felder, Peter

title: Die hundertjährige Translationsfeier der beiden Wettinger Katakombenheiligen **Marianus** und Getulius

[Integration eines Freihandzeichen-Tools in das Trainings- und Prüfungssystem CaseTrain](#) # 2

 [doi:10.3205/10CBT27](#) Text : ConferencePaper

Ifland, Marianus • Hörnlein, Alexander • Ott, Julian • Puppe, Frank

creator: Ifland, **Marianus**

[Konzeption und Evaluation eines fallbasierten Trainingssystems im universitätsweiten Einsatz \(CaseTrain\)](#) # 3

 [doi:10.3205/MIBE000086](#) Text : JournalArticle

Hörnlein, Alexander • Ifland, Marianus • Klügl, Peter • Puppe, Frank

creator: Ifland, **Marianus**

[Akzeptanz medizinischer Trainingsfälle als Ergänzung zu Vorlesungen](#) # 4

 [doi:10.3205/ZMA000754](#) Text : JournalArticle

Hörnlein, Alexander • Mandel, Alexander • Ifland, Marianus • Lüneburg, Edeltraud • Deckert, Jürgen • (et. al.)

creator: Ifland, **Marianus**

[Aufwandsanalyse für computerunterstützte Multiple-Choice Papierklausuren](#) # 5

 [doi:10.3205/ZMA000767](#) Text : JournalArticle

Mandel, Alexander • Hörnlein, Alexander • Ifland, Marianus • Lüneburg, Edeltraud • Deckert, Jürgen • (et. al.)

creator: Ifland, **Marianus**

[Die bildlichen Darstellungen im Atlas Marianus des Wilhelm Gumpenberg und eine Wallfahrtsbilderreihe in der Bischöflichen Sammlung Freiburg](#) # 6

 [doi:10.5169/SEALS-339455](#)

Ronner, Christel

title: Die bildlichen Darstellungen im Atlas **Marianus** des Wilhelm Gumpenberg und eine

[Data from: Contrasting insights provided by single and multispecies data in a regional comparative phylogeographic study](#) # 7

 [doi:10.5061/DRYAD.M7RC3](#) Dataset : DataPackage

DataCite Metadata Schema 3.1

XML Schema

Documentation

Examples

- [Full DataCite XML example](#)
- [Example for a simple dataset](#)
- [Example with complicated values](#)
- [Example with DataCollector as Contributor and a geoLocation box](#)
- [Example with GeoLocation](#)
- [Example with HasMetadata as related resource](#)
- [Example with IsIdenticalTo as related resource](#)
- [Example with ResearchGroup as Contributor and Methods as Description](#)
- [Example with Collection as ResourceType](#)
- [Example with Video as ResourceType](#)
- [Example for a workflow ResourceType](#)

DataCite Metadata Schema 3.1

📅 October 16, 2014

XML Schema:

<https://doi.org/10.5438/0011>

Documentation:

<https://doi.org/10.5438/0010>

DataCite Metadata Schema

Mandatory elements

- Creator
- Title
- Publication Year
- Publisher
- Identifier

Optional elements

- Contributor (with type)
- Date (with type)
- Alternate identifier
- Subject, Description
- Type, Format, Version
- Rights, Size, Language
- Geo-location
- Related identifiers

Metadata for Reuse

Documenting data for reuse

You will know you have documented your data well if

- should you need to repeat the work, you can
- a peer researcher can understand the data without consulting you
- a peer researcher can replicate/verify your results
- you can defend what you did to a reviewer
- in ten years' time, you can still understand the data

Think how you would explain to a new team member

- how you gathered/processed your data
- how you recorded/encoded your data
- which files are which

Basic level: unstructured metadata

Readme files: plain text documents with the following sections:

- Methodology
- Third-party inputs
- Workflow
- Outputs
- File structure and conventions

When ready to archive data, add the following:

- Inventory of files
- Citation information
- Licence information
- Relationships

```
#####  
# Methodologies  
  
Full details of the methods used to create the datasets are provided in the following references:  
* Pink, C.J., Swaminathan, S.K., Dunham, I., Rogers, J., Ward, R., et al. (2010). The mouse genome reference sequence. Nature Reviews Genetics 11: 575-582.  
* Pink, C.J. and Hurst, L.D. (2010). Timing of replication is a major determinant of gene length in the mouse genome. Genome Biology 11: R111.  
  
#####  
# Input Files  
  
Input files must be stored in a directory named Input_Files located in the root directory of the archive.  
* Mouse_Jul2007_EXin.tfa Mouse genome build m99 obtained from Ensembl  
* Rat_Nov2004_EXin.tfa Rat genome Assembly Nov 2004 (BGI) obtained from Ensembl  
* HMD_RatS.rpt Rat genome Assembly Nov 2004 (BGI) citation: Eppig, J. T., J. A. I. et al. (2004).  
  
#####  
# Scripts  
  
Two tcl script libraries are provided. For both, the main script is located in the root directory of the archive.  
* Scripts_With_selection_filter: contains scripts to create the datasets with a selection filter.  
* Scripts_No_selection_filter: contains scripts to create the datasets without a selection filter.  
  
The scripts are dependent on a working installation of the LAGI software.  
  
#####  
# Output Files  
  
Both sets of scripts create and store intermediate files to enable the user to track the progress of the analysis.  
The final datasets are created in tab-separated .txt files and for clarity of publication, these files have been renamed as follows:  
* mm_rn_Ki_RT_dataset_with_filter.txt  
* mm_rn_Ki_RT_dataset_no_filter.txt  
  
For preservation purposes .xlsx, and .csv versions of the data are also provided.  
  
#####  
# Dataset variables  
  
Ortholog Internal reference number to enable tracking of orthologous genes  
Mouse_Chromosome Mouse chromosomal location  
Rat_Chromosome Rat chromosomal location  
Mouse_RefSeq Mouse RefSeq ID  
Rat_RefSeq Rat RefSeq ID  
GT_SKEW Extent of GT skew in the mouse intron
```

Source: Pink, C., Hurst, L., Lercher, M., (2015). doi:10.15125/BATH-00091

Structured metadata for survey data

The accepted standard for survey data is **DDI**.

- DDI Codebook 2.5: <http://www.ddialliance.org/Specification/DDI-Codebook/2.5/>

Outline

- *Document description* (describing the record itself)
- *Study description*
 - **Citation information**
 - **Study information**: abstract, keywords, spatiotemporal coverage, unit of analysis
 - **Methodology/processing**: collection mode, sampling, weighting, cleaning, . . .
 - **Data access**: archive, conditions of access
 - **Other study description materials**: related resources

Structured metadata for survey data

Resources to help

- **HASSET** (Humanities And Social Science Electronic Thesaurus): <https://hasset.ukdataservice.ac.uk/>
- **Colectica Designer**:
<http://www.colectica.com/software/designer>
- **Easy DDI Organizer**: <http://csrda.iss.u-tokyo.ac.jp/international/ddi/edo/>,
<https://github.com/Easy-DDI-Organizer/EDO>
- **CED²AR** (Comprehensive Extensible Data Documentation and Access Repository):
<https://github.com/ncrncornell/ced2ar>

Structured metadata for sensor data

For describing sensors, there is **Sensor Model Language**:

<http://www.sensorml.com/>

- Minimal record provides a **description** and **identifier** of the sensor, its **output** (what it measures), and its **location**.
- Can extend it with **classifications**, **constraints**, **capabilities**, and links to further **documentation**

Resources to help

- **SWEET** (Semantic Web for Earth and Environmental Terminology): <http://sweet.jpl.nasa.gov/>
- **SensorML Ontology Registry and Repository**
<http://www.sensorml.com/ontologies.html>

Structured metadata for sensor data

MOLES 3 (Metadata Objects for Linking Environmental Sciences): <https://epubs.stfc.ac.uk/work/65066>

- Based on ISO 19156 Observations and Measurements:
<http://www.opengeospatial.org/standards/om>
- Compatible with ISO 19115
- Complex scheme with multiple entities:
 - Observation Collection, Observation, Observation Process
 - Observable Property, Feature of Interest
 - Acquisition, Computation, Result
 - Project, Utilities (citation information)

Structured metadata for photographs and models

For photographs of buildings, there is **VRA Core**:

<http://www.loc.gov/standards/vracore/>

Can be used for both the image and the real thing, e.g.

- agent (architect, photographer)
- dates (designed, built, taken)
- measurements
- technique
- worktype
- relation (e.g. image of . . .)

Specification recommends ontologies for certain elements, and has optional controlled vocabularies for others.

Structured metadata for photographs and models

For models, less clear separation between data and metadata:

- gbXML exchange format: <http://www.gbxml.org/>
- ISO 16739 Industry Foundation Classes:
<http://www.buildingsmart-tech.org/specifications/ifc-overview>

Enough to use standard data format and generic discovery metadata?

Conclusions

What metadata should you collect? Depends . . .

- Who will ask for it?
- What will they use it for?

When should you collect it?

- As soon as you can – don't wait until the end!

How should you collect it?

- Extract/copy, try to avoid retyping
- Keep central 'master' copies of metadata you will reuse across datasets

How should you record it?

- Use existing standards where you can – local profile?
- Look at the Metadata Standards Directory:
<http://rd-alliance.github.io/metadata-directory/>
- Be consistent
- If in doubt, use a README file – transform it later?

How should you record it?

- Use existing standards where you can – local profile?
- Look at the Metadata Standards Directory:
<http://rd-alliance.github.io/metadata-directory/>
- Be consistent
- If in doubt, use a README file – transform it later?

Thank you for listening

Questions?