Metadata

Fundamentals of Data Management

Adam Carter
Project Manager, EPCC
A.Carter@epcc.ed.ac.uk
+44 131 650 6009

Lecture Overview

- This lecture will focus on **Metadata**
- We'll also introduce the ideas of **Semantics** and **Ontologies**
  - which we'll build upon more in a later lecture about the Semantic Web

- What Is Metadata?
- Where Is Metadata Found?
- Why Should Metadata Be Used?

Metadata is…

- Data about data
  - One person’s metadata is another person’s data
- Information that makes data useful
- A description of what the data is, for use by others, and by you when you’ve forgotten
  - And more and more, by automated systems
- A means to *understand* the data, and possibly *reproduce* the data (or make an equivalent observation)

Kinds of Metadata

- System Metadata / Structural Metadata
  - File ownership, modification date, how it’s packaged, etc.
- “Content Metadata” / “Descriptive Metadata”
  - What the data relates to
  - Where the data relates to
  - When the data relates to
  - Who the data relates to
  - How the data were collected / created
  - Why the data were collected / created
  - Who collected /created the data
  - When the data was collected / created
  - Where the data were collected
  - …

Metadata Categorisation

- Structural/Control Metadata and Guide Metadata
  - Bretherton & Singley – 1994
  - *doi:10.1109/SSDM.1994.336950*
- Technical, Business and Process
  - Ralph Kimball
  - *urn:isbn:978-0-470-14977-5*
- Descriptive, Structural and Administrative
  - National Information Standards Organisation
  - *urn:isbn:1-880124-62-9*

Where is the metadata?

- Sometimes it’s embedded alongside the data
- Sometimes it’s in metadata files, indexes and catalogues

Semantics

- The *meaning* of the data
  - and how we convey this in the data and its metadata
- E.g. a date in a file might mean
  - The date that the data describes
  - The date that the data was stored
  - That the data pertains to some point in time during the day
  - That the data is an average over the day
  - That the first data point in the data set relates to a time on the stated day
- Concepts described in data or metadata might have specific meanings that should be exactly defined
  - Does “rain” include “sleet”? What about hail?

Ontology

- A controlled vocabulary
  - A means to describe semantics
  - Precise definitions for a set of terms
  - Can be used in metadata and the data itself
- c.f. “Folksonomy”
  - Tagging
  - Uncontrolled
  - Responsive, Dynamic
  - #EUDAT #RDA

Ontology versus Vocabulary

“There is no clear division between what is referred to as “vocabularies” and “ontologies”. The trend is to use the word “ontology” for more complex, and possibly quite formal collection of terms, whereas “vocabulary” is used when such strict formalism is not necessarily used or only in a very loose sense. Vocabularies are the basic building blocks for inference techniques on the Semantic Web”

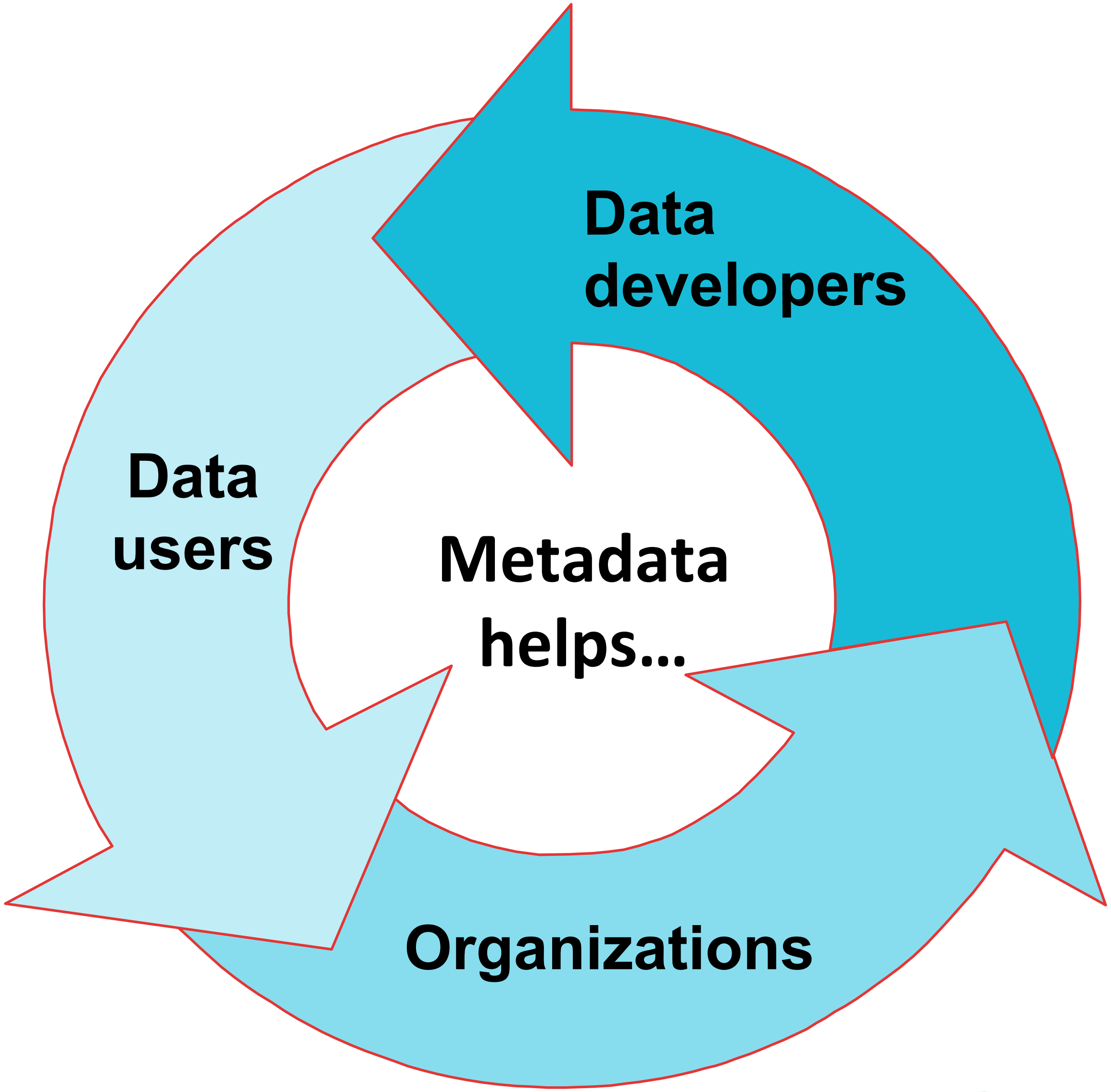
from http://www.w3.org/standards/semanticweb/ontology

[Accessed: 2014-03-24]

Why should you use metadata?

- It can make your data more discoverable
  - People can search on the metadata
- It can make your data more reusable
  - …because it's understandable
  - Reusable for the "same" purpose (e.g. to aid validation of the results), and potentially others
  - Facilitates finding related data
- It makes your data more reproducible
  - If you know how/why/where it was collected, it helps others to reproduce your research/experiment in order to validate it

The Value of Metadata



Data developers

Data users

Metadata helps…

Organizations

What makes good metadata

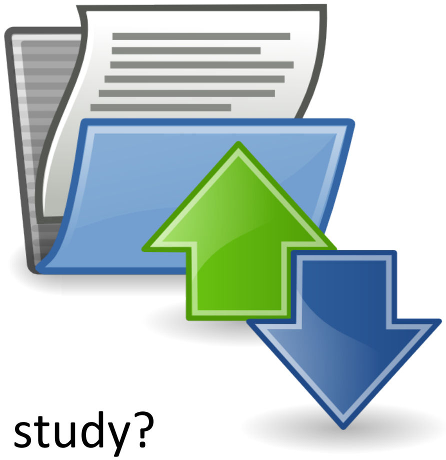
- Metadata is good if it allows your data to be found and understood by all those who might want to make use of it
- Complete
- Accurate
- Precise
- Conforming to standards
  - Semantic: Meaning of Terms
  - Which metadata are mandatory
  - Formatting / Syntax
- Accessible
  - Online, addressable (can be linked to), harvestable

Working with Data

- When you ***provide*** data to someone else, what types of information would you want to include with the data?

- When you ***receive*** a dataset from an external source, what types of details do you want to know about the data?

Working with Data



- **Providing data**:
  - Why were the data created?
  - What limitations, if any, do the data have?
  - What does the data mean?
  - How should the data be cited if it is re-used in a new study?

- **Receiving data**:
  - What are the data gaps?
  - What processes were used for creating the data?
  - Are there any fees associated with the data?
  - In what scale were the data created?
  - What do the values in the tables mean?
  - What software do I need in order to read the data?
  - What projection are the data in?
  - Can I give these data to someone else?

Who should create metadata?

- Ideally the same person/people who created the data.
  - They understand it best!
- Sometimes those responsible for the data's distribution and curation are well-placed to add additional metadata
  - particularly structural metadata

Important Metadata Standards

- There are many standards available to document data. Each has a different focus, yet ask for similar information about the data set.
- Your choice will depend on:
  - your field of practice
  - your motivation for using metadata
- Dublin Core Metadata Initiative
  - DCMI Metadata Terms
  - Dublin Core Metadata Element Set
- Metadata Encoding and Transmission Standard (METS)
- OAI-PMH – A metadata harvesting standard

Important Metadata Standards

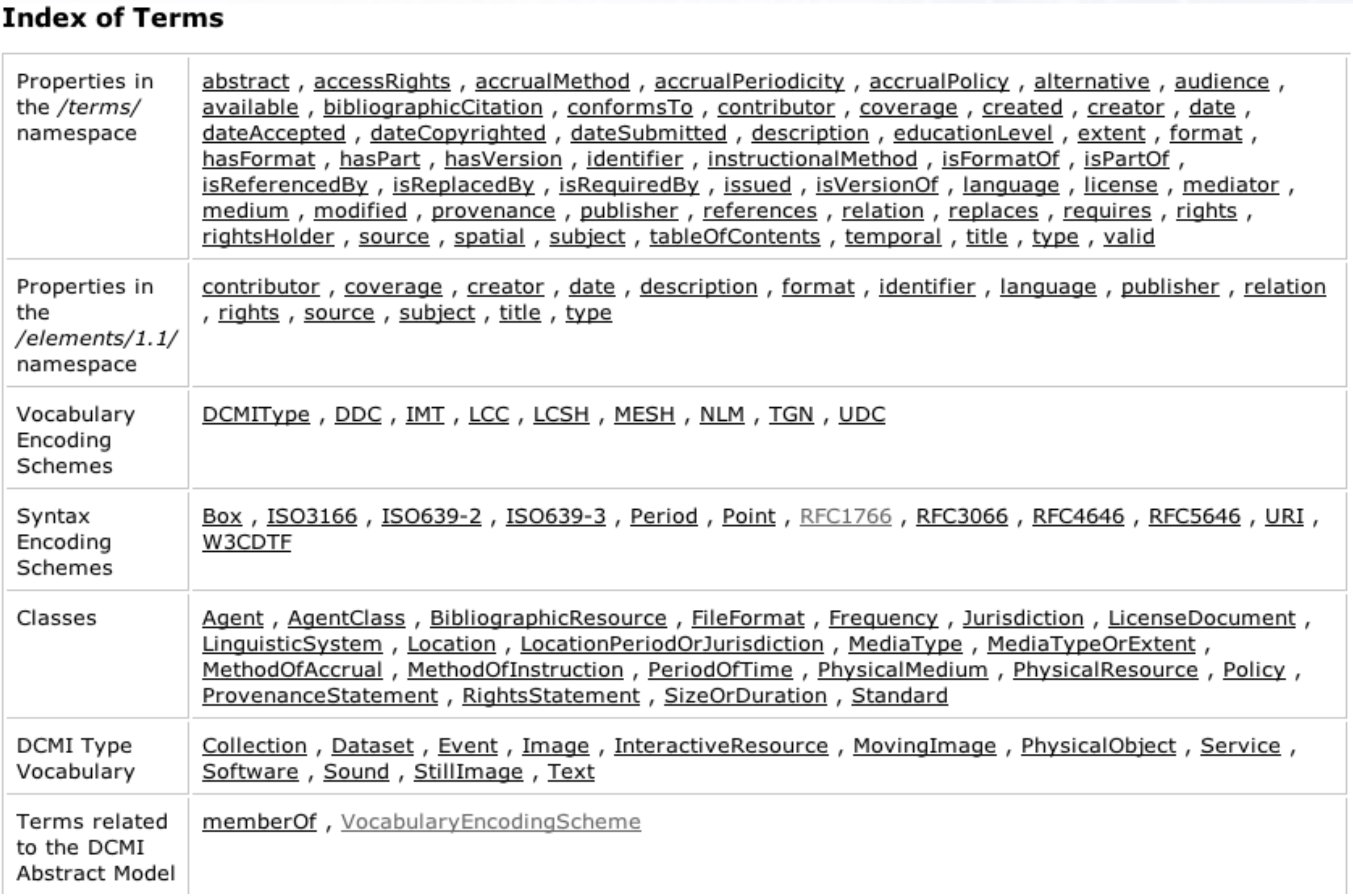
- There are **many** standards available to document data. Each has a different focus, yet ask for similar information about the data set.
- Your **choice** will depend on:
  - your field of practice
  - your motivation for using metadata

It’s still not always easy to make the choice. Initiatives such as the RDA’s Metadata Standards Directory Working Group are looking at ways to help people find the right metadata standard to use by compiling a directory. See: https://rd-alliance.org/group/metadata-standards-directory-working-group.html.

Metadata Standards: Examples

- Dublin Core Element Set
  - Emphasis on web resources, publications
  - http://dublincore.org/documents/dces/
  - Standardised in
    - ISO Standard 15836:2009 and ANSI/NISO Standard Z39.85-2012

Contributor	Coverage	Creator	Date	Description
Format	Identifier	Language	Publisher	Relation
Rights	Source	Subject	Title	Type



Index of Terms

Properties in the */terms/* namespace	abstract , accessRights , accrualMethod , accrualPeriodicity , accrualPolicy , alternative , audience , available , bibliographicCitation , conformsTo , contributor , coverage , created , creator , date , dateAccepted , dateCopyrighted , dateSubmitted , description , educationLevel , extent , format , hasFormat , hasPart , hasVersion , identifier , instructionalMethod , isFormatOf , isPartOf , isReferencedBy , isReplacedBy , isRequiredBy , issued , isVersionOf , language , license , mediator , medium , modified , provenance , publisher , references , relation , replaces , requires , rights , rightsHolder , source , spatial , subject , tableOfContents , temporal , title , type , valid
Properties in the */elements/1.1/* namespace	contributor , coverage , creator , date , description , format , identifier , language , publisher , relation , rights , source , subject , title , type
Vocabulary Encoding Schemes	DCMIType , DDC , IMT , LCC , LCSH , MESH , NLM , TGN , UDC
Syntax Encoding Schemes	Box , ISO3166 , ISO639-2 , ISO639-3 , Period , Point , RFC1766 , RFC3066 , RFC4646 , RFC5646 , URI , W3CDTF
Classes	Agent , AgentClass , BibliographicResource , FileFormat , Frequency , Jurisdiction , LicenseDocument , LinguisticSystem , Location , LocationPeriodOrJurisdiction , MediaType , MediaTypeOrExtent , MethodOfAccrual , MethodOfInstruction , PeriodOfTime , PhysicalMedium , PhysicalResource , Policy , ProvenanceStatement , RightsStatement , SizeOrDuration , Standard
DCMI Type Vocabulary	Collection , Dataset , Event , Image , InteractiveResource , MovingImage , PhysicalObject , Service , Software , Sound , StillImage , Text
Terms related to the DCMI Abstract Model	memberOf , VocabularyEncodingScheme

Metadata Standards: Examples

- FGDC* Content Standard for Digital Geospatial Metadata (CSDGM)
  - Emphasis on geospatial data
  - With Profiles & Extensions:
    - Biological Data Profile (BDP) of the CSDGM
    - Profile to the CSDGM emphasis on biological data (and geospatial)
  - http://www.fgdc.gov/metadata/geospatial-metadata-standards

  - *The **Federal Geographic Data Committee** (a US (government) interagency committee)

Metadata Standards: Examples

- ISO 19115/19139 Geographic information: Metadata
  - Emphasis on geospatial data and services
  - http://www.fgdc.gov/metadata/geospatial-metadata-standards#fgdcendorsedisostandards
- Ecological Metadata Language (EML)
  - Focus on ecological data
  - http://knb.ecoinformatics.org/eml_metadata_guide.html
- Darwin Core
  - Emphasis on museum specimens
  - http://rs.tdwg.org/dwc/index.htm

Metadata Standards: Examples

- Geography Markup Language (GML)
  - Emphasis on geographic features (roads, highways, bridges)
  - http://www.opengeospatial.org/standards/gml

Others…

- DDI: The Data Documentation Initiative
  - http://www.ddialliance.org/
- CDWA: Categories for the Description of Works of Art
  - http://www.getty.edu/research/publications/electronic_publications/cdwa/index.html

Semantic Annotation

- Annotating existing data, often data that has been created by others, and particularly derived or long-tail data (which is sometimes prone to errors)
- The data’s subsequent users want to annotate errors and create references to accepted ontologies and more up-to-date data from elsewhere
- Many of the technologies that can be used to annotate information on the semantic web can also be used in this context
- The same ontologies can also be used
- EUDAT has a Working Group active in this area

Further Reading on Metadata & Semantics

- http://www.niso.org/publications/press/UnderstandingMetadata.pdf
- http://www.dataone.org/education-modules
  - Lessons 7 & 8
- https://rd-alliance.org/working-groups/metadata-standards-directory-working-group.html
- http://www.eudat.eu/system/files/Semantics%20at%20the%20Second%20EUDAT%20Conference.pdf
- http://www.eudat.eu/User%20Documentation%20-%20B2FIND.html

Summary

- Metadata is “data about data” or “documentation of data”
- Metadata allows data to be discovered, accessed, understood and re-used
- Metadata standards provide structure and consistency to data documentation
- Standards and tools vary – select according to defined criteria such as data type, organisational guidance, and available resources
- Standards which have associated semantics add additional value

Acknowledgements & Re-Use

We suggest the attribution:

Based on a presentation prepared by Adam Carter for EPCC's MSc In High Performance Computing with Data Science

Contains content prepared as part of EUDAT's training on the Fundamentals of Data Infrastructures. See [www.eudat.eu/training.](http://www.eudat.eu/training)

Slides 11 and 12 are derived from content created by DCMI and made available under the terms of CC-BY version 3.0 unported. © DCMI 1995-2014. Used under license. Content taken from http://dublincore.org/documents/dcmi-terms/#terms-RFC1766.

Some slides (particularly those showing example metadata standards) are based on slides "Metadata and EUDAT" originally created by Shaun de Witt, STFC.

Slides labelled DataOne are from https://www.dataone.org/education-modules and are released into the public domain as CC0.

DataONE Education Module: Metadata. DataONE. Retrieved Sep 19, 2014. From http://www.dataone.org/sites/all/documents/L07_Metadata.pptx