



# EUDAT's Fundamentals of Data Infrastructures Metadata, Semantics, Ontologies

Adam Carter

EPCC, The University of Edinburgh

# What is Metadata?

- Data About Data, “Information that makes data useful”
- 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



# 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

# 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

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

# 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](http://knb.ecoinformatics.org/eml_metadata_guide.html)
- Darwin Core
  - Emphasis on museum specimens
  - <http://rs.tdwg.org/dwc/index.htm>



# Metadata Standards: Examples

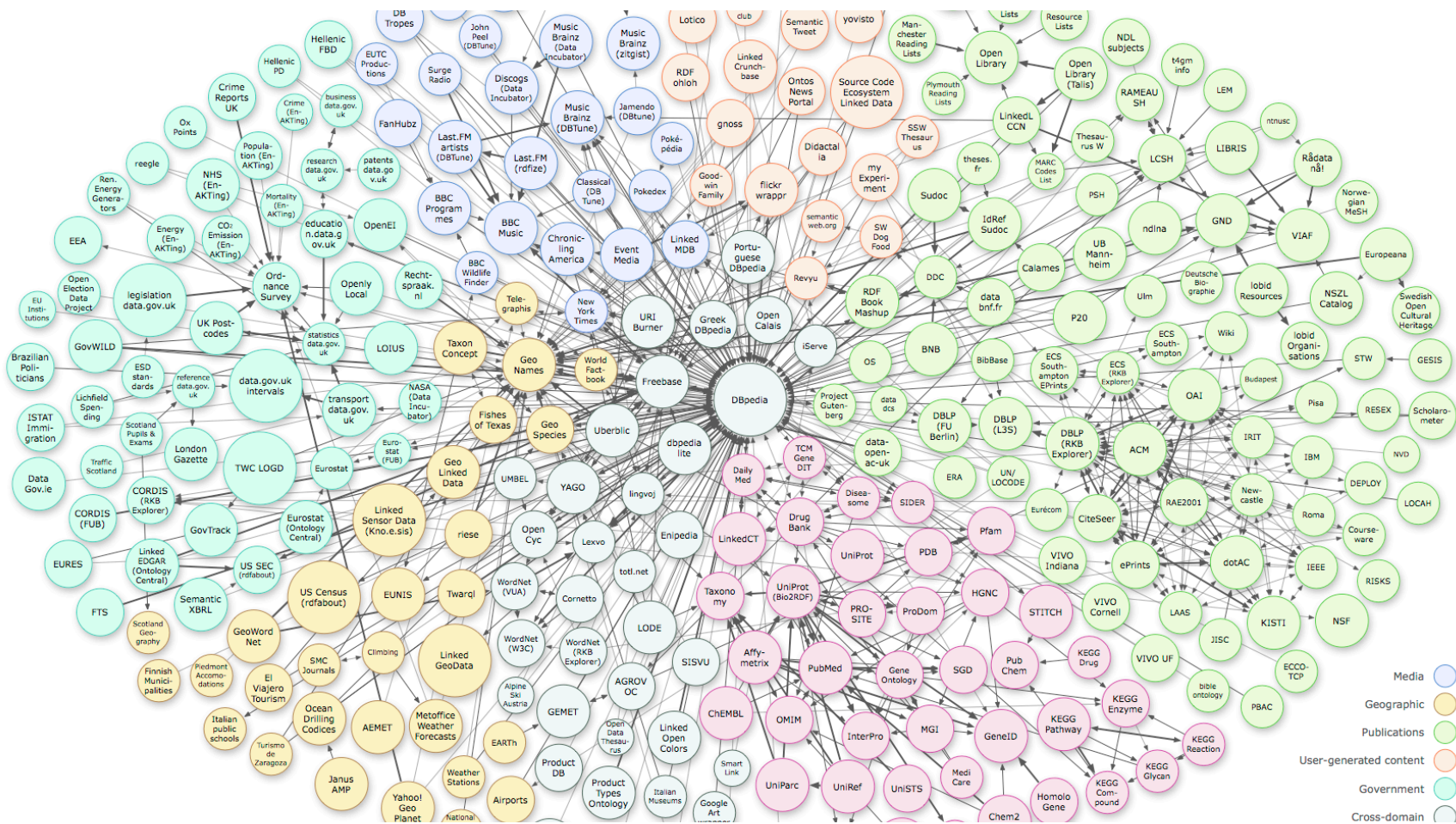
- Geography Markup Language (GML)
  - Emphasis on geographic features (roads, highways, bridges)
  - <http://www.opengeospatial.org/standards/gml>



## Others...

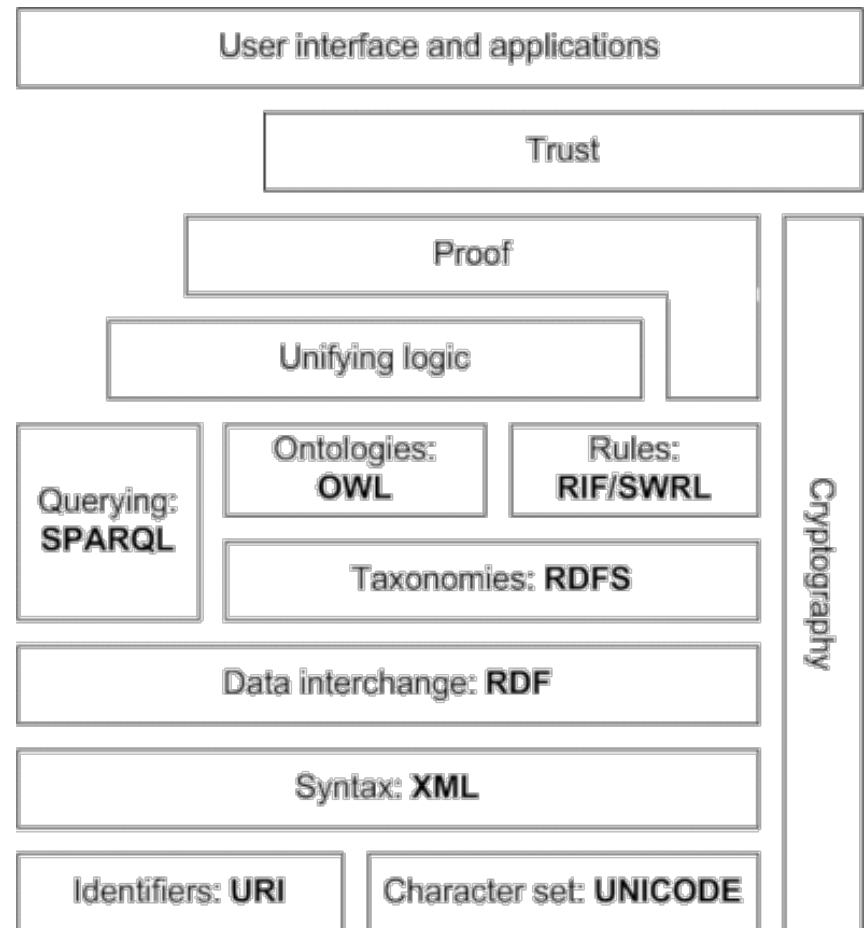
- DDI: The Data Documentation Initiative
  - <http://www.ddialliance.org/>
- CDWA: Categories for the Description of Works of Art
  - [http://www.getty.edu/research/publications/electronic\\_publications/cdwa/index.html](http://www.getty.edu/research/publications/electronic_publications/cdwa/index.html)

# Linked Open Data – The Semantic Web



# The Semantic Web

- A collaborative movement led by the W3C
- The set of (machine-readable) resources
  - Linked Open Data
- The formats and technologies that enable the above
- Based around RDF Triples:
  - Subject, Predicate, Object
  - where each of these is a URI



# 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>



# Acknowledgements & Re-Use

© 2014 The University of Edinburgh, and others (see below)



You are free to use this presentation and its content under the terms of CC-BY version 4.0.

We suggest the attribution:

*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” created by Shaun de Witt, STFC.