An Infrastructure for Open, Linked Governmental Data Provision towards Research Communities and Citizens

Keith G Jeffery, Anne Asserson,

Nikos Houssos, Brigitte Jörg

- Research and Research Information
- Metadata
- Problems with Metadata Formats
- CERIF
- A 3-layer Model for Metadata

Research and Research Information

- Research leads to wealth creation and improvement in the quality of life.
- Research Information needs to be collected, made available, communicated and curated.

- Researchers: managing CV, bibliography, generating web pages and finding collaborators.
- Research managers: evaluation, benchmarking, managing intellectual property
- Innovators: ideas through to products and services.
- Media: ‘research stories’ and by citizens interested in research and in ‘citizen science’.

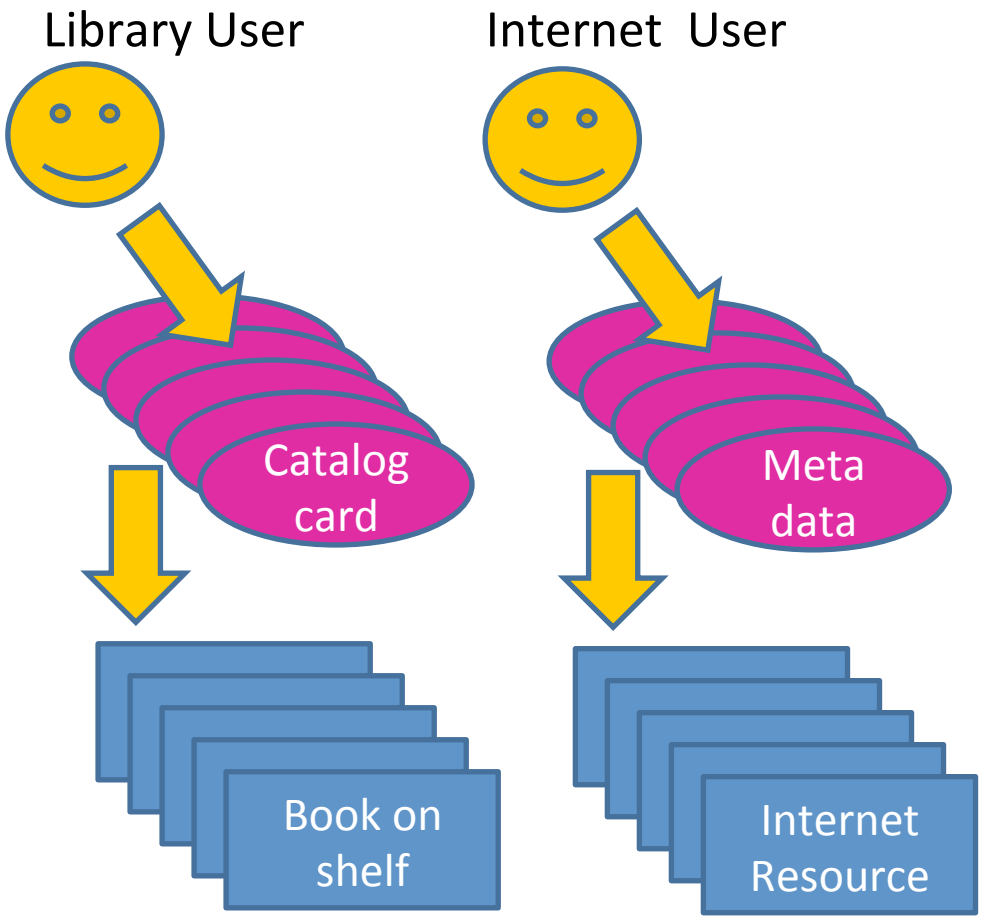
Research and Research Information

- One research product is research datasets – and associated software.
- Discovery and use needs metadata
- Context of the dataset
  - why it was collected, by whom, under what conditions and using what equipment at which organisation.
  - How the dataset relates to the purposes of the project, the funding and related scholarly publications (both white and grey).
- All of this contextual information assists the end-user in judging the applicability and quality of the dataset for their (re-)purposing.

- Research and Research Information
- Metadata
- Problems with Metadata Formats
- CERIF
- A 3-layer Model for Metadata

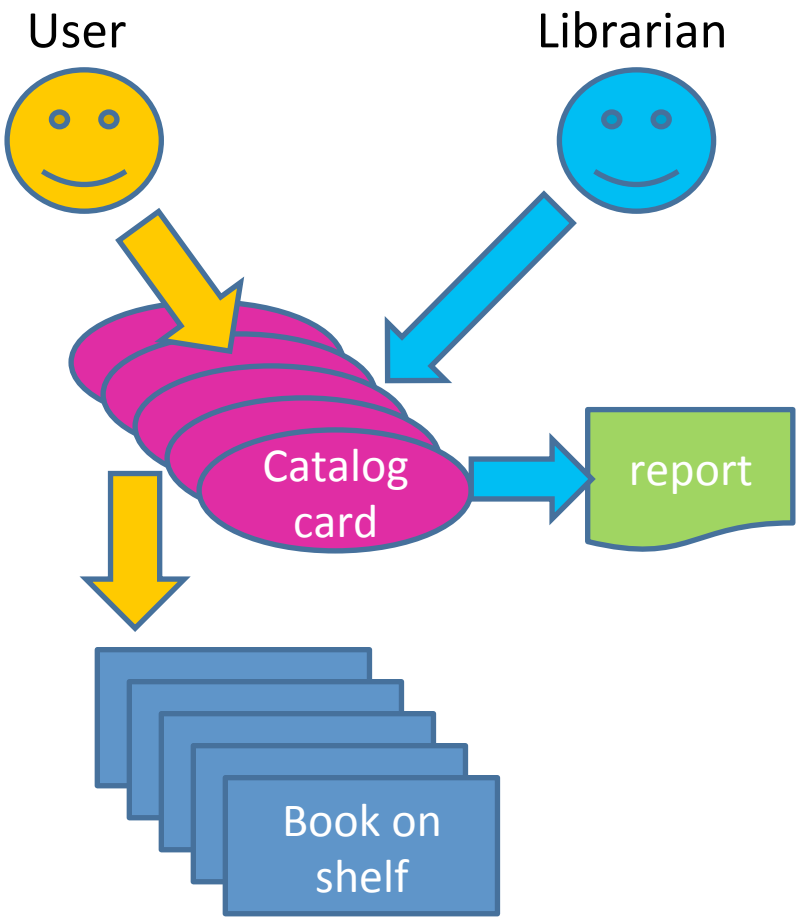
Metadata

- 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**



- Research and Research Information
- Metadata
- Problems with Metadata Formats
- CERIF
- A 3-layer Model for Metadata

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

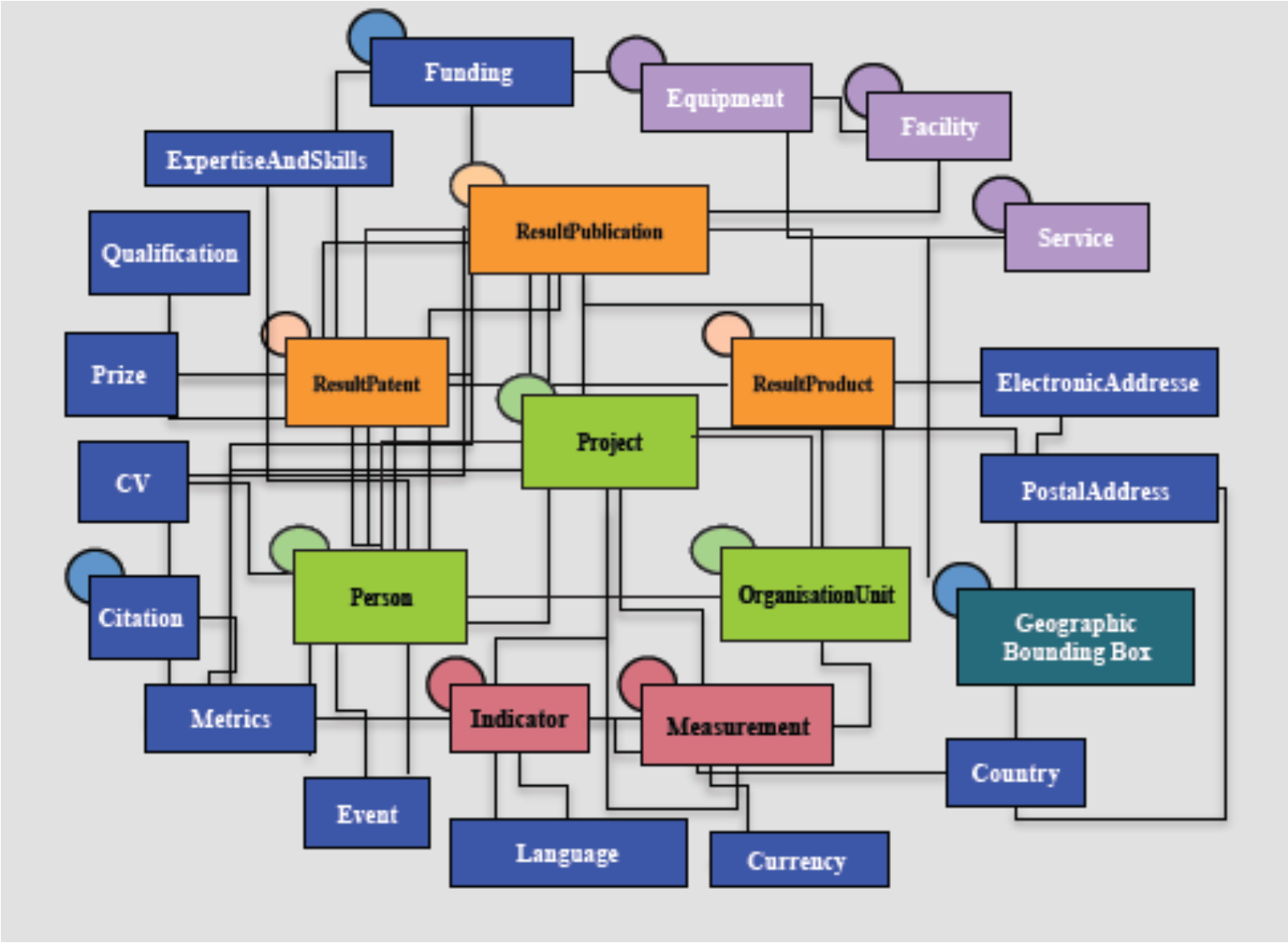
- they violate basic principles of information integrity
  - elements do not depend functionally on the uniquely identified metadata record.
- they store event flags or dates in the metadata
  - e.g. ‘date of publication’.
- they do not handle well multilinguality and multiple linguistic versions of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

- Research and Research Information
- Metadata
- Problems with Metadata Formats
- CERIF
- A 3-layer Model for Metadata

CERIF History

- **C**ommon **E**uropean **R**esearch **I**nformation **F**ormat
- Developed by an EC-organised group of government-appointed experts representing member states (and EEA);
- CERIF91 was not unlike Dublin Core
  - Experience 1990-1996 highlighted problems
- CERIF2000
  - Extended Entity-Relationship Model
  - Formal syntax and declared semantics
- EU Recommendation to Member States
  - i.e. a 'standard'
- 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

Contextual Metadata: CERIF



CERIF Features

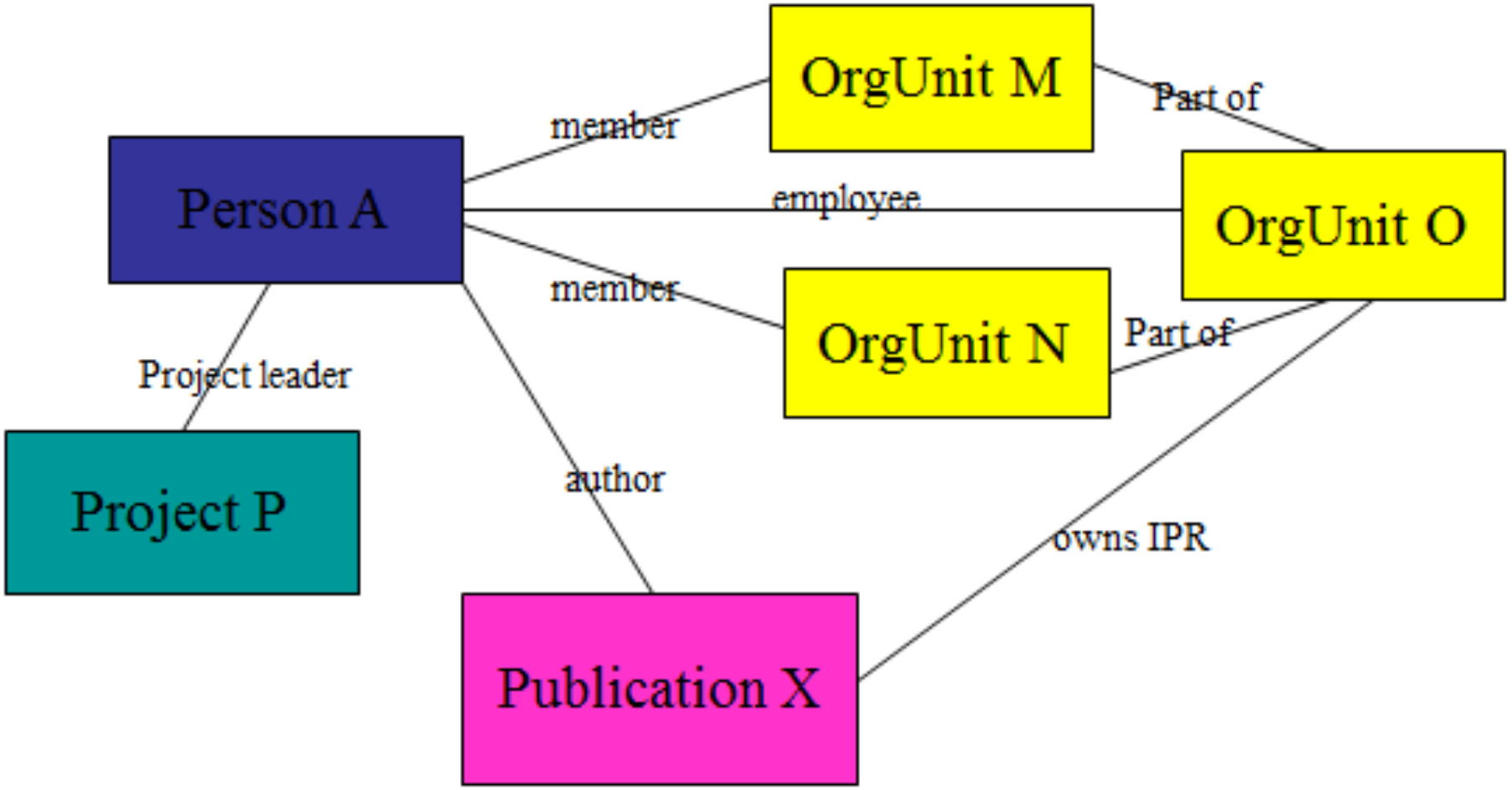
- it separates base entities (e.g. project, person, organisation, publication) from linking entities which link together instances of 2 base entities with a role (author, employee, project leader) and temporal interval of validity.
  - This is much more advanced in semantics and integrity than hypermedia models, the use of XLINK or LOD (Linked Open Data);
- it has formal syntax and declared semantics: it separates all terms into a semantic layer referenced from the syntax (so in link entities the role is a pointer to the semantic layer and in base entities list-restricted attribute values such as country code are in the semantic layer).
  - This ensures consistency and integrity;
- the linking mechanism also applies in the semantic layer so terminology schemes and the terms within them can be related with role and temporal duration.
  - This allows semantic crosswalking for interoperability;
- the richness of CERIF means it can act as a superset interoperation hub for other metadata formats, generating them congruently from the CERIF format.
  - This permits interoperation;

CERIF Features

Base entity	Temporal range	Role	Base entity

- Person A (DT1 - DT2) (is author of) Publication X
- OrgunitO (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
- OrgunitM (DT1-DT2) (is part of) Orgunit O
- OrgunitN (DT1-DT2) (is part of) Orgunit O

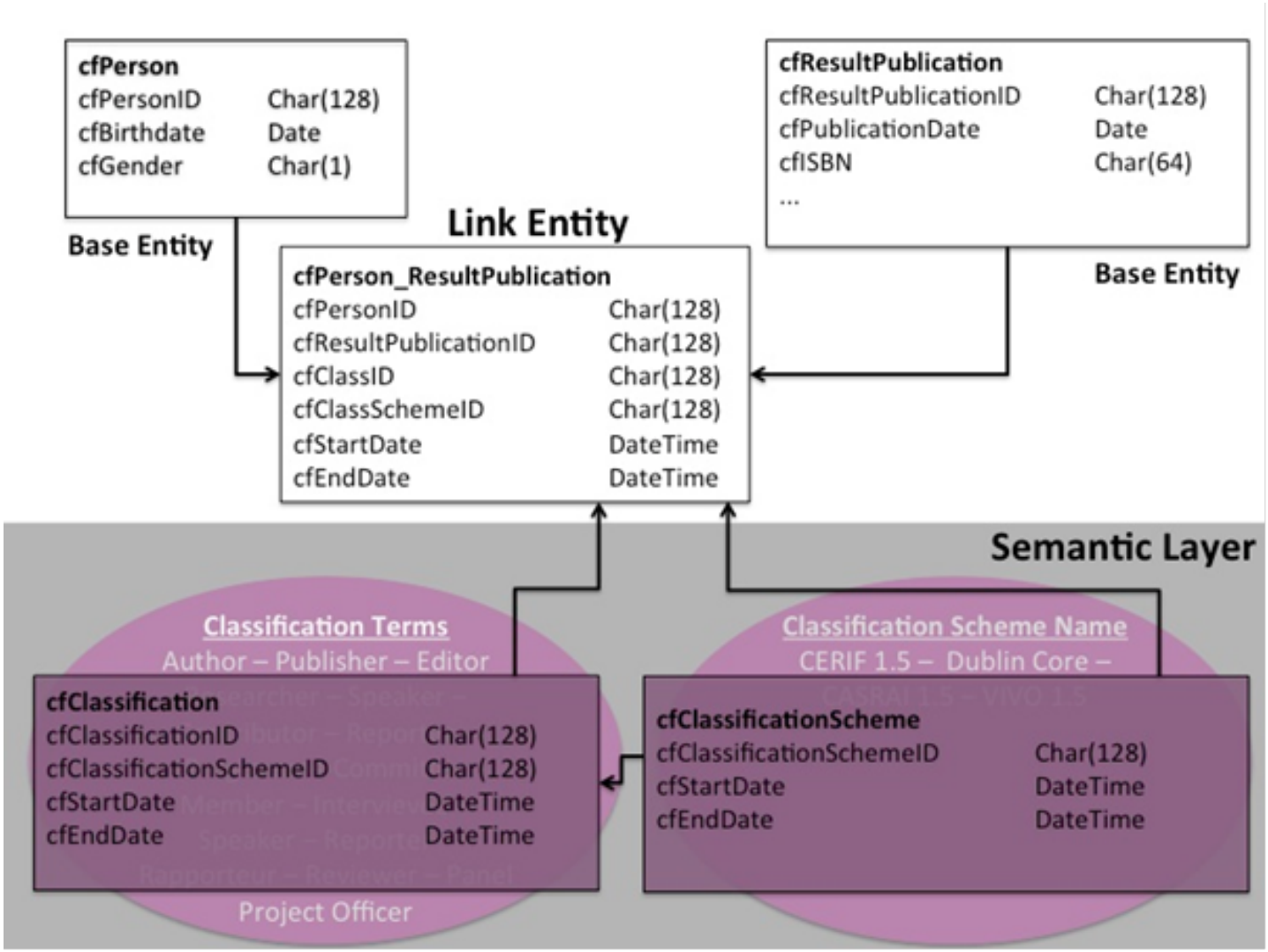
CERIF Features



CERIF Features: Semantic Layer

- The ‘role’ in link entities
- And restricted attribute value lists in base entities
- Stored in the semantic layer of CERIF using the usual linking relations technique
- And referenced from the main database
- ➔ ensures consistency: all semantics in one place
- ➔ allows semantic crosswalking between different schemes

CERIF Features



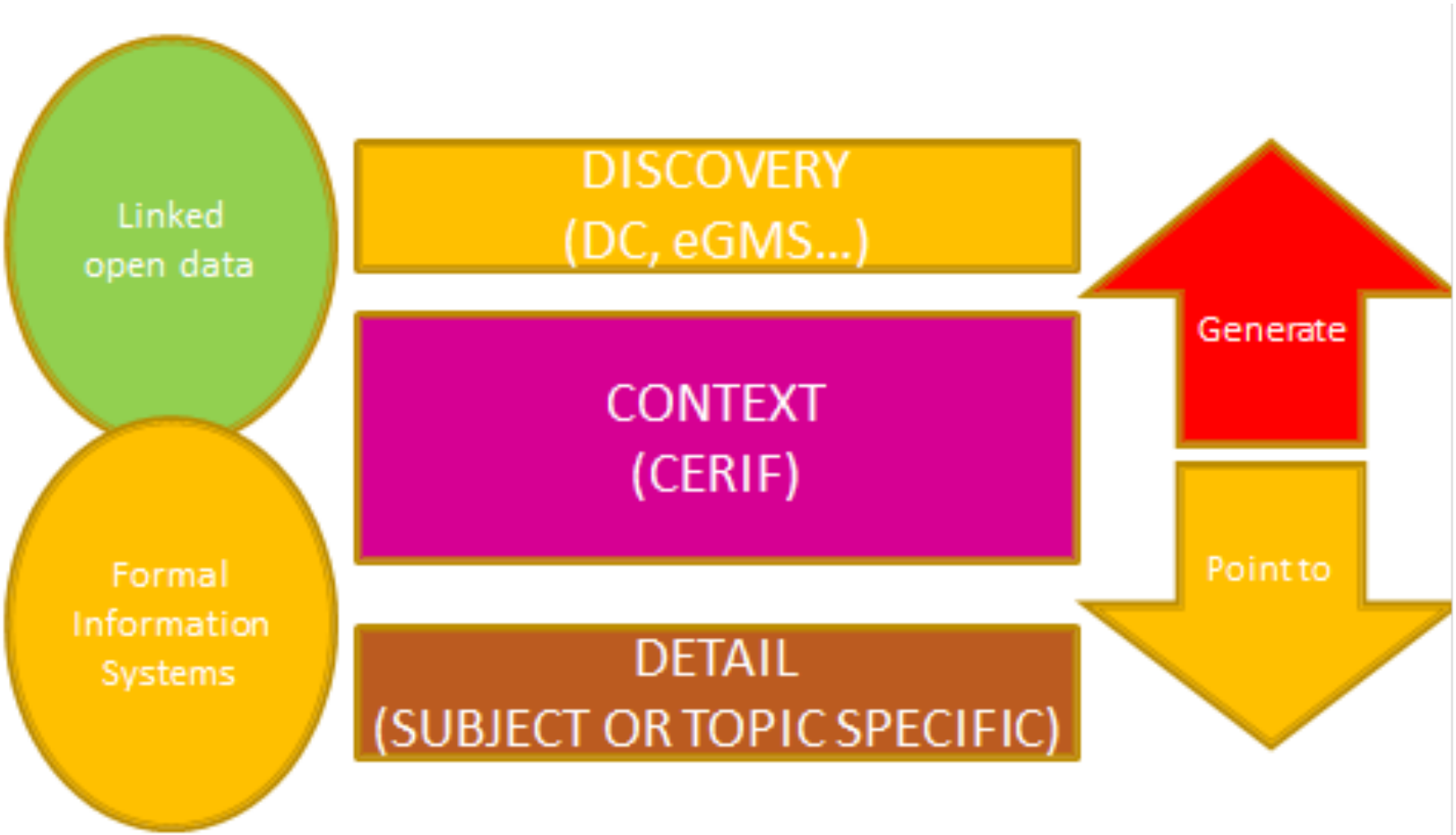
- Research and Research Information
- Metadata
- Problems with Metadata Formats
- CERIF
- A 3-layer Model for Metadata

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

Conclusion

- The 3 layer model for metadata developed within the ENGAGE project:
  - Brings together open government data with open research data
  - Brings together a LOD / semantic web environment with a more formal information processing environment
  - Provides the required metadata for all purposes