“Data is the new Oil” (Ann Winblad)

Keith G Jeffery

keith.jeffery@keithgjefferyconsultants.co.uk

Data is the New Oil

- 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**

STFC Rutherford Appleton Laboratory

CAREER

- Late 60s First UK relational system: G-EXEC
- 70s Filematch: interoperation
- Early 80s Online grants, library, science
- Late80s IDEAS, EXIRPTS
- 90s CERIF
- 90s W3C Standards
  - CGI, SVG, SMIL, OWL, SKOS
- 00s e-Science
  - GRIDs, CLOUDs

Agenda

- Kinds of data
  - Open government data, open data, big data
- Need for metadata
  - problems with exiting standards and experience of ENGAGE
- CERIF
  - for research information
  - wider
- CERIF / INSPIRE proposed mapping
- Metadata in RDA context
- Conclusion

Data Characterisation

- Data
  - Structured
  - Semi-structured
  - Unstructured

  - Static
  - Dynamic
  - Streamed

  - Secure / open
  - Private / public

  - Toll free/Toll

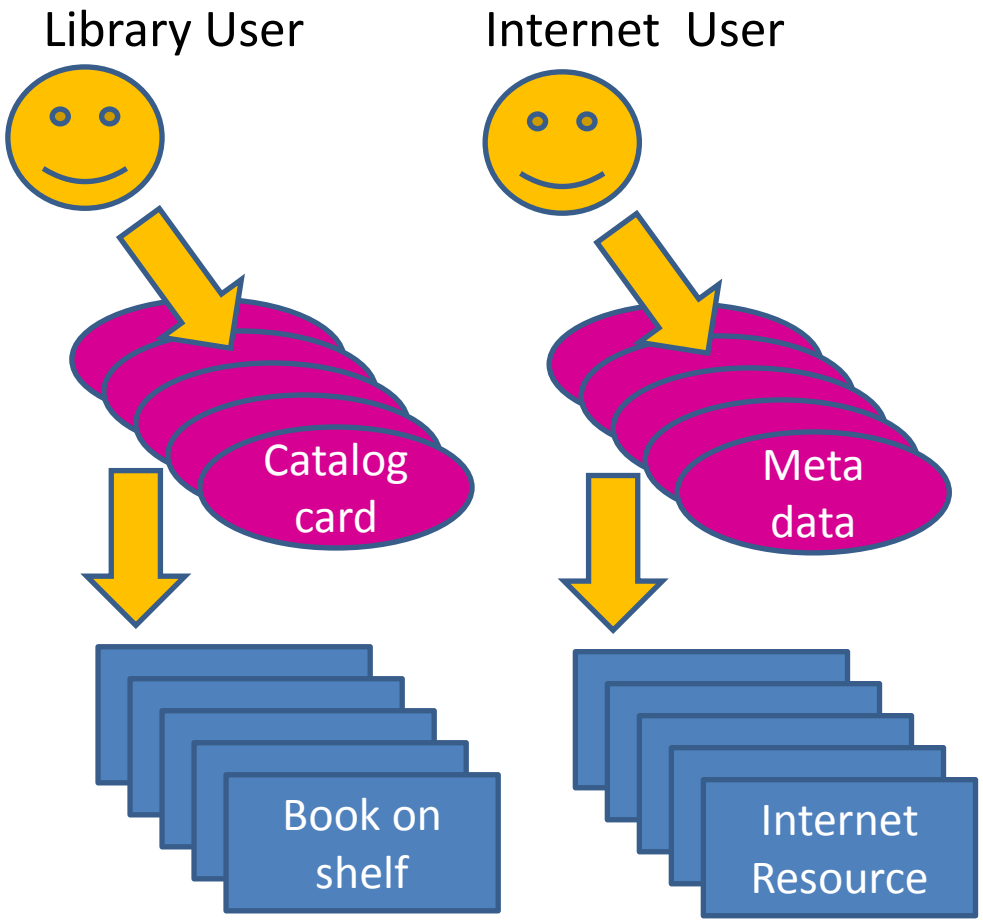
- Open government data
- Open data
- Big data

Agenda

- Kinds of data
  - Open government data, open data, big data
- Need for metadata
  - problems with exiting standards and experience of ENGAGE
- CERIF
  - for research information
  - wider
- CERIF / INSPIRE proposed mapping
- Metadata in RDA context
- Conclusion

Metadata

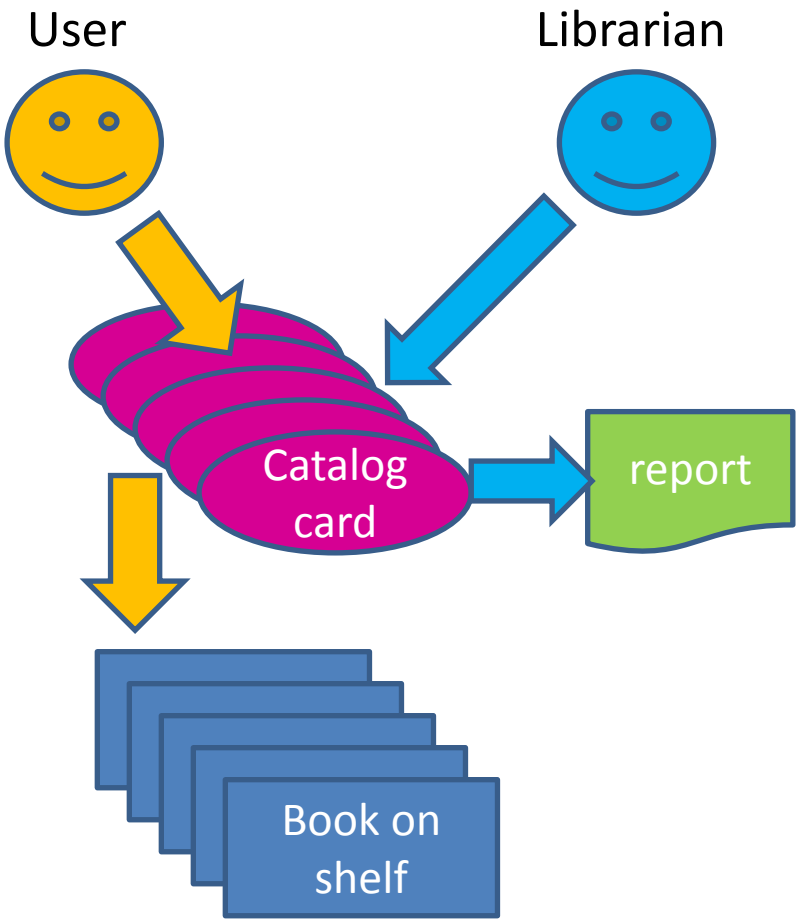
- Data about data (DCMI defintion)
  - Unhelpful!
- Analogy of user of library
- Somehow describes internet resources for the end-user



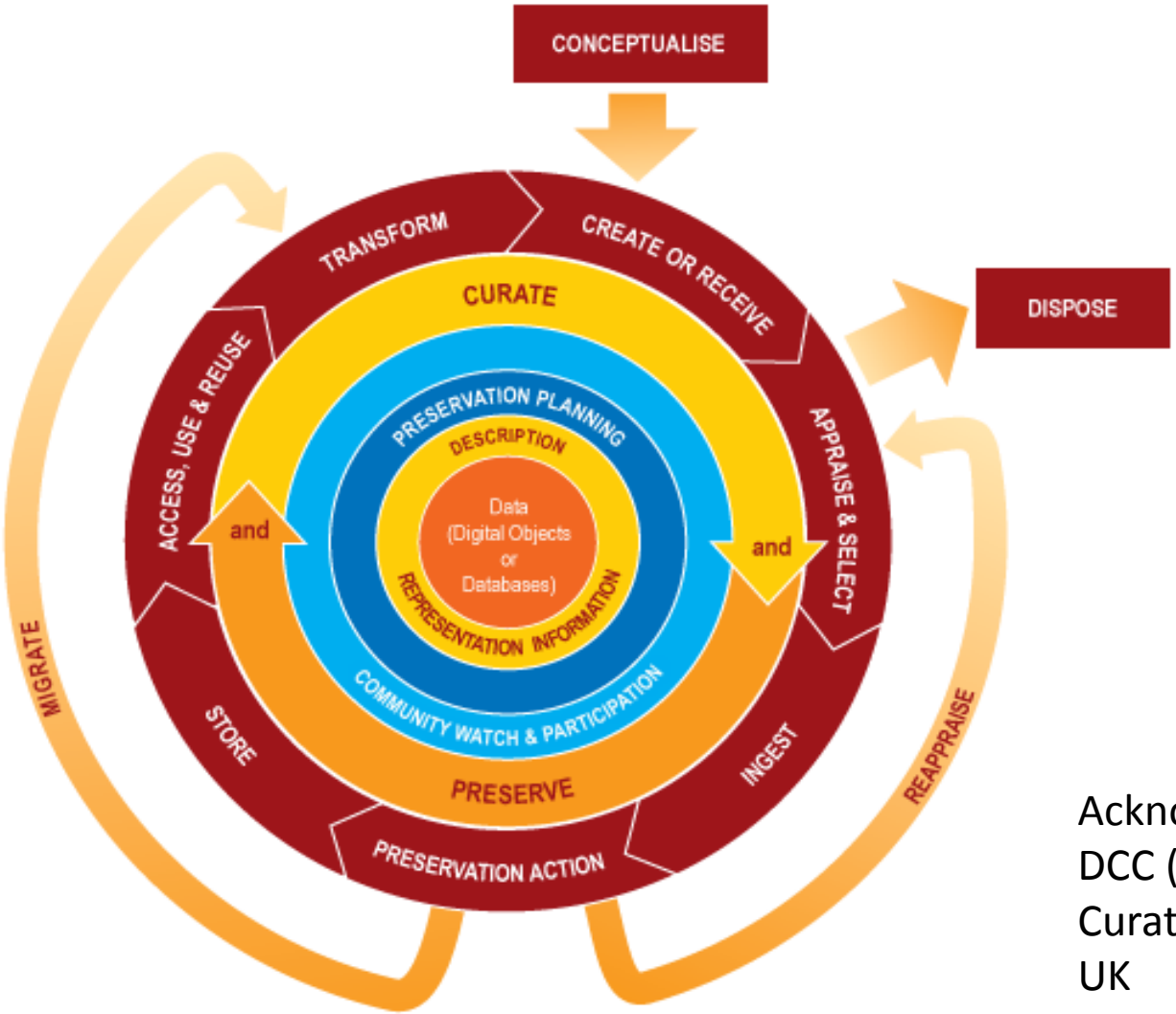
Metadata

- 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

Metadata Standards

- 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

- RDF
- ontologies

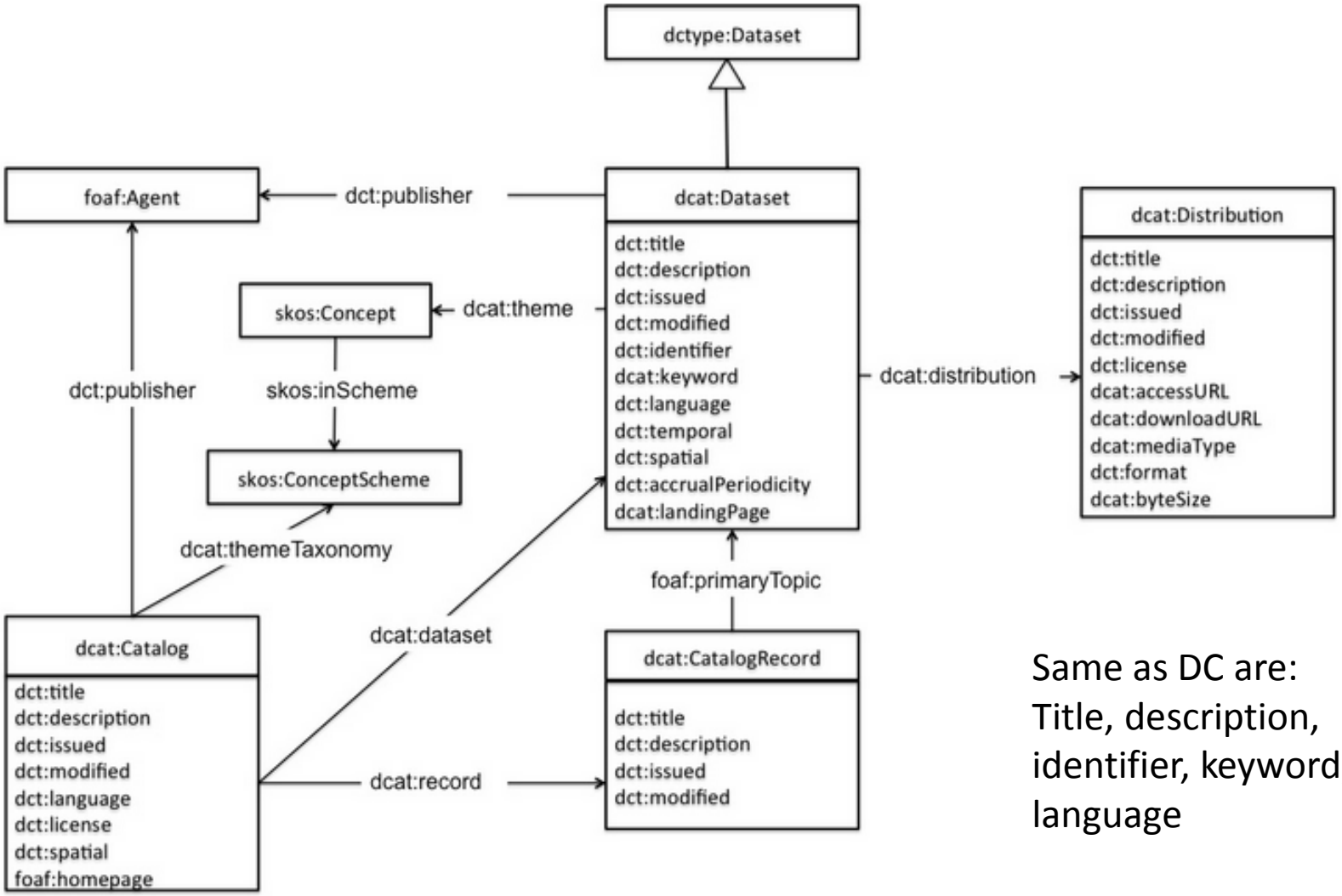
Black signifies same as DC

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

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

- **C**ommon **E**uropean **R**esearch **I**nformation **F**ormat
- 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
- 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)

5	Capture both the dates and actors of events	Accurate real-world representation, provenance, versioning	YES	Only dates	Only dates	Only dates
6	Recursive relationships	Compound objects, Derived objects	YES	YES	NO	NO
7	Extensible relationship semantics	Complex objects, accurate semantics	YES	NO	NO	NO
8	Representation and crosswalking between vocabularies	Co-existence of different vocabularies	YES	NO	NO	YES/NO
9	Multilingual values for the same metadata field	Multi-lingual environment (e.g. Europe)	YES	YES	YES	YES
10	Translated flag for multi-linguality	Warn metadata consumers (including programs) for machine translated values	YES	NO	NO	NO

The Problem with ‘flat’ metadata

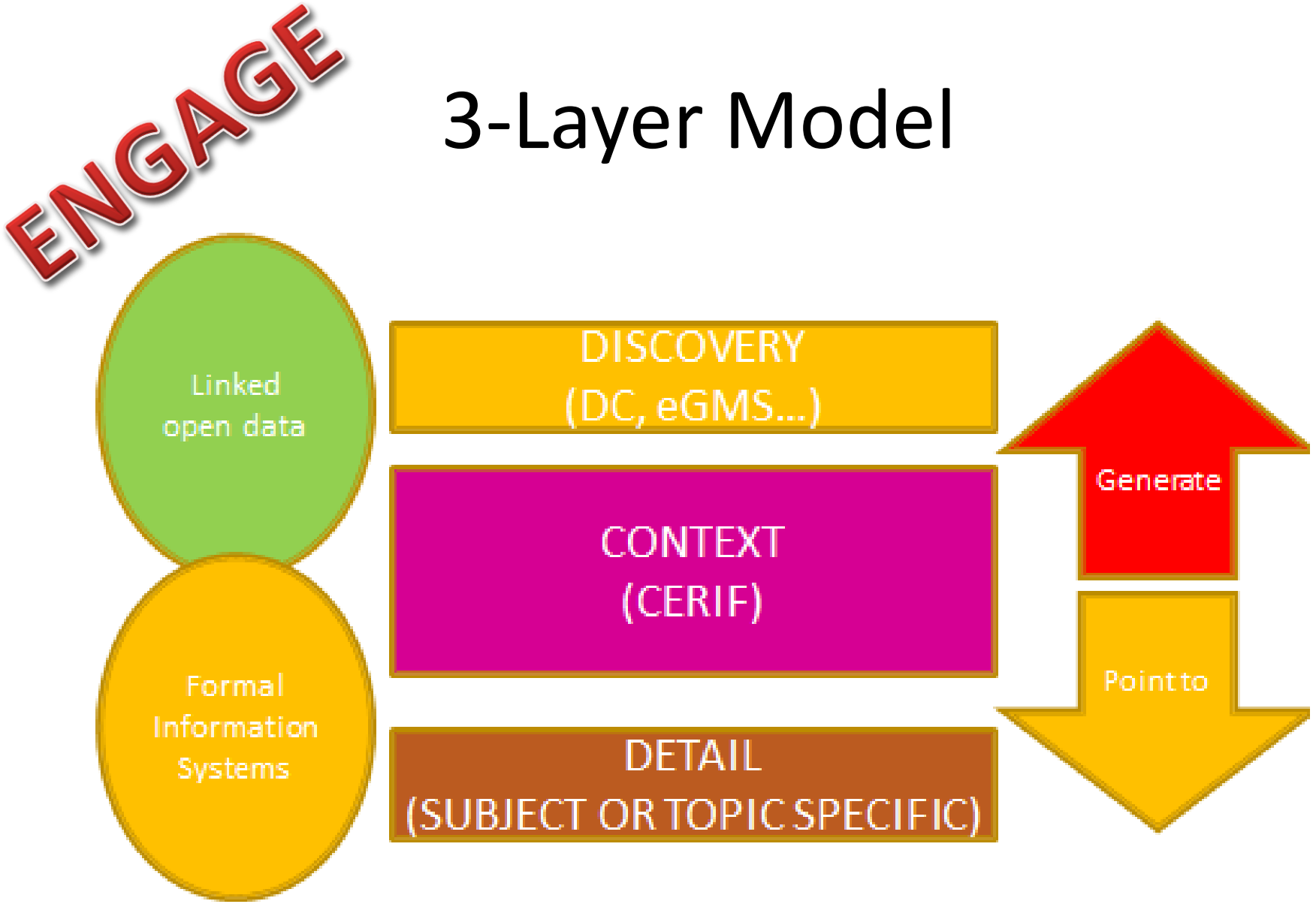
- they violate basic principles of information integrity
  - elements do not depend functionally on the uniquely identified metadata record.
- they store event flags or dates in the metadata
  - e.g. ‘date of publication’, ’received (Y/N)’
- they do not handle well multilinguality and multiple linguistic 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;

3-Layer Model

- Need to interoperate at discovery level with other commonly-used metadata standards
- Need to navigate user to detailed domain-specific metadata on datasets to allow further (re-)processing
- Between these two need to understand the **CONTEXT** of the described objects (not only data)

- So use CERIF as the middle contextual layer
- Generate discovery level (above)
- Point to detailed level (below)

3-Layer Model



3-Layer Model

Agenda

- Kinds of data
  - Open government data, open data, big data
- Need for metadata
  - problems with exiting standards and experience of ENGAGE
- CERIF
  - for research information
  - wider
- CERIF / INSPIRE proposed mapping
- Metadata in RDA context
- Conclusion

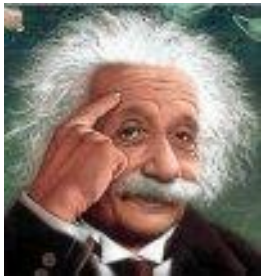
euroCRIS
Current Research Information Systems

Information of Interest

Institution

Books

Person / CV



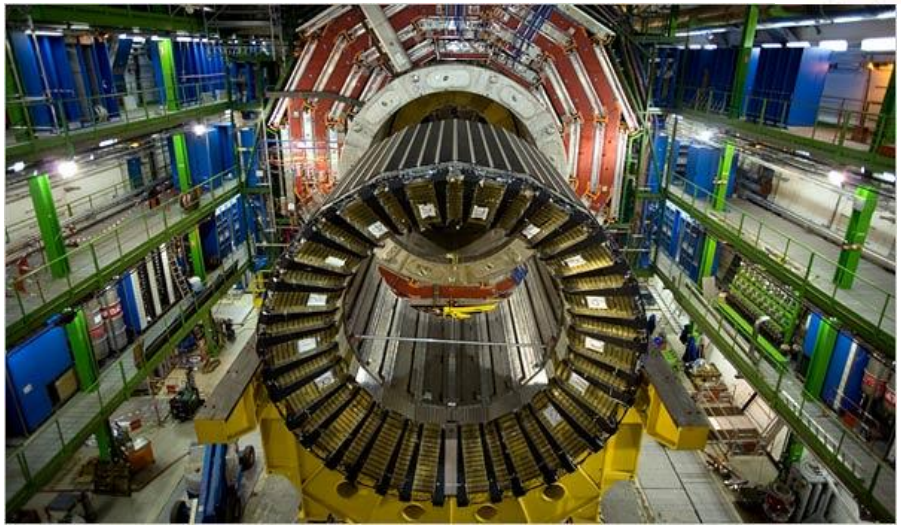
Publisher

Research Group

Patent

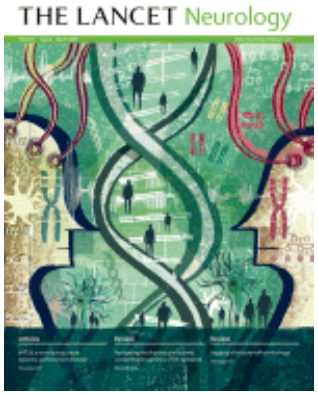
Event

Project

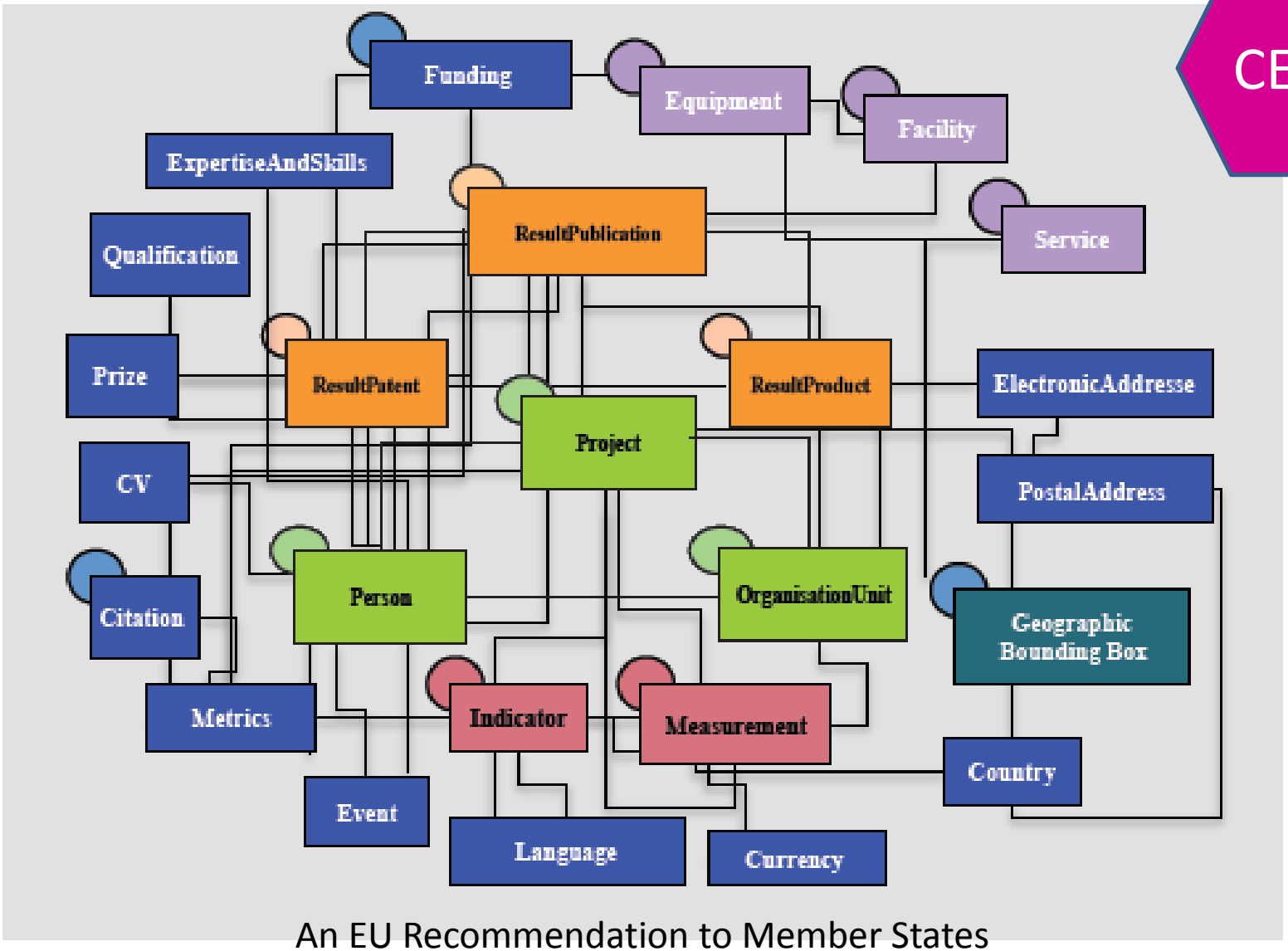


Equipment

Journal/article



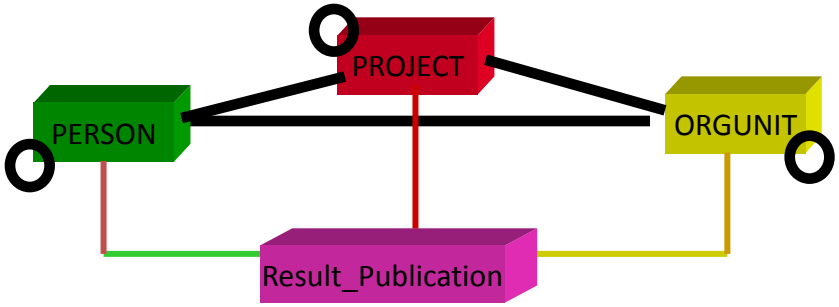
Contextual Metadata: CERIF



An EU Recommendation to Member States

RESULT_PUBLICATION

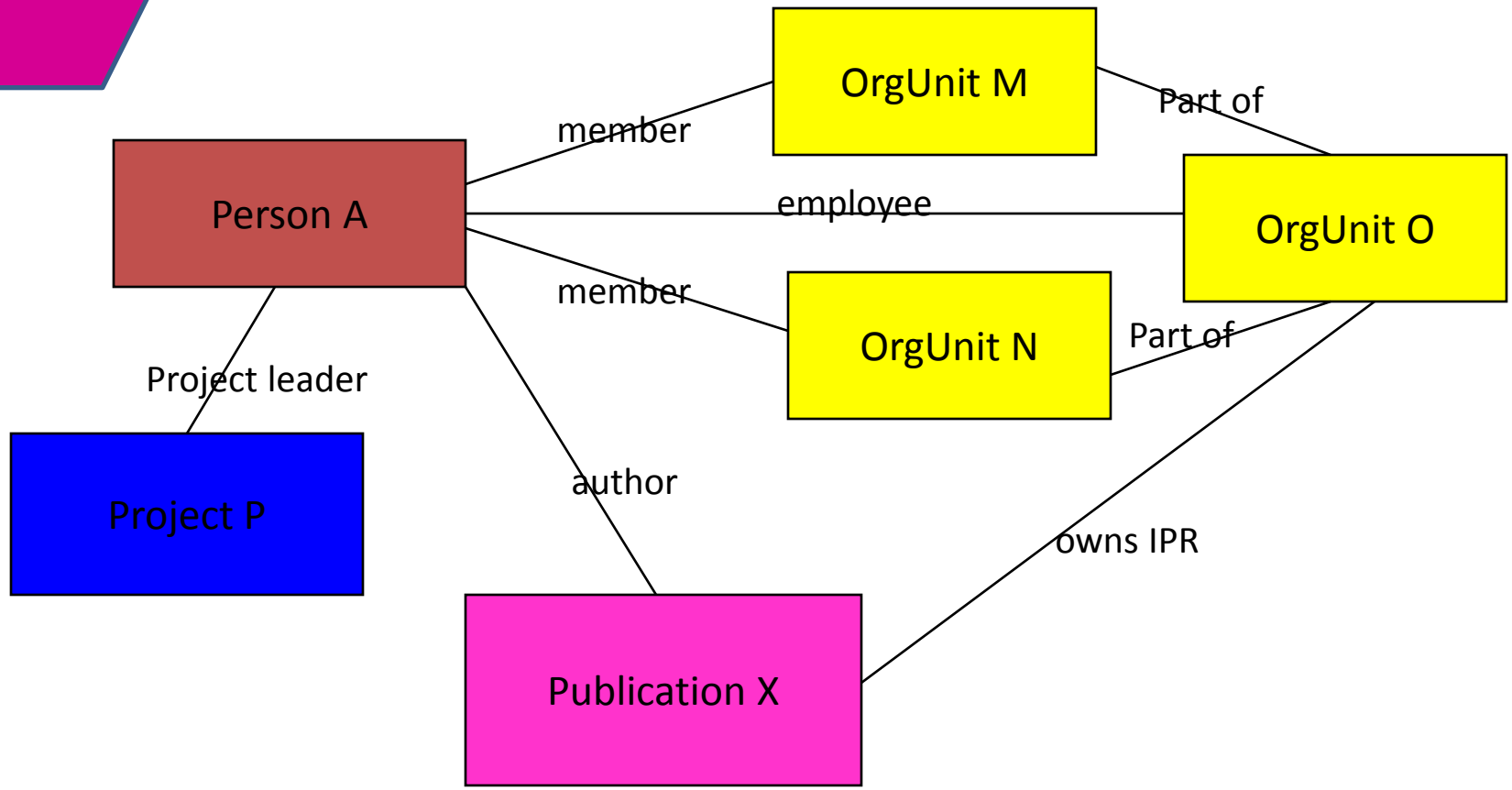
CERIF Expressiveness



Can Express:

Person A	(DT1 - DT2)	(is author of)	Publication X
Orgunit O	(DT1 - DT2)	(is owner of IPR in)	Publication X
Person A	(DT1 - DT2)	(is employee of)	Orgunit O
Person A	(DT1 - DT2)	(is project leader of)	Project P
Person A	(DT1-DT2)	(is member of)	Orgunit M
Person A	(DT1-DT2)	(is member of)	Orgunit N
Orgunit M	(DT1-DT2)	(is part of)	Orgunit O
Orgunit N	(DT1-DT2)	(is part of)	Orgunit O

Result_Publication Instance Diagram



euroCRIS
Current Research Information Systems

CERIF Features

CERIF

- 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 characterset
- 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

Allows the user to judge better relevance, quality

- Repository metadata DC (Dublin Core) insufficient
- (as recognised by OpenAIREPlus when adopted CERIF)

Agenda

- Kinds of data
  - Open government data, open data, big data
- Need for metadata
  - problems with exiting standards and experience of ENGAGE
- CERIF
  - for research information
  - wider
- CERIF / INSPIRE proposed mapping
- Metadata in RDA context
- Conclusion

INSPIRE-CERIF

Fairly straighforward
Already have CERIF-DC
INSPIRE geospatial elements ➔ CERIF: GeoBbox

- Identifier (Title, ID, Abstract, Locator)
- Classification
- Keyword
- Geo B Box, Country
- Temporal (dates)
- Lineage
- Resolution
- Conformity
- Constraints (use)
- Responsible party

- Title, ID, Abstract, URI
- Class Scheme, Class
- Keyword
- Geo B Box, Country
- Linking relations start/end date/time
- Linking relations temporal/classification
- Measurement
- Linking relation to certifier
- Linking relation to licence
- Linking relation to OrgUnit/Person

Agenda

- Kinds of data
  - Open government data, open data, big data
- Need for metadata
  - problems with exiting standards and experience of ENGAGE
- CERIF
  - for research information
  - wider
- CERIF / INSPIRE proposed mapping
- Metadata in RDA context
- Conclusion

Metadata RDA

- Metadata Interest Group
- Metadata Standards Directory Working Group
- Data In Context Interest Group

- Working with Provenance Group
- and groups on repositories, types…
- An various domain-specific groups

Agenda

- Kinds of data
  - Open government data, open data, big data
- Need for metadata
  - problems with exiting standards and experience of ENGAGE
- CERIF
  - for research information
  - wider
- CERIF / INSPIRE proposed mapping
- Metadata in RDA context
- Conclusion

Conclusion

- Data is the new Oil

- Metadata is the catalyst to make it useful