Flying the flag

In support of metadata standards

Alex Ball

2017-05-25

University of Bath

Bloodbath at St Justin's

From the back cover of a crime novel:

> *Trapped in the crypt of St Justin's are five Justinians and the corpse of the Chaplain, murdered in stealth by one of them! . . . The Porter accuses the Bursar! The Bursar accuses the Principal, who in turn accuses the Butler! The Butler suspects the Dean or the Principal! The Dean claims the Principal is innocent! Which of these five theories is right? That would be telling! How many of them are right? To reveal even that would give the game away completely! . . .*

— logic puzzle by Bob Hargrave

Who killed the Chaplain?

	Theory					
Suspect	1	2	3	4	5	Total
Bursar	✓	✗	✗	✗	✓	2
Butler	✗	✗	✓	✗	✓	2
Dean	✗	✗	✗	✓	✓	2
Porter	✗	✗	✗	✗	✓	1
Principal	✗	✓	✗	✓	✗	2

Who killed the Chaplain?

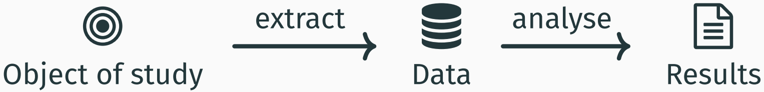
	Theory					
Suspect	1	2	3	4	5	Total
Bursar	✓	✗	✗	✗	✓	2
Butler	✗	✗	✓	✗	✓	2
Dean	✗	✗	✗	✓	✓	2
Porter	✗	✗	✗	✗	✓	**1**
Principal	✗	✓	✗	✓	✗	2

It’s surprising how powerful metadata can be.

Metadata

What is metadata?

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



Object of study

extract

Data

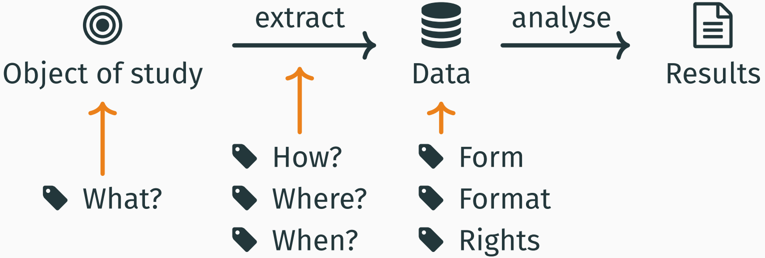
analyse

Results

- Context determines whether something is data or metadata

What is metadata?

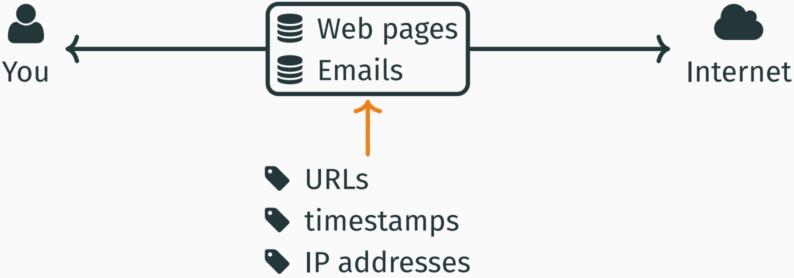
- Literally ‘data about data’
- Information that helps you work with other information



- Context determines whether something is data or metadata

Example: Internet traffic

Your perspective:



You

Web pages

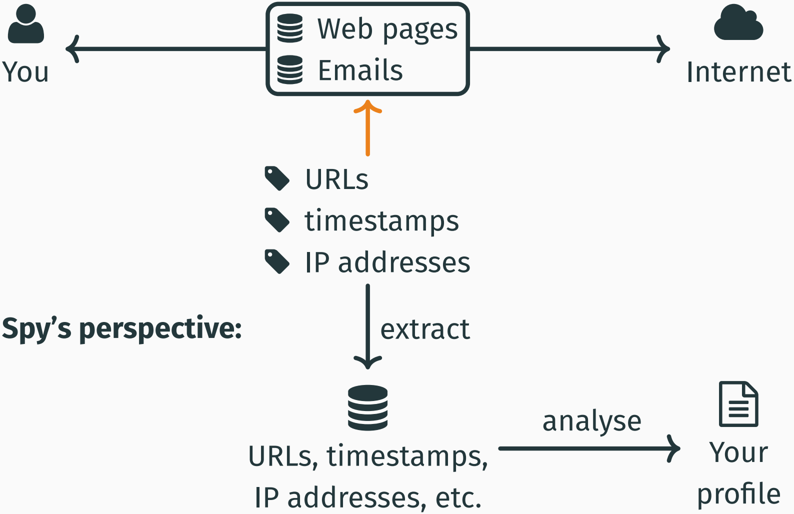
Emails

Internet

- URLs
- timestamps
- IP addresses

Example: Internet traffic

Your perspective:



You

Web pages

Emails

Internet

URLs

timestamps

IP addresses

extract

Spy’s perspective:

URLs, timestamps,
IP addresses, etc.

analyse

Your
profile

Types of metadata

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

Reference	Identifying, citing, searching for a known resource
Discovery	Speculative searching
Provenance	Assessing authenticity or trustworthiness
Contextual	Relating data to other resources, agents, activities
Rights	Securing data against unauthorized/illegal actions
Packaging	Arranging components of a resource
Fixity	Checking integrity
Structural	Loading/opening a file
Semantic	Unlocking the meaning of a resource

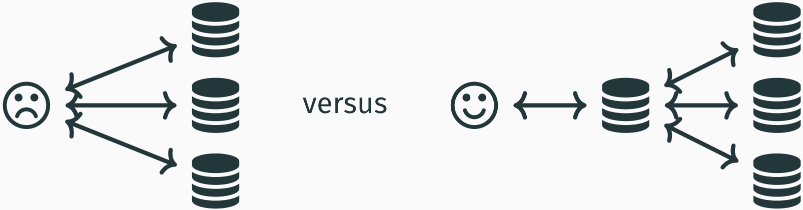
Types of metadata

In the research context, we are mostly concerned with

- Discovery metadata – help other researchers find the data, and give credit for them → impact
- Contextual metadata – keeping the institution and funder happy, conveying quality and relevance
- Structural & semantic metadata – ensure that researchers can understand and use/reuse the data

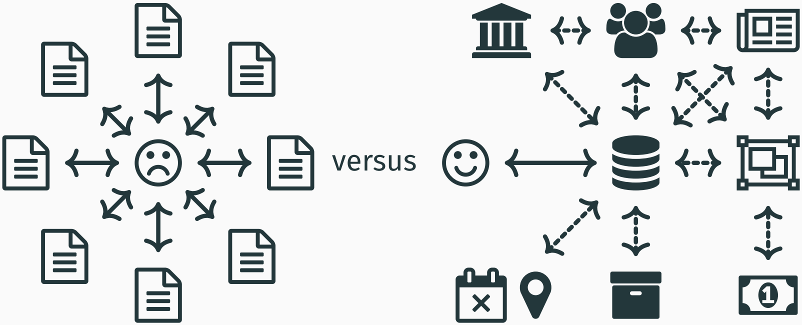
Why should I use a metadata standard?

Better discovery

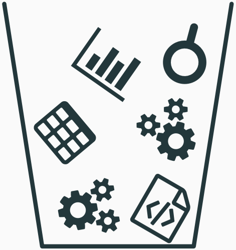


versus

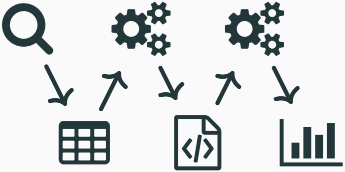
Better context



Better reuse



versus

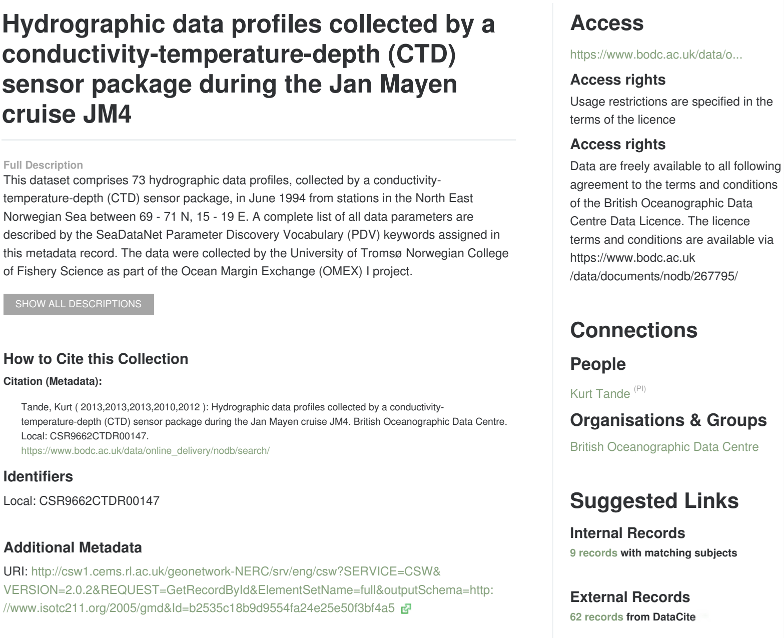


Better ecosystem

- Less working things out from scratch
- More complete metadata
- Benefits of practising
- Better documentation of the standards
- Concentration of development attention and effort
- Better time-saving tools
- etc., etc.

Research Data Discovery

UK Research Data Discovery Service (phase 1)



Hydrographic data profiles collected by a conductivity-temperature-depth (CTD) sensor package during the Jan Mayen cruise JM4

Full Description

This dataset comprises 73 hydrographic data profiles, collected by a conductivity-temperature-depth (CTD) sensor package, in June 1994 from stations in the North East Norwegian Sea between 69 - 71 N, 15 - 19 E. A complete list of all data parameters are described by the SeaDataNet Parameter Discovery Vocabulary (PDV) keywords assigned in this metadata record. The data were collected by the University of Tromsø Norwegian College of Fishery Science as part of the Ocean Margin Exchange (OMEX) I project.

SHOW ALL DESCRIPTIONS

How to Cite this Collection

Citation (Metadata):

Tande, Kurt (2013,2013,2013,2010,2012): Hydrographic data profiles collected by a conductivity-temperature-depth (CTD) sensor package during the Jan Mayen cruise JM4. British Oceanographic Data Centre. Local: CSR9662CTDR00147.
https://www.bodc.ac.uk/data/online_delivery/nodb/search/

Identifiers

Local: CSR9662CTDR00147

Additional Metadata

URI: http://csw1.cems.rl.ac.uk/geonetwork-NERC/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetRecordById&ElementSetName=full&outputSchema=http://www.isotc211.org/2005/gmd&Id=b2535c18b9d9554fa24e25e50f3bf4a5

Access

https://www.bodc.ac.uk/data/o...

Access rights

Usage restrictions are specified in the terms of the licence

Access rights

Data are freely available to all following agreement to the terms and conditions of the British Oceanographic Data Centre Data Licence. The licence terms and conditions are available via https://www.bodc.ac.uk/data/documents/nodb/267795/

Connections

People

Kurt Tande (PI)

Organisations & Groups

British Oceanographic Data Centre

Suggested Links

Internal Records

9 records with matching subjects

External Records

62 records from DataCite

Metadata requirements

Discovery metadata

- Search by title, description
- Search by subject, keywords
- Search by spatial coordinates

Contextual metadata

- Browse via researchers
- Browse via projects
- Browse via funders
- Links to semantic metadata

Collaborators

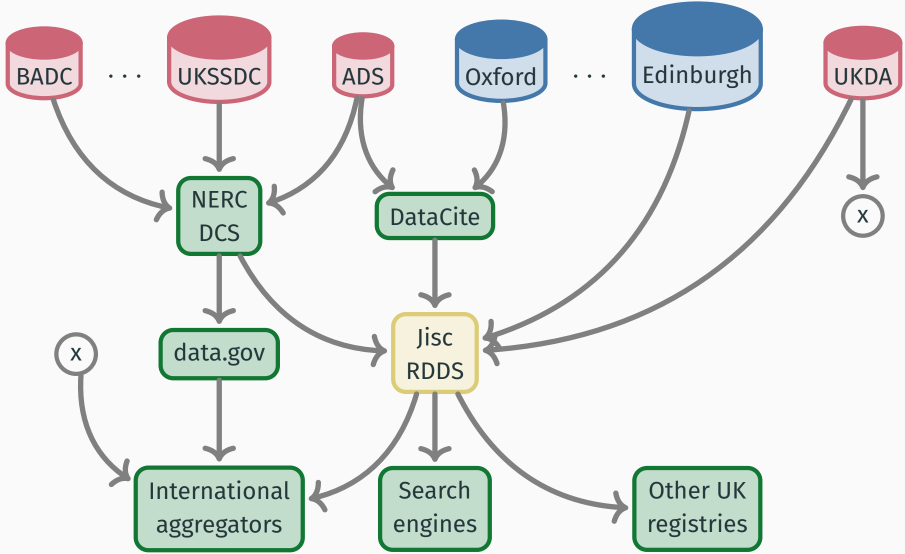
Data centres

- UK Data Archive
- NERC Data Catalogue Service
  - BADC
  - BODC
  - EIDC
  - NEODC
  - NGDC
  - PDC
  - UKSSDC
  - Archaeology Data Service

Universities

- Edinburgh
- Glasgow
- Hull
- Lincoln
- Leeds
- Oxford
- Oxford Brookes
- St Andrews
- Southampton

Metadata flows



RIF-CS data model

COLLECTION
repository

COLLECTION
registry

COLLECTION
collection

COLLECTION
dataset

COLLECTION
catalogueOrIndex

PARTY
group

PARTY
person

PARTY
administrative-Position

ACTIVITY
program

ACTIVITY
project

ACTIVITY
course

ACTIVITY
event

ACTIVITY
award

SERVICE
create

SERVICE
generate

SERVICE
transform

SERVICE
report

. . .

Metadata crosswalks

DDI Codebook 2.5

- UK Data Archive

DataCite 3

- Archaeology Data Service
- Oxford

EPrints 3/ReCollect

- Glasgow
- Leeds
- Southampton

MODS 3.5

- Edinburgh
- Hull
- St Andrews

OAI-PMH Dublin Core

- Oxford Brookes
- Lincoln

UK Gemini 2.2

- NERC Data Catalogue Service

Lessons learned

- We only wrote 6 crosswalks out of a possible 18
  - Standards cut our workload by a third!
  - Savings would have been greater on national rollout
- Could generate detailed records using even simple standards
  - For details, see Ball (2014)
- Problems mainly due to differences in data model:
  - Needed information on people, groups, projects: not much of this in metadata schemes designed for documents/datasets
  - Hard to infer personal identity without more information
  - Had to work with what we were given

DataCite Metadata Schema v4.0

Mandatory elements

- Creator
- Title
- Publication year
- Publisher
- Identifier
- Resource type

Recommended elements

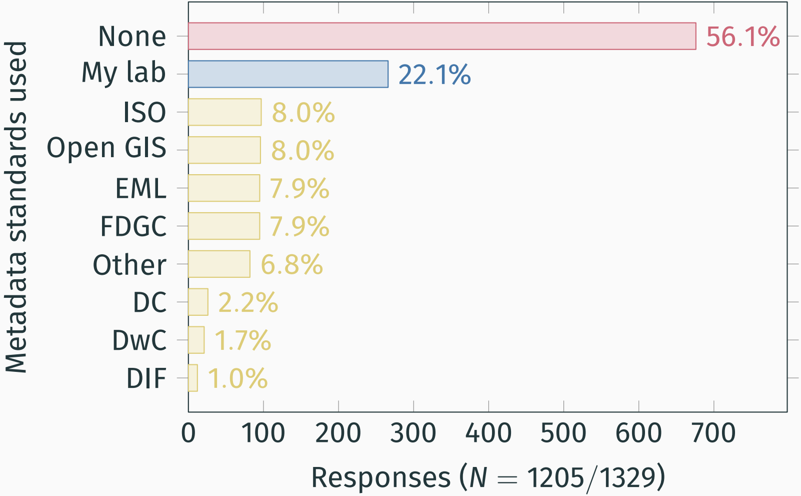
- Subject, Description
- Contributor (with type, affiliation)
- Date (with type)
- Geo-location
- Related identifiers

Optional elements

- Alternate identifier
- Format, Version, Size
- Rights, Language
- Funding reference

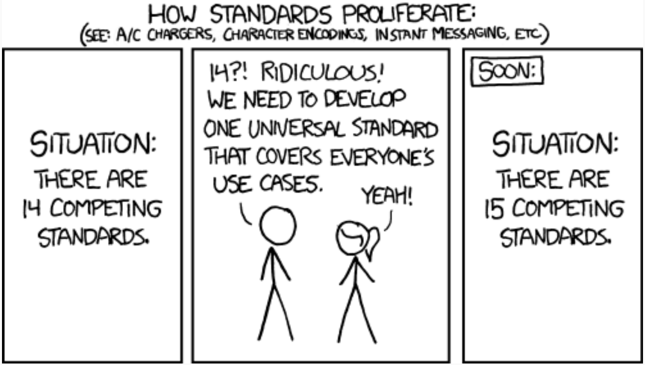
So why doesn’t everyone use a metadata standard?

No suitable standard?



Source: Tenopir, C. et al. (2011), ‘Data Sharing by Scientists: Practices and Perceptions’, *PLoS ONE* 6/6: e21101. doi: 10.1371/journal.pone.0021101

Too many standards?


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

Source: 🅭🅯🄏 Randall Munroe

‘The nice thing about standards is that you have so many to choose from’ — Tanenbaum (1988)

Isn’t that, like, really *hard*?

Just fill out this simple form . . .

```
<mods xmlns="http://www.loc.gov/mods/v3"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.loc.gov/mods/v3
http://www.loc.gov/standards/mods/v3/mods-3-4.xsd"> <titleInfo> <title> Title goes
here </title> </titleInfo> <name type="personal"> <namePart>Author name goes
here</namePart> <role> <roleTerm type="text">Author</roleTerm> </role> </name>
<typeOfResource>dataset</typeOfResource> <genre>Dataset</genre> <originInfo>
<publisher>Publisher name goes here</publisher> </originInfo> <language>
<languageTerm type="text">Language name</languageTerm> <languageTerm type="code"
authority="iso639-2b">ISO 639-2b code</languageTerm> </language>
<physicalDescription> <internetMediaType>MIME type goes here, repeat as
necessary</internetMediaType> <digitalOrigin>born digital</digitalOrigin>
<extent>Number of records in your database, or size of file in bytes</extent>
</physicalDescription> <abstract> Abstract goes here </abstract> <subject
authority="scheme name goes here"> <topic>Keyword goes here, repeat as
necessary</topic> <cartographics>Spatial coordinates<cartographics/>
<temporal>Temporal extente</temporal> <geographic>Spatial extent in
words</geographic> </subject> <identifier>ID goes here</identifier> <location> <url
usage="primary display" access="object in context">Location of record</url> <url
access="raw object">Location for download</url> </location> <accessCondition
type="useAndReproduction"> Usage restrictions or permissions </accessCondition>
<relatedItem ID="relatedMaterials"> <location> <url usage="primary display"
access="object in context">Record of related item</url> </location> </relatedItem>
<note type="citation"> Sample citation goes here </note> <note
type="software">Required software goes here</note> <subject ID="location"
displayLabel="Description of spatial extent again"> <cartographics> <coordinates>
List of coordinates, comma separated </coordinates> </cartographics> <topic>Type of
coordinates goes here</topic> </subject> </mods>
```

Metadata Standards Catalog

RDA Metadata Standards Directory WG

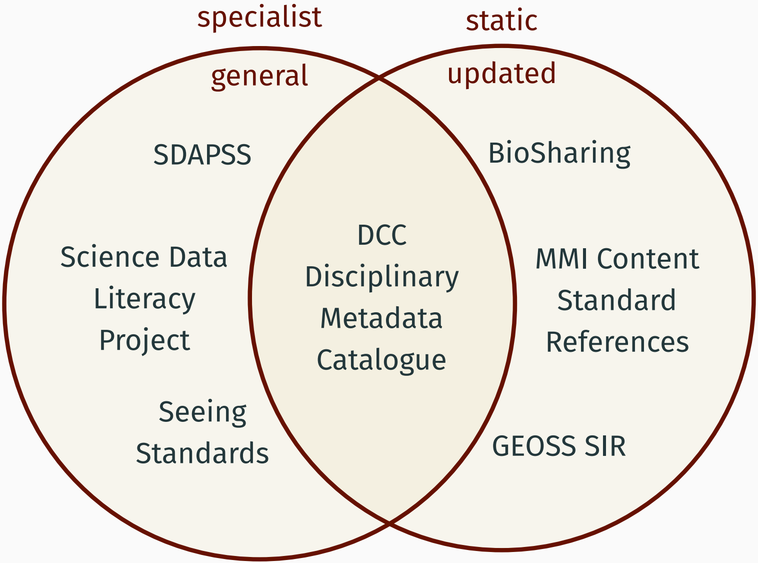
Key facts

- Ran 1 August 2013 – 1 February 2015
- 150 members from many countries and disciplines

Goals

1. Develop an RDA Metadata Standards Directory listing standards relevant for research data
   - Comprehensive
   - Easy for anyone to contribute or update
2. Define and develop use cases for research metadata
3. Develop plan for long-term growth and maintenance of the directory

Existing work



specialist

general

static

updated

SDAPSS

Science Data Literacy Project

Seeing Standards

DCC Disciplinary Metadata Catalogue

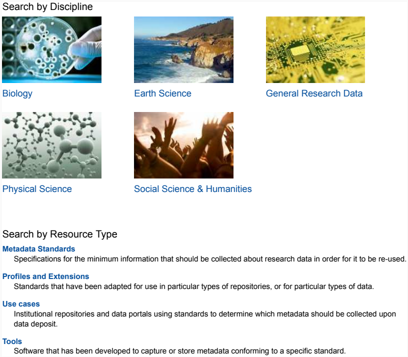
BioSharing

MMI Content Standard References

GEOSS SIR

The Metadata Standards Directory

DCC Disciplinary Metadata



Search by Discipline

Biology

Earth Science

General Research Data

Physical Science

Social Science & Humanities

Search by Resource Type

Metadata Standards
Specifications for the minimum information that should be collected about research data in order for it to be re-used.

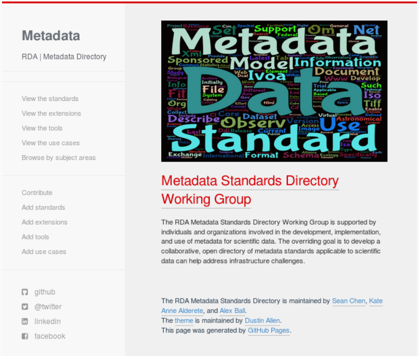
Profiles and Extensions
Standards that have been adapted for use in particular types of repositories, or for particular types of data.

Use cases
Institutional repositories and data portals using standards to determine which metadata should be collected upon data deposit.

Tools
Software that has been developed to capture or store metadata conforming to a specific standard.

http://www.dcc.ac.uk/resources/metadata-standards

RDA Metadata Standards Directory



Metadata

RDA | Metadata Directory

View the standards
View the extensions
View the tools
View the use cases
Browse by subject areas

Contribute
Add standards
Add extensions
Add tools
Add use cases

github
@twitter
linkedin
facebook

Metadata Standards Directory Working Group

The RDA Metadata Standards Directory Working Group is supported by individuals and organizations involved in the development, implementation, and use of metadata for scientific data. The overriding goal is to develop a collaborative, open directory of metadata standards applicable to scientific data can help address infrastructure challenges.

The RDA Metadata Standards Directory is maintained by Sean Chen, Kate Anne Alderete, and Alex Ball.
The theme is maintained by Dustin Allen.
This page was generated by GitHub Pages.

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

But there is more to be done . . .

- Search, not just browse
- Access data with machine-to-machine protocols
- Richer information
  - versions, mapping directionality, endorsements
  - greater use of entity relationships
- More services
  - Extracting what you need from compliant metadata . . .
  - Calculating migration pathways . . .
  - Comparing elements in different schemes . . .
  - Generating ‘first-pass’ converters . . .

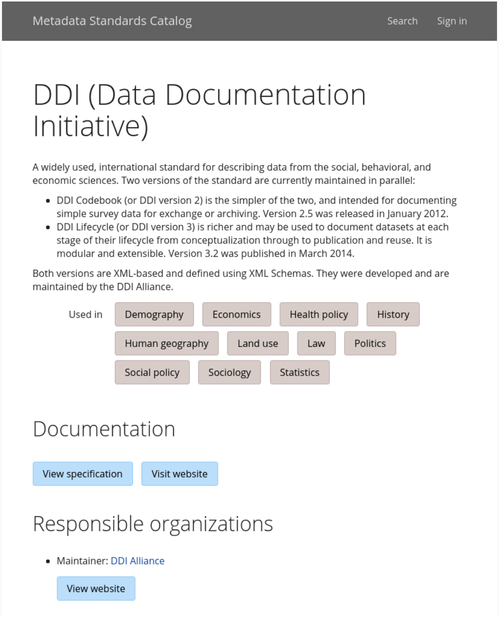
The Metadata Standards Catalog

Is this the right one for me?

- Name
- Description
- Research area
- Data type
- Maintainer, funder
- Endorsements

How do I use it?

- User guide
- Specification



DDI (Data Documentation Initiative)

A widely used, international standard for describing data from the social, behavioral, and economic sciences. Two versions of the standard are currently maintained in parallel:

- DDI Codebook (or DDI version 2) is the simpler of the two, and intended for documenting simple survey data for exchange or archiving. Version 2.5 was released in January 2012.
- DDI Lifecycle (or DDI version 3) is richer and may be used to document datasets at each stage of their lifecycle from conceptualization through to publication and reuse. It is modular and extensible. Version 3.2 was published in March 2014.

Both versions are XML-based and defined using XML Schemas. They were developed and are maintained by the DDI Alliance.

Used in: Demography, Economics, Health policy, History, Human geography, Land use, Law, Politics, Social policy, Sociology, Statistics

Documentation

Responsible organizations

- Maintainer: DDI Alliance

The Metadata Standards Catalog

How do I refer to it/find it again?

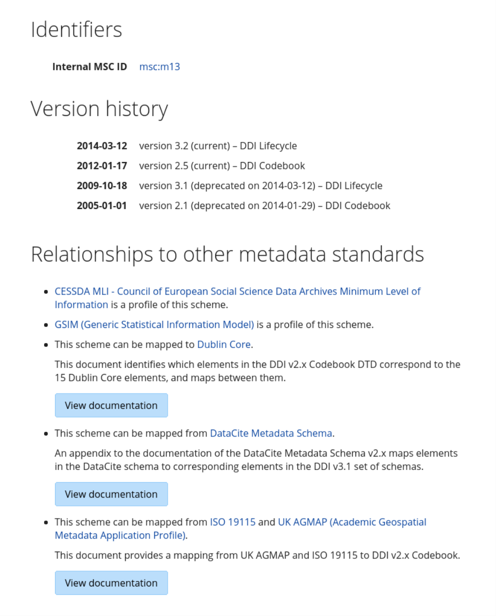
- Identifiers

Is this the right one for me?

- Version history
- Parent/child schemes

Can I convert existing metadata to it? Will I be locked in?

- Mappings to/from other schemes



Identifiers

Internal MSC ID	msc:m13

Version history

2014-03-12	version 3.2 (current) – DDI Lifecycle
2012-01-17	version 2.5 (current) – DDI Codebook
2009-10-18	version 3.1 (deprecated on 2014-03-12) – DDI Lifecycle
2005-01-01	version 2.1 (deprecated on 2014-01-29) – DDI Codebook

Relationships to other metadata standards

- CESSDA MLI - Council of European Social Science Data Archives Minimum Level of Information is a profile of this scheme.
- GSIM (Generic Statistical Information Model) is a profile of this scheme.
- This scheme can be mapped to Dublin Core.

  This document identifies which elements in the DDI v2.x Codebook DTD correspond to the 15 Dublin Core elements, and maps between them.

  View documentation
- This scheme can be mapped from DataCite Metadata Schema.

  An appendix to the documentation of the DataCite Metadata Schema v2.x maps elements in the DataCite schema to corresponding elements in the DDI v3.1 set of schemas.

  View documentation
- This scheme can be mapped from ISO 19115 and UK AGMAP (Academic Geospatial Metadata Application Profile).

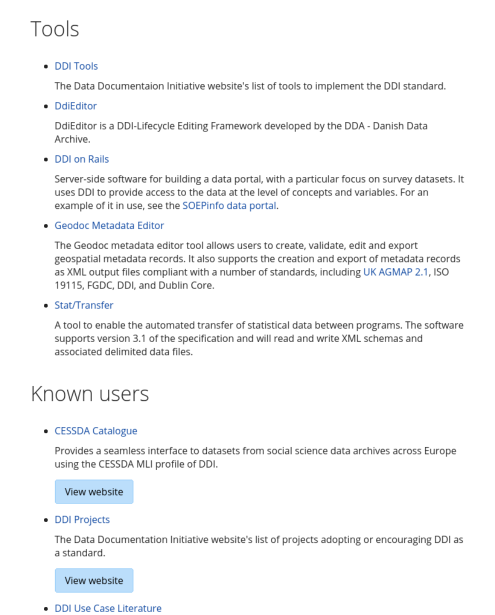
  This document provides a mapping from UK AGMAP and ISO 19115 to DDI v2.x Codebook.

  View documentation

The Metadata Standards Catalog

How do I use it?

- Software
- Services
- Known users
- Sample records



Tools

- DDI Tools

  The Data Documentaion Initiative website's list of tools to implement the DDI standard.

- DdiEditor

  DdiEditor is a DDI-Lifecycle Editing Framework developed by the DDA - Danish Data Archive.

- DDI on Rails

  Server-side software for building a data portal, with a particular focus on survey datasets. It uses DDI to provide access to the data at the level of concepts and variables. For an example of it in use, see the SOEPinfo data portal.

- Geodoc Metadata Editor

  The Geodoc metadata editor tool allows users to create, validate, edit and export geospatial metadata records. It also supports the creation and export of metadata records as XML output files compliant with a number of standards, including UK AGMAP 2.1, ISO 19115, FGDC, DDI, and Dublin Core.

- Stat/Transfer

  A tool to enable the automated transfer of statistical data between programs. The software supports version 3.1 of the specification and will read and write XML schemas and associated delimited data files.

Known users

- CESSDA Catalogue

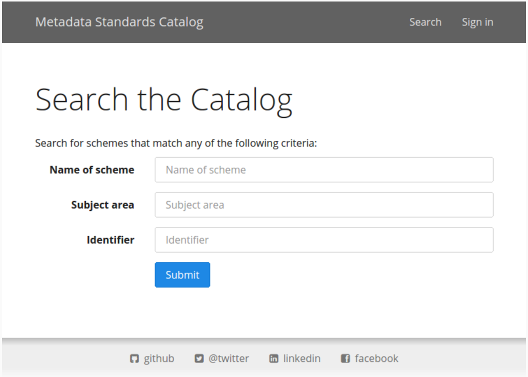
  Provides a seamless interface to datasets from social science data archives across Europe using the CESSDA MLI profile of DDI.

- DDI Projects

  The Data Documentation Initiative website's list of projects adopting or encouraging DDI as a standard.

- DDI Use Case Literature

The Metadata Standards Catalog



Metadata Standards Catalog

Search the Catalog

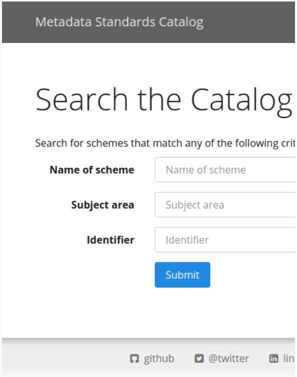
Search for schemes that match any of the following criteria:

Name of scheme: Name of scheme

Subject area: Subject area

Identifier: Identifier

The Metadata Standards Catalog



Metadata Standards Catalog

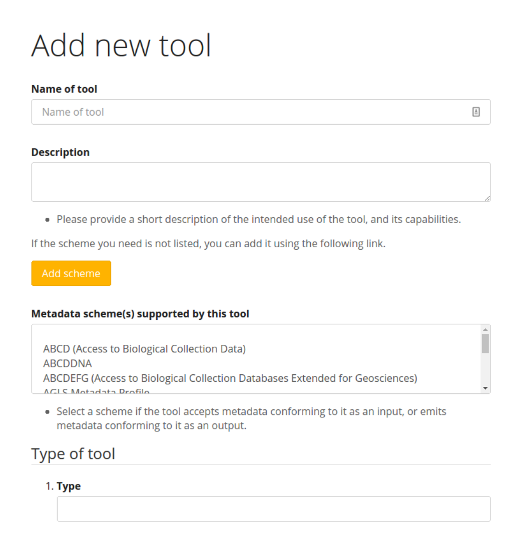
Search the Catalog

Search for schemes that match any of the following crit

Name of scheme: Name of scheme

Subject area: Subject area

Identifier: Identifier



Add new tool

Name of tool

Name of tool

Description

- Please provide a short description of the intended use of the tool, and its capabilities.

If the scheme you need is not listed, you can add it using the following link.

Metadata scheme(s) supported by this tool

ABCD (Access to Biological Collection Data)
ABCDDNA
ABCDEFG (Access to Biological Collection Databases Extended for Geosciences)
AGLS Metadata Profile

- Select a scheme if the tool accepts metadata conforming to it as an input, or emits metadata conforming to it as an output.

Type of tool

1. **Type**

Future developments

- GUI Highlight standards bodies
- GUI Dynamic filtering while browsing
- GUI Side-by-side specifications
- GUI Version history as timeline
- GUI Search by article DOI
- GUI Show maturity rating for schemes

- API Make changes to database via API
- API Query standards by their elements
- API Query by element value encoding
- API Query by article DOI
- API Calculate crosswalks

https://www.rd-alliance.org/groups/metadata-standards-catalog-working-group.html

Canonical metadata packages

Dataset

Unique Identifier
Name/title
Description
Keywords
Spatial coordinates
Temporal coordinates
Location (e.g. URL)
Medium/format
Availability (e.g. licence)
Schema
Quality
Provenance

Person

Originator

Activity

Project

Related publications
Related software
Citations

Facility
Equipment

Unpacking the elements

Example: spatial coordinates

- X, Y, Z in declared coordinate system
  - May be connected with temporal coordinate
- Precision
- Accuracy
- Resolution

Need to unpack all elements and validate the result

- Join in: https://www.rd-alliance.org/groups/metadata-ig.html
- Hope to publish as an RDA output
- Basis for converters?

Call to action

Metadata → better data

- Even bad documentation is better than nothing
- The more structure, the better
  - Clear headings and sections in documentation
  - Consistent metadata
- Look for metadata standards you can use
  - Metadata Standards Directory/Catalog
- Not an exact fit? Create a local profile
  - Avoid completely bespoke schemes
- Be consistent

Thank you for listening

Grazie per l’attenzione

Any questions?

References

- Ball, A. (2014), *UK Research Data Registry Mapping Schemes*, version 09 (Edinburgh, UK: Digital Curation Centre, 9 May), http://www.dcc.ac.uk/sites/default/files/documents/registry/uk-rdr-mapping-v09.pdf.
- DataCite Metadata Working Group (2016), *DataCite Metadata Schema for the Publication and Citation of Research Data*, version 4.0 (DataCite e.V.). doi: 10.5438/0012.
- Tanenbaum, A. S. (1988), *Computer Networks*, (2nd edn., Upper Saddle River, NJ: Prentice-Hall).
- Tenopir, C. et al. (2011), 'Data Sharing by Scientists: Practices and Perceptions', *PLoS ONE* 6/6: e21101. doi: 10.1371/journal.pone.0021101.