A multi-level metadata approach for a Public Sector Information data infrastructure

Nikos Houssos[1,2], Brigitte Jörg[1,3], Brian Matthews[4]

[1] euroCRIS
[2]National Documentation Centre, Greece
[3]German Research Center for Artificial Intelligence (DFKI), Germany
[4] Science and Technology Facilities Council, UK

11th International Conference on Current Research Information Systems (CRIS 2012), Prague, 6-9 June 2012

Agenda

- Introduction
- A 3-level metadata approach for PSI data sets
- Mapping to CERIF from current PSI metadata schemata
- Architecture of a metadata architecture for PSI datasets
- Publishing as Linked Open Data
- Summary – conclusions

Public Sector Information

- Data produced by governmental organisations – typically referring to datasets
- Examples: geospatial, demographic, statistical, environmental, public safety, financial data
- Growing international movement: open access to PSI datasets in a way that facilitates reuse
- Opening up PSI datasets can potentially lead to substantial economic gains

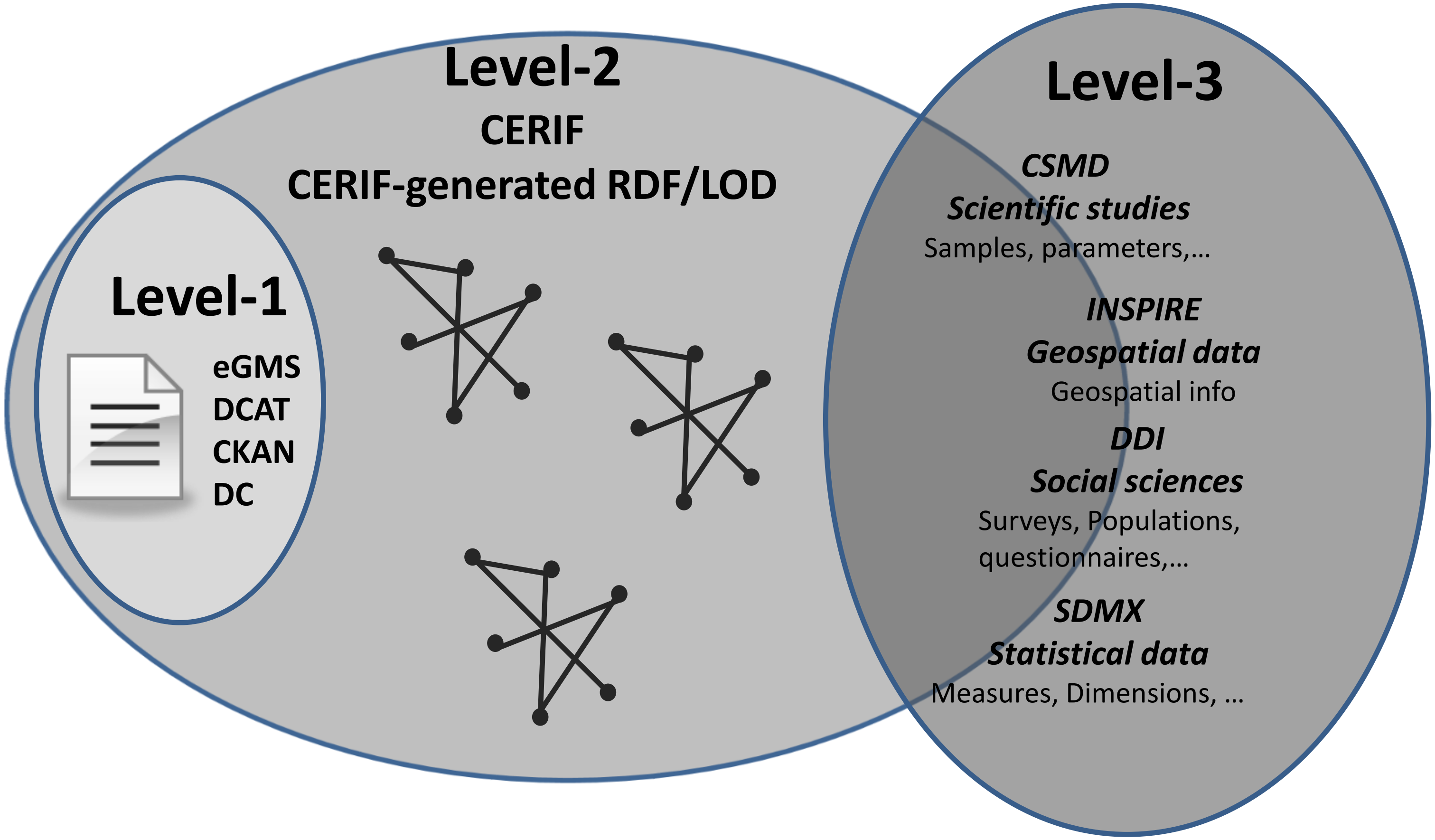
ENGAGE

- FP7-INFRASTRUCTURES project
- Subprogramme area:
  - INFRA-2011-1.2.2. Data infrastructures for e-Science
- Main goal: deployment and use of an **advanced service infrastructure**, incorporating **distributed** and **diverse public sector information resources** as well as **data curation**, **semantic annotation** and **visualisation** tools, capable of supporting scientific collaboration and governance-related research from multi-disciplinary scientific communities, while also empowering the deployment of **open governmental data** towards citizens

A 3-level metadata approach

- Level-1. Discovery metadata. Flat schemata (analogous to Dublin core). Enables basic search by non-sophisticated users.
- Level-2. Usage metadata. A structured, semantically-rich model for contextual metadata. Enables advanced domain-independent services.
- Level-3. Domain metadata. Detailed domain-specific metadata. Allows advanced services provided by specialised tools.

A 3-level metadata approach



Level-2

CERIF

CERIF-generated RDF/LOD

Level-1

eGMS
DCAT
CKAN
DC

Level-3

CSMD

Scientific studies

Samples, parameters,…

INSPIRE

Geospatial data

Geospatial info

DDI

Social sciences

Surveys, Populations, questionnaires,…

SDMX

Statistical data

Measures, Dimensions, …

Rich contextual metadata is important!

- Captures context, purpose, provenance, coverage, etc.
- Allows the user to:
  - Discover a dataset
  - Evaluate utility and re-use potential
  - Reuse it!
- Enables advanced services
  - Sophisticated search/discovery and navigation, mining, visualisation, reporting

Motivation

- Developing ENGAGE involves addressing a typical information integration problem with challenges across many dimensions
- Focus of the present contribution: How metadata is represented and managed

Design choices

- Level-2 is the appropriate level to do integration
- Expresiveness of metadata representation.
- “Lowest common denominator” vs. “Conceptual model”
- The “Conceptual model” approach is preferable since it avoids losing information when integrating information – at the cost of a more detailed mapping per input source

Selection of a global conceptual model

- Existing PSI dataset metadata schemata are not adequate for the task
- Two major candidates:
  - CERIF
  - A new OWL ontology

CERIF vs. OWL

- CERIF advantages:
  - No need to build model from scratch, reuse the CERIF entities and structures.
  - Ability to represent temporal aspects of relationships in a way that is simple and easy to implement. In RDF and OWL the respective feature is subject to ongoing research and not available in mainstream tools (e.g. triplestores, SPARQL endpoint implementations).
  - Maturity of the tools and platforms for the development of production-scale systems on top of relational databases. Tools and platforms for RDF/OWL are improving fast, but their relational databases counterparts are much more mature.
- OWL advantages:
  - Easier to provide the information as Linked Open Data.
  - Can inherently support inference.

CERIF as a conceptual model for PSI datasets metadata

- CERIF has significant advantages as the Level-2 canonical model for contextual metadata representation within ENGAGE
- But can CERIF represent PSI datasets metadata in its current form?

Mapping to CERIF from current PSI metadata schemata

- Selected schemata
  - CKAN. Used in the popular CKAN software platform. It is a simple, flat model that does not include capabilities for modelling complex linkages with entities in the context of datasets (e.g. persons, organisations, projects) and also lacks features to represent semantic relationships
  - eGMS. The UK e-Government Metadata Standard, an application profile of Dublin Core.
  - DCAT. An RDF Schema vocabulary for representing PSI data catalogues, currently being developed within the W3C. DCAT has a structure that is to a limited extent normalised. Not able to capture different roles/semantics in the relationships among entities.

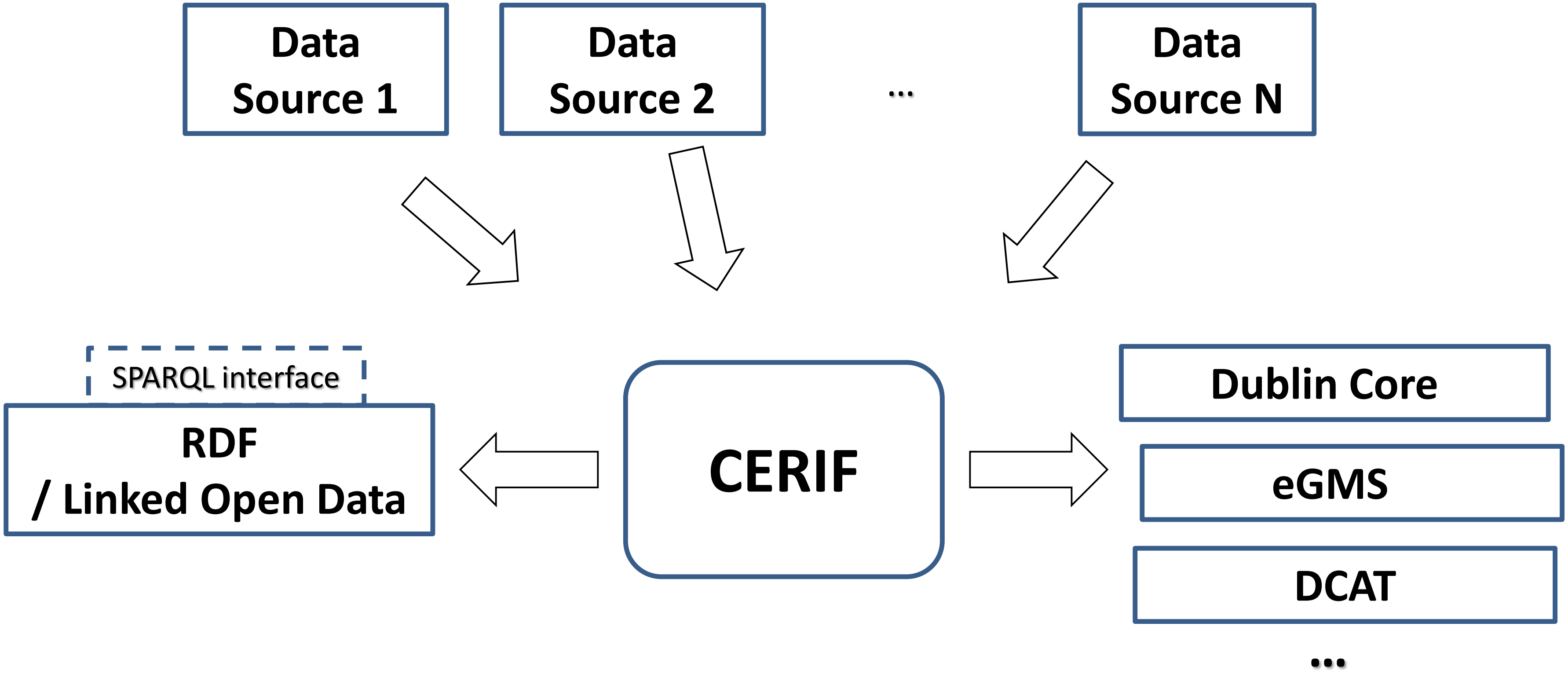
Overall mapping approach

- Datasets are modelled using the CERIF cfResultProduct entity
- Individual digital resources (e.g. files) are modelled using the CERIF cfMedium entity
- Entities such as APIs, web services, feeds, are modelled using the CERIF cfService entity
- The CERIF Semantic Layer is used for categorical data fields and for semantics of relationships

Conclusions from the mapping exercise

- Information in common PSI datasets metadata schemata can be represented in CERIF in a straightforward way, without loss of the semantics.
- Most required data elements can be represented in CERIF as relationships or classifications, without the need for new, explicit data fields. This is mainly due to the flexibility of the CERIF Semantic Layer.
- The inherent highly normalised, graph-based structure of CERIF avoids significant limitations in simple, flat models.

An architecture for PSI metadata



Important aspects

- Publishing Linked Open Data
  - Straightforward with CERIF
  - The euroCRIS LOD Task Group is defining standard ways to achieve this
  - Successful use case of CERIF generating LOD in the VOA3R project.
- Linking with Level-3 metadata
  - Specialised, domain specific standards like CSMD, SDMX, DDI, EML, INSPIRE will be used for Level-3
  - A part of the Level-3 metadata schema may concern contextual metadata that can be represented in CERIF at Level-2 and be used for domain-independent services

Summary

- A 3-level approach to managing metadata in a PSI infrastructure has been adopted by the ENGAGE project.
- “Conceptual model” approach to information integration.
- Domain-independent integration at Level-2 using CERIF as the canonical model for contextual metadata.
- Generation of Linked Data and simple, common schemata from CERIF.
- It has been demonstrated that CERIF is able to represent PSI datasets metadata
  - Detailed mapping has been performed from major PSI data models to CERIF
- The proposed architecture to be implemented and evaluated within ENGAGE within the next two years.

Thank you for your attention!

- More info:

nhoussos AT ekt.gr

brigitte.joerg AT gmail.com

brian.matthews AT stfc.ac.uk