From Open Data to Data-Intensive Science through CERIF

Keith G Jeffery\textsuperscript{a}, Anne Asserson\textsuperscript{b}, Nikos Houssos\textsuperscript{c}, Valerie Brasse\textsuperscript{d}, Brigitte Jörg\textsuperscript{e}

\textsuperscript{a} Keith G Jeffery Consultants, Shrivenham, SN6 8AH, UK
\textsuperscript{b} University of Bergen, Bergen, 5009, Norway
\textsuperscript{c} Hellenic Documentation Centre, Athens, GR-11635, Greece
\textsuperscript{d} IS4RI, Strasbourg, France
\textsuperscript{e} JeiBee Ltd, London, UK
“Data is the new Oil”
(Ann Winblad)

• Like oil has been, data is
  – Abundant
  – Unrefined
  – Needs refining to extract value
  – Has great value when refined
  – Can be used in many ways

• So how do we gain value from data?
  – We manage it

• And what is required for that management?
  – Data management plan
  – Appropriate metadata covering all aspects of the data lifecycle
Structure

- The Concept of Open Data
- The Jungle of Open Data
- Metadata
- Open Government Data plus Research Data
- The e-Infrastructure Requirement
- CERIF for Data-Intensive Science
The Concept of Open Data

- **Open Government Data**
  - Motivation
    - Transparency, Commercialism
  - Technology
    - W3C LOD / RDF / CKAN
    - (but mainly .pdf, .csv, .xls)
  - Derivation
    - Summarised from publicly funded research data
  - Restrictions
    - licence

- **Research Data**
  - Motivation
    - Peer review, Re-use
  - Technology
    - Datasets / RDBMS
    - Schema level, discovery
  - Derivation
    - Observation, experiment, simulation
  - Restrictions
    - Commercial
    - national security
    - embargo
The Concept of Open Data

• Open Government Data
  • Motivation
    – Transparency, Commercialism
  • Technology
    – W3C LOD / RDF / CKAN
    – (but mainly .pdf, .csv, .xls)
  • Derivation
    – Summarised from publicly funded research data
  • Restrictions
    – licence

• Research Data
  • Motivation
    – Peer review, Re-use
  • Technology
    – Datasets / RDBMS
    – Schema level discovery
  • Derivation
    – Observation, experiment, simulation
  • Restrictions
    – Commercial
    – national security
    – embargo

With these problems no agreed standards for interoperability
Structure

- The Concept of Open Data
- The Jungle of Open Data
- Metadata
- Open Government Data plus Research Data
- The e-Infrastructure Requirement
- CERIF for Data-Intensive Science
The Jungle of Open Data

• Heterogeneity
  – Media / formats
  – Volume
  – Velocity (of change)
  – Domain
  – Language
  – Character set
  – Quality
  – Access restrictions
  – Availability guarantees
Open Data: We need to know

- Unique Identifier (for later use including citation)
- Location (URL)
- Description
- Keywords (terms)
- Temporal coordinates
- Geospatial coordinates
- Originator (organisation(s) / person(s))
- Project
- Facility / equipment
- Quality
- Availability (licence, persistence)
- Provenance
- Citations
- Related publications (white or grey)
- Related software
- Schema
- Medium / format
Open Data: We need to know

- Unique Identifier (for later use including citation)
- Location (URL)
- Description
- Keywords (terms)
- Temporal coordinates
- Geospatial coordinates
- Originator (organisation(s) / person(s))
- Project
- Facility / equipment
- Quality
- Availability (licence, persistence)
- Provenance
- Citations
- Related publications (white or grey)
- Related software
- Schema
- Medium / format

These are NOT properties or elements, they are relationships
Open Data: We need to know

- Unique Identifier (for later use including citation)
- Location (URL)
- Description
- Keywords (terms)
- Temporal coordinates
- Geospatial coordinates
- Originator (organisation(s) / person(s))
- Project
- Facility / equipment
- Quality
- Availability (licence, persistence)
- Provenance
- Citations
- Related publications (white or grey)
- Related software
- Schema
- Medium / format

These are NOT properties or elements, they are relationships

MANUAL
From human-readable to machine-understandable

AUTOMATED
Structure

• The Concept of Open Data
• The Jungle of Open Data
• Metadata
• Open Government Data plus Research Data
• The e-Infrastructure Requirement
• CERIF for Data-Intensive Science
Metadata

• Data about data (DCMI definition)
  – Unhelpful!
• Analogy of user of library
• Somehow describes internet resources for the end-user
Metadata

• Consider a library
  – Catalogue cards
  – Books on shelves

• To researcher or reader the catalogue cards are metadata
  – Describe the book and point to where it is on the shelf
  – Descriptive and navigational metadata

• To librarian catalogue cards are data
  – use catalogue cards to count number of books on ‘information technology

• So do not distinguish data and metadata except by how used
Data Lifecycle

Acknowledgement
DCC (Digital Curation Centre, UK)
Metadata

- Description
- Location
- Contextualisation
- Preservation
- Provenance
- Schema

- Discovery
- Context
- Detail

- Re-use
- Interoperation
There are hundreds of specific formats used as a ‘standard’ within a specific community but ones used widely are:

- **DC (Dublin Core):** used to describe web pages & web resources
- **CKAN (Comprehensive Knowledge Archive Network):** used in government open data sites – based on DC
- **eGMS; e-Government Metadata Standard** – based on DC
- **DCAT (Data Catalog):** used for datasets on the web – based on DC
- **INSPIRE:** used for datasets with geospatial coordinates
  - EU Directive and standard; some overlap with DC but extended
- **CERIF (Common European research Information Format):** used for all research information

All but CERIF are ‘flat’ or ‘linear’
Metadata Standards: DC

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

- Text
- HTML
- XML
- RDF
- Namespaces
- Ontologies
Metadata Standards: CKAN

- Title
- Unique Identifier
- Groups
- Description
- Revision History
- Licence
- Tags
- Multiple Formats
- API key
- Extra Fields

Black signifies same as DC

- RDF
- ontologies

©Keith G Jeffery et al
CRIS14 Rome May 2014
Metadata Standards: e-GMS

- Accessibility
- Addressee
- Aggregation
- Audience
- Contributor
- Coverage
- Creator
- Date
- Description
- Digital signature
- Disposal
- Format
- Identifier

- Language
- Location
- Mandate
- Preservation
- Publisher
- Relation
- Rights
- Source
- Status
- Subject
- Title
- Type

Black signifies same as DC

©Keith G Jeffery et al

CRIS14 Rome May 2014
Metadata Standards: DCAT

Same as DC are: Title, description, identifier, keyword, language
Metadata Standards: INSPIRE

• EU Directive (2008, 2009)
• For Geospatial datasets
  – Initiated by ESA
• Essentially DC plus geospatial information
• Geospatial information very detailed – coordinate system, precision etc
Metadata Standards: CERIF

- **Common European Research Information Format**
- Data Model for exchange and storage of information about research
- CERIF91 (1987-1990) quite like the later Dublin Core (late 1990s)
- CERIF2000 (1997-1999) used full E-E-R modelling
  - Base entities
  - Linking entities with role and temporal interval
- 2002 EC requested euroCRIS to maintain, develop and promote CERIF [www.eurocris.org](http://www.eurocris.org)
- Now in use in 43 countries and national standard for research information in 10
# Metadata Comparison (1)

| # | Feature | Use case | CERIF | Dublin Core | CKAN | DCAT |
|---|---------|----------|-------|-------------|------|------|
| 1 | Representation of graph structures | Realistic representation of domain of discourse, Generation of Linked Open Data | YES | YES | NO | YES |
| 2 | Typed values enforced for values that are entity instances | Unambiguous identification of types and instances | YES | NO | NO | YES |
| 3 | Explicit representation of resources (e.g. data files) | Different physical embodiments of what the metadata describes | YES | NO | YES | YES |
| 4 | Time-stamping of relationships | Accurate real-world representation, provenance, versioning | YES | NO | NO | NO |
## Metadata Comparison (2)

|   | Capture both dates and actors of events | Accurate real-world representation, provenance, versioning | YES | Only dates | Only dates | Only dates |
|---|----------------------------------------|--------------------------------------------------------|-----|------------|------------|------------|
| 5 |                                        |                                                        |     | YES        | Only dates | Only dates |
|   | Recursive relationships                 | Recursive relationships                                 | YES | YES        | NO         | NO         |
| 6 |                                        | Recursive relationships                                | YES | YES        | NO         | NO         |
|   | Extensible relationship semantics       |                                                        |     | YES        | NO         | NO         |
| 7 |                                        | Extensible relationship semantics                      | YES | NO         | NO         | NO         |
|   | Representation and crosswalking between vocabularies | Representation and crosswalking between vocabularies | YES | NO         | NO         | YES/NO     |
| 8 |                                        |                                                        |     | YES        | YES        | YES        |
|   | Multilingual values for the same metadata field | Multilingual values for the same metadata field | YES | YES        | YES        | YES        |
| 9 |                                        |                                                        |     | YES        | NO         | NO         |
| 10| Translated flag for multi-linguality   | Translated flag for multi-linguality                 | YES | NO         | NO         | NO         |
The Problem with ‘flat’ metadata

• they violate basic principles of information integrity
  – elements do not depend functionally on the uniquely identified metadata record.
• they store event flags or dates in the metadata
  – e.g. ‘date of publication’, ’received (Y/N)’
• they do not handle well multilinguality and multiple linguistic versions of the same text field;
• they do not manage well versioning and provenance
  – this requires time-stamped relationships between one research information entity and another
• they do not allow multiple classification schemes for the same entity or – more generally – multiple terminology schemes for the same attribute of an entity;
• they do not provide mechanisms for crosswalking between different vocabularies;
• they do not provide extension mechanisms that preserve interoperability;
CERIF

Dataset is here
CERIF Features

- Developed by international community – consensus
- Flexible and extensible
- Separation of base and link entities
  - Flexible / extensible
  - Rich semantics (role)
  - Temporal: it is the relationships that have duration
- Multi character set
- Multilingual
- Formal Syntax
  - Efficient, accurate computer processing
- Declared Semantics
  - Including crosswalks for interoperation
Repositories and CERIF

• To view content (white or grey) in repositories through contextualised, structured metadata
  – E.g. Relate publication to:
    • Persons
    • Organisations
    • Projects
    • Funding
    • Facilities
    • Equipment
    • Event
    • Patent
    • Product
  
• Repository metadata DC (Dublin Core) insufficient
• (as recognised by OpenAIREPlus when adopted CERIF)

Allows the user to judge better relevance, quality
Metadata RDA

- Metadata Interest Group
- Metadata Standards Directory Working Group
- Data In Context Interest Group
- Working with Provenance Group
- And data citation group
- And groups on repositories, types...
- An various domain-specific groups
Structure

• The Concept of Open Data
• The Jungle of Open Data
• Metadata
• Open Government Data plus Research Data
• The e-Infrastructure Requirement
• CERIF for Data-Intensive Science
The Vision: Metadata Stack

- **DISCOVERY (DC, eGMS...)**
- **CONTEXT (CERIF)**
- **DETAIL (SUBJECT OR TOPIC SPECIFIC)**

Linked open data

Formal Information Systems

Generate

Point to
Open Data and Information Processing

Manual download
Manual connection to software
Manual integration

Example: summary data in semantic web/LOD environment (RDF) with associated processing

generate
provide access to

Relational (Links)
Integrity, performance

Example: research datasets in Relational DB environment with associated analysis, visualisation, data mining ....

Automated download
Automatic connection to software
Automated integration
3-Layer Model

- DISCOVERY METADATA
  (OKAN, e_GMS, DC..)
- CONTEXTUAL METADATA
  (CERIF)
- DETAILED METADATA
  (DOMAIN-SPECIFIC)

SERVICES USING
DISCOVERY
METADATA

SERVICES USING
CONTEXTUAL
METADATA

SERVICES TIED TO
DOMAIN-SPECIFIC
ACTIVITY

Target Dataset(s)
Structure

• The Concept of Open Data
• The Jungle of Open Data
• Metadata
• Open Government Data plus Research Data
• The e-Infrastructure Requirement
• CERIF for Data-Intensive Science
The Vision: The Models

Complete cohort of researchers, research managers, innovators, media

- **User Model**: interaction with data, processing, persons
- **Processing Model**: providing what the user requires
- **Data Model**: representing research
- **Resource Model**: representing ICT

Complete ICT environment for research
Structure

• The Concept of Open Data
• The Jungle of Open Data
• Metadata
• Open Government Data plus Research Data
• The e-Infrastructure Requirement
• CERIF for Data-Intensive Science
CONCLUSION

• Unique Identifier (for later use including citation)
• Location (URL)
• Description
• Keywords (terms)
• Temporal coordinates
• Geospatial coordinates
• Originator (organisation(s)/ person(s))
• Project
• Facility / equipment
• Quality
• Availability (licence, persistence)
• Provenance
• Citations
• Related publications (white or grey)
• Related software
• Schema
• Medium / format