Metadata

Interest Group

RESEARCH DATA ALLIANCE

Towards a metadata Rosetta Stone

The RDA MIG Metadata Element Set

Alex Ball

University of Bath

IG RDMinEng Seminar: Building Metadata Standards Within Engineering Disciplines and Communities

9 October 2020

Principles

Metadata are essential for FAIR data

Findable

1. Globally unique, persistent ID
2. Rich metadata
3. Metadata indexed in a catalogue
4. Data ID present in metadata

Accessible

1. Retrieval by ID
   - 1.1 using open, free protocol
   - 1.2 supporting access controls
2. Metadata always accessible

Interoperable

1. Standard, formal representation
2. Vocabularies also FAIR
3. Qualified links to other meta/data

Re-usable

1. Accurate, relevant attributes
   - 1.1 clear licence
   - 1.2 provenance information
   - 1.3 meet community standards

RDA Metadata Principles and their Use

Extracts from Jeffery and Koskela (2014):

1. > The only difference between metadata and data is mode of use.

   Metadata denotes a **role** played, not a type of **thing**.

2. > Metadata is not just for data, it is also for users, software services, computing resources.

   To describe **data**, you also have to describe **other things**.

3. > Metadata is not just for description and discovery; it is also for contextualisation (relevance, quality, restrictions [rights, costs]) and for coupling users, software and computing resources to data (to provide a Virtual Research Environment).

   Different types of metadata support different **tasks**.

RDA Metadata Principles and their Use

4. > Metadata must be machine-understandable as well as human understandable for autonomicity (formalism).

More **formalism** mean less human **intervention** required.

5. > Management (meta)data is also relevant (research proposal, funding, project information, research outputs, outcomes, impact . . .).

. . . for assessing quality, relevance, integrity, compliance . . .

RDA Metadata Principles and their Use

4. > Metadata must be machine-understandable as well as human understandable for autonomicity (formalism).

   More **formalism** mean less human **intervention** required.

5. > Management (meta)data is also relevant (research proposal, funding, project information, research outputs, outcomes, impact . . .).

   . . . for assessing quality, relevance, integrity, compliance . . .

Vision:

- For given set of use cases, define **metadata package**
- **Map** between each metadata scheme and the package
- Use mappings to autogenerate **converters**; finesse by hand

Elements

Work towards metadata packages

1. Data in Context Interest Group led collection of **metadata use cases**
2. Metadata Interest Group **analysed elements** used
   - Analysis spreadsheet and explanatory slides: `https://rd-alliance.org/use-case-analysis.html`
3. Metadata Element Set debated at RDA Plenary meetings
4. Now in the process of ‘unpacking’ the elements

Recommended Metadata Element Set

Dataset

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

Person/Organization

- Originator

Activity

- Project

- Related publications
- Related software
- Citations

- Facility
- Equipment

Work towards metadata packages

1. Data in Context Interest Group led collection of **metadata use cases**
2. Metadata Interest Group **analysed elements** used
   - Analysis spreadsheet and explanatory slides: `https://rd-alliance.org/use-case-analysis.html`
3. Metadata Element Set debated at RDA Plenary meetings
4. Now in the process of ‘unpacking’ the elements

Unpacking the Element Set

Google Drive folder for MIG Metadata Element Set discussion:

- Unique Identifier
- Description
- Keywords
- Spatial coordinates
- Temporal coordinates
- Location
- Medium/format
- Availability
- Schema
- Quality
- Provenance
- Originator
- Project
- Related publications
- Related software
- Citations
- Facility/equipment
- Gaps to consider

Element set priorities

Survey with n = 37 (ordered by mean; median only differs where shown)

1. Unique Identifier
2. Description
3. Location
4. Originator
5. Keywords
6. Availability
7. Temporal coordinates
8. Spatial coordinates
9. Medium/format
10. Provenance (2↓)
11. Quality
12. Project
13. Schema
14. Citations
15. Related publications
16. Facility/equipment (2↑)
17. Related software

Gaps to consider Repository name (data publisher);
Title (main, alternative, abbreviated);
Methodology; Sampling procedure.

Example: unpacking Unique Identifier

Semantics:

- One or more strings that can be used to identify the resource
- Belong to a **scheme** (which may have a resolver/bridge)
- May need to be qualified by
  - role/purpose for which this ID is used
  - provenance (e.g. who coined it and when)
  - scope (e.g. version, granule)
- Should be
  - universally unique
  - permanent/persistent
  - unstructured
  - resolvable but not intrinsically an address
  - otherwise meaningless

Example: unpacking Unique Identifier

Syntax:

- A ‘base’, context-agnostic identifier

  Scheme Specified using a controlled vocabulary

  Value The identifier string itself

- Context-specific identifiers

  Scheme Specified using a controlled vocabulary

  Value The identifier string itself

  . . . and other qualifiers as required (TBD)

Example: unpacking Description

Semantics:

- Portion of free text describing the resource
- Written in a particular **language**
- Plays a **role** or describes a given **aspect**:
  - Title
  - Abstract
  - Methodology
  - Unstructured manifest of contents
  - Technical information
  - Note
- Will be **encoded** in given way

Example: unpacking Description

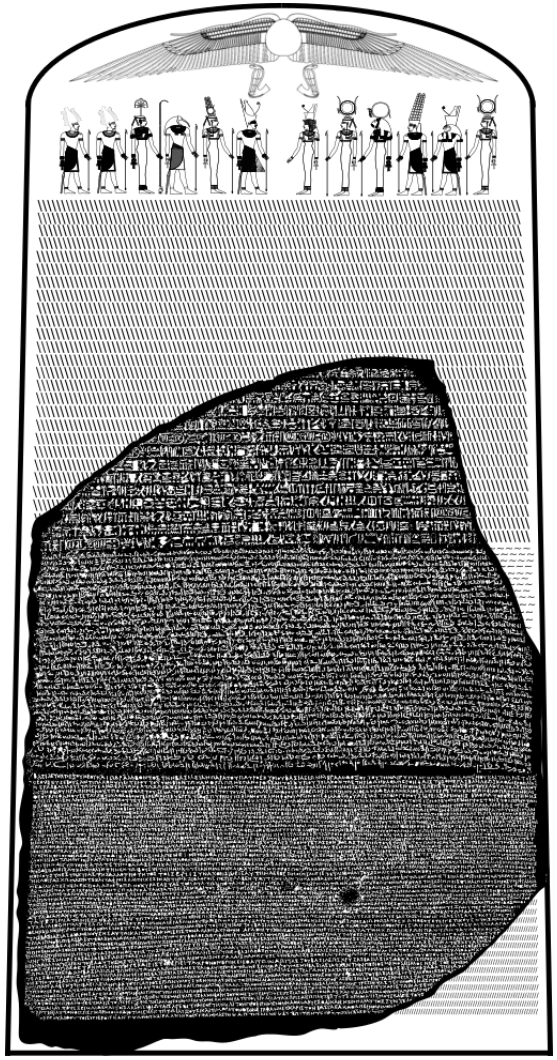
Syntax:

Type	Specified using a controlled vocabulary (`title`, `abstract`...)
Language	Specified as a language code conformant with IETF BCP 47
Format	Specified using a controlled vocabulary (`plain`, `html5`...)
Value	Long string (i.e. may contain line break characters)

End Goal

Ambition for the Metadata Element Set

- Starting point for developing new domain standards
- Tool for analysing existing metadata schemes
- ‘Rosetta Stone’ for interconverting between arbitrary standards



The Metadata Standards Catalog

https://rdamsc.bath.ac.uk/

Is this the right one for me?

How do I use it?

How do I refer to it/find it again?

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

Metadata Standards Catalog Search Sign in

CSMD (Core Scientific Metadata Model)

A study-data oriented model, primarily in support of the ICAT data managment infrastructure software. The CSMD is designed to support data collected within a large-scale facility's scientific workflow; however the model is also designed to be generic across scientific disciplines.

Sponsored by the Science and Technologies Facilities Council, the latest full specification available is v 4.0, from 2013.

Used in Biochemistry Chemistry Crystallography Materials engineering

Documentation

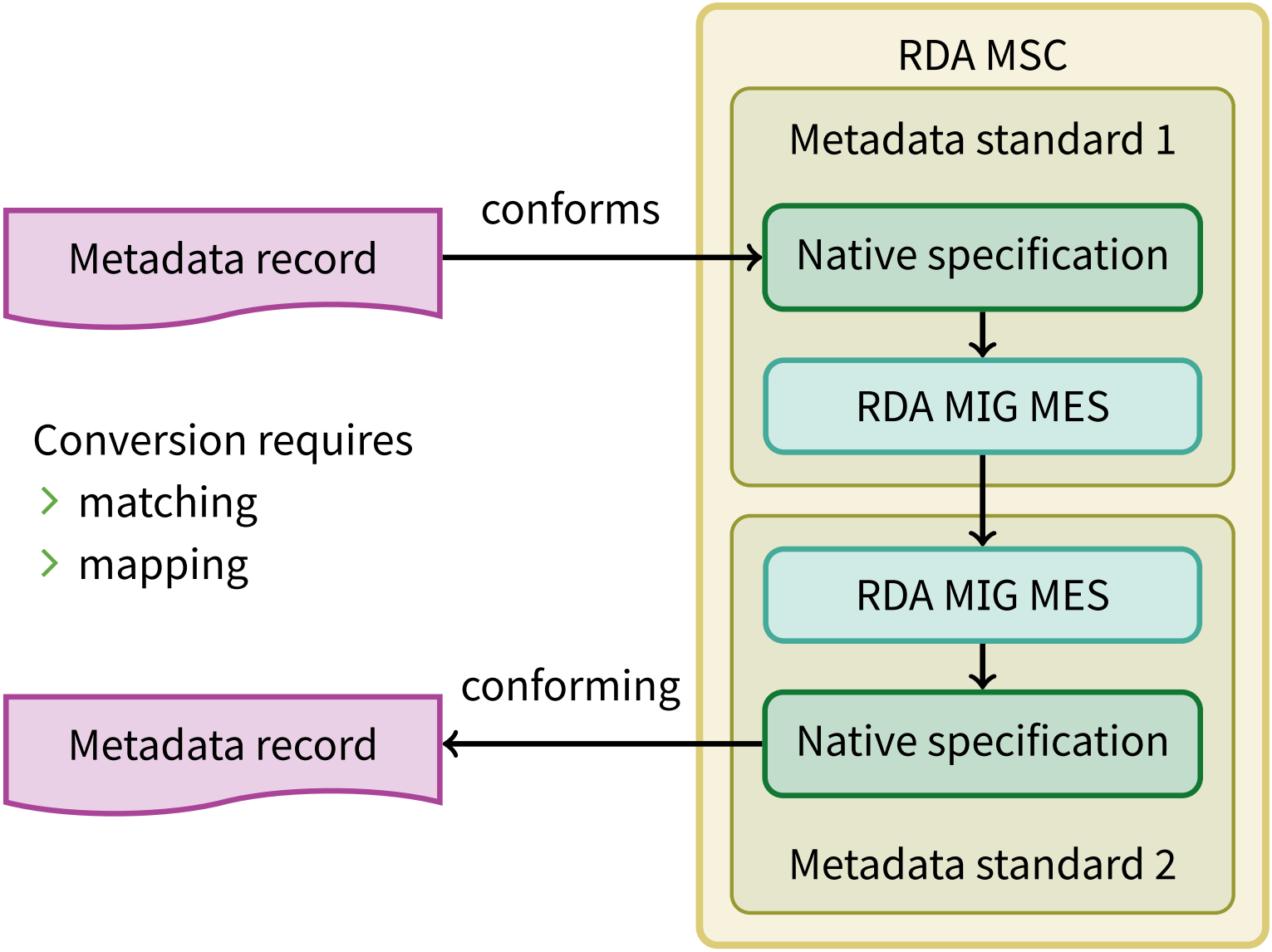
Visit website

Identifiers

Internal MSC ID msc:m8

Relationships to other metadata standards

Bringing it all together

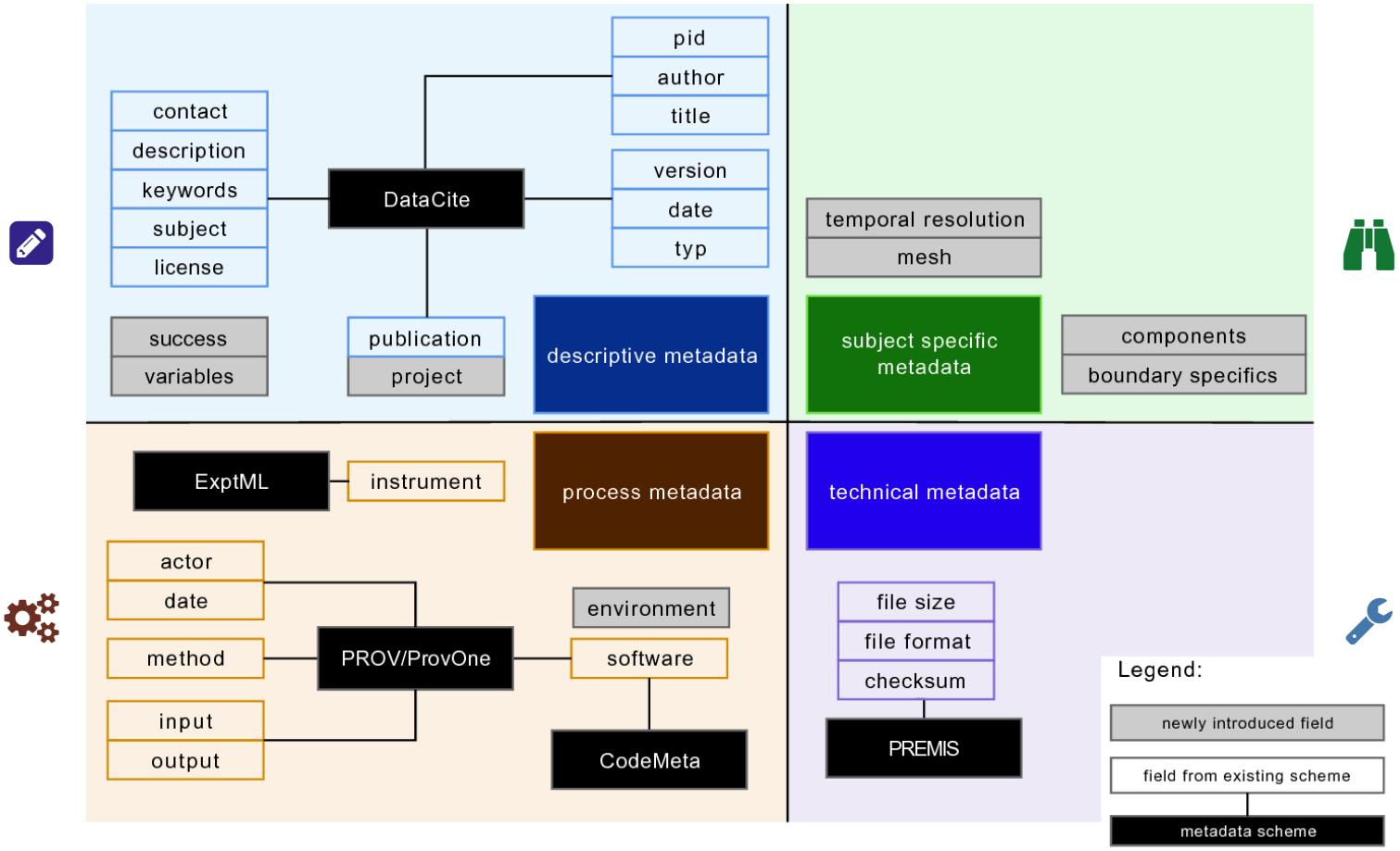


Conversion requires

- matching
- mapping

Worked Example

EngMeta: Metadata for computational engineering



Schembera and Iglezakis (2019)

EngMeta in terms of RMES

Unique Identifier

- (P)ID
- Associated IDs

Description

- Title
- Description
- Data generation method
- Further information

Location

- File name/location

Keywords

- Keywords
- Subject
- Data type
- Object of research

Spatial coordinates

- Spatial resolution

Temporal coordinates

- Temporal resolution

EngMeta in terms of RMES

Schema

- Measured variables
- Controlled variables
- Phases
- Components
- Parameters
- Definition of boundaries

Quality

- Indication of success
- Explanation of failure

Medium/format

- File type

Availability

- Legal information

Originator

- Contact person
- Producer/author
- Contributor

EngMeta in terms of RMES

Provenance

- Dates
- Version
- Provenance
- Processing step

Facility/equipment

- Processing step

Project

- Project
- Funding information

Related publications

- Publication

Related software

- Associated resources

Citations

- Context

Gaps to consider

- File size
- Checksum

RESEARCH DATA ALLIANCE

Thank you for your attention

Metadata Interest Group:

https://www.rd-alliance.org/groups/metadata-ig.html

References

- Jeffery, K. and Koskela, R. (2014), *RDA Metadata Principles and their Use*, (Research Data Alliance, 14 Nov.), https://www.rd-alliance.org/metadata-principles-and-their-use.html, accessed 5 Oct. 2020.
- Schembera, B. and Iglezakis, D. (2019), 'The Genesis of EngMeta: A Metadata Model for Research Data in Computational Engineering', in *Metadata and Semantic Research* (Communications in Computer and Information Science, 846; Cham: Springer), 127–32. doi: 10.1007/978-3-030-14401-2_12.