Flying the flag
In support of metadata standards

Alex Ball
2017-05-25

University of Bath
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?

<table>
<thead>
<tr>
<th>Suspect</th>
<th>Theory</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bursar</td>
<td>✔️</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✔️</td>
<td>2</td>
</tr>
<tr>
<td>Butler</td>
<td>✗</td>
<td>✗</td>
<td>✔️</td>
<td>✗</td>
<td>✗</td>
<td>✔️</td>
<td>2</td>
</tr>
<tr>
<td>Dean</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>2</td>
</tr>
<tr>
<td>Porter</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✔️</td>
<td>1</td>
</tr>
<tr>
<td>Principal</td>
<td>✗</td>
<td>✔️</td>
<td>✗</td>
<td>✗</td>
<td>✔️</td>
<td>✗</td>
<td>2</td>
</tr>
</tbody>
</table>

It's surprising how powerful metadata can be.
## Who killed the Chaplain?

<table>
<thead>
<tr>
<th>Suspect</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bursar</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>2</td>
</tr>
<tr>
<td>Butler</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>2</td>
</tr>
<tr>
<td>Dean</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>2</td>
</tr>
<tr>
<td><strong>Porter</strong></td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>1</td>
</tr>
<tr>
<td>Principal</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>2</td>
</tr>
</tbody>
</table>

It’s surprising how powerful metadata can be.
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:

You → Web pages → Internet

↓

URLs
timestamps
IP addresses

Your
profile
extract
analyse
Spy's perspective:
Example: Internet traffic

Your perspective:

You → Web pages, Emails → Internet

Spy’s perspective:

Internet → extract → 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
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.

How to Cite this Collection
Citation (Metadata):
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)
British Oceanographic Data Centre
9 records

Organisations & Groups
British Oceanographic Data Centre
62 records from DataCite

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

BADC → UKSSDC → ADS → Oxford → Edinburgh → UKDA

NERC DCS

DataCite

Jisc RDDS

data.gov

International aggregators

Search engines

Other UK registries

International aggregators

Other UK registries
RIF-CS data model

COLLECTION
- repository
- registry
- collection
- dataset
- catalogueOrIndex

PARTY
- group
- person
- administrative-Position

ACTIVITY
- program
- project
- course
- event
- award

SERVICE
- create
- generate
- transform
- 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?

Metadata standards used

- None: 56.1%
- My lab: 22.1%
- ISO: 8.0%
- Open GIS: 8.0%
- EML: 7.9%
- FDGC: 7.9%
- Other: 6.8%
- DC: 2.2%
- DwC: 1.7%
- DIF: 1.0%

Responses (N = 1205/1329)

Too many standards?

‘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 . . .

```xml
<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>
  </physicalDescription>
  <extent>Number of records in your database, or size of file in bytes</extent>
  <abstract>Abstract goes here</abstract>
  <subject authority="scheme name goes here">
    <topic>Keyword goes here, repeat as necessary</topic>
    <cartographics>Spatial coordinates</cartographics>
  </subject>
  <temporal>Temporal extent</temporal>
  <geographic>Spatial extent in words</geographic>
  <identifier>ID goes here</identifier>
  <location>
    <url usage="primary display" access="object in context">Location of record</url>
    <url usage="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>
  </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

- SDAPSS
- Science Data Literacy Project
- Seeing Standards
- DCC Disciplinary Metadata Catalogue
- BioSharing
- MMI Content Standard References
- GEOSS SIR

Specialist vs. General

Updated vs. Static
While data curators, and increasingly researchers, know that good metadata is key for research data access and re-use, figuring out precisely what metadata to capture and how to capture it is a complex task. Fortunately, many academic disciplines have supported initiatives to formalise the metadata specifications the community deems to be required for data re-use. This page provides links to information about these disciplinary metadata standards, including profiles, tools to implement the standards, and use cases of data repositories currently implementing them.

For those disciplines that have not yet settled on a metadata standard, and for those repositories that work with data across disciplines, the General Research Data section links to information about broader metadata standards that have been adapted to suit the needs of research data.
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 . . .
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.

Documentation

- View specification
- Visit website

Responsible organizations

- Maintainer: DDI Alliance

View website
How do I refer to it/find it again?

- Identifiers
- Version history
- Parent/child schemes

Is this the right one for me?

- Identifiers
- 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: mscm13

**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
How do I use it?

- Software
- Services
- Known users
- Sample records

Tools

- **DDI Tools**
  The Data Documentation 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 [SOEPlinfo 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.

  View website

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

  View website

- **DDI Use Case Literature**
Search the Catalog

Search for schemes that match any of the following criteria:

Name of scheme

Subject area

Identifier

Submit
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.

Add scheme

Metadata scheme(s) supported by this tool

• 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

- Highlight standards bodies
- Dynamic filtering while browsing
- Side-by-side specifications
- Version history as timeline
- Search by article DOI
- Show maturity rating for schemes
- Make changes to database via API
- Query standards by their elements
- Query by element value encoding
- Query by article DOI
- Calculate crosswalks

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique Identifier</td>
<td>Originator</td>
</tr>
<tr>
<td>Name/title</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>Keywords</td>
<td></td>
</tr>
<tr>
<td>Spatial coordinates</td>
<td>Activity</td>
</tr>
<tr>
<td>Temporal coordinates</td>
<td></td>
</tr>
<tr>
<td>Location (e.g. URL)</td>
<td></td>
</tr>
<tr>
<td>Medium/format</td>
<td></td>
</tr>
<tr>
<td>Availability (e.g. licence)</td>
<td>Related publications</td>
</tr>
<tr>
<td>Schema</td>
<td></td>
</tr>
<tr>
<td>Quality</td>
<td>Related software</td>
</tr>
<tr>
<td>Provenance</td>
<td>Citations</td>
</tr>
<tr>
<td>Facility</td>
<td></td>
</tr>
<tr>
<td>Equipment</td>
<td></td>
</tr>
</tbody>
</table>
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
• 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?

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.
