Houssos, Nikos; Jörg, Brigitte; Matthews, Brian: *A multi-level metadata approach for a Public Sector Information data infrastructure*. In: Jeffery, Keith G; Dvořák, Jan (eds.): E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production: Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic). Pp. 19-31. ISBN 978-80-86742-33-5. Available from: www.eurocris.org.

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

Nikos Houssos[a b], Brigitte Jörg[a c], Brian Matthews[d]

[a] euroCRIS; [b] National Documentation Centre / National Hellenic Research Foundation, Greece; [c] German Research Center for Artificial Intelligence, Germany; [d] Science and Technology Facilities Council, UK

Summary

This paper describes an approach for representing and handling metadata about Public Sector Information data sets in a large scale data infrastructure as designed within ENGAGE, a project of the FP7 Research Infrastructures programme. A multi-level approach is adopted, allowing management of metadata at various levels of expressive power, and thus enabling different use cases and requirements to be served through a single platform. CERIF is being investigated as a common conceptual model to ensure information integration from diverse sources without loss of meaning and furthermore as the basis for the generation of Linked Open Data. Through detailed mappings from common metadata schemata used for PSI it is shown that CERIF is a data model suitable for representing rich contextual metadata for the domain of governmental datasets.

1 Introduction

Governmental organizations around the world continuously produce a vast array of datasets of diverse types and thematic domains, commonly referred to as Public Sector Information (PSI). Indicative examples include geospatial, demographic, statistical, environmental, public safety and financial data. There is a growing international movement towards providing this wealth of information publicly in a way that enables re-use. The primary motivation is that PSI can potentially be utilised in a large variety of commercial and non-profit activities leading to considerable economic gains; a recent study estimates this at EUR 40 billion per year in the EU (Vickery, 2011). Examples of such activities include the creation of new products and services as well as the support of research. The latter is of significance, since the available data can have a pivotal role in the advancement of research, from social and economic sciences to natural sciences and engineering.

The EC FP7 ENGAGE project (ENGAGE, 2011) is a major effort in this field with the goal of ENGAGE is deploying and using an advanced service infrastructure, incorporating distributed and diverse public sector information resources. This will be augmented with data curation, semantic annotation and visualisation tools, to support scientific collaboration and governance-related research from multi-disciplinary communities, while also enabling the deployment of open governmental data towards citizens. An important factor in providing open governmental data is the metadata that accompanies the dataset. Without appropriate metadata, it is difficult for users to discover a dataset, evaluate its utility and reuse potential and ultimately re-use it. Therefore, the representation and management of metadata is one of the main issues addressed by ENGAGE.

The present paper concerns the metadata approach adopted for the ENGAGE infrastructure. Three levels of metadata are identified, accommodating different usage requirements. At the core of the PSI metadata is a conceptual model for datasets, based on the EU recommended CERIF specifica-

tion[1] (Jörg, 2010), able to provide rich contextual information and to capture the semantic relationships of datasets with each other and with other entities.

The rest of the paper is structured as follows: Section 2 provides an overview of the proposed multi-level metadata approach, presents the rationale for the overall solution and key design choices and elaborates on the selection of a global conceptual model for PSI metadata that led to a CERIF-based architecture that is fully compatible with Semantic Web technologies and the Linked Open Data initiative. Section 3 presents a detailed effort to investigate whether information captured by existing PSI metadata schemata can be represented by CERIF, and Section 4 discusses some architectural issues.. The paper concludes with a summary of its main contribution

2 A 3-level metadata approach for PSI datasets

The task of representing, in a unified way, metadata about PSI data sets at the level of heterogeneity envisioned for ENGAGE is challenging, since government data about a diverse range of topics could be expected to be made available within the ENGAGE platform. The approach adopted for handling metadata is therefore based on a three-level scheme. Varying degrees of detail and the need to address different requirements are reflected in these discrete levels, as follows:

Level-1. **Discovery Metadata.** Simple, 'flat' metadata schemata, analogous to Dublin Core. This level of information is useful to assist non-sophisticated users to perform basic searches and find data sets using a very limited and easy to learn vocabulary ("metadata pidgin") (Baker, 2000). Examples of such schemata are Dublin Core (plain and qualified), eGMS, DCAT (Data Catalog Vocabulary) and data models used by software platforms like CKAN. Schemata of this type are relatively easy to populate for a data set and simple to understand and use for basic discovery services. However, they do not capture well semantic interrelations among entities. They are used to form catalogues with one "catalogue card" describing each data set and constitute a common denominator for information about data sets, leading to a loss of semantic information when integrating data from heterogeneous sources.

Level-2. **Usage Metadata.** A structured, linked entity model for contextual metadata, able to capture the semantic relationships of data sets with each other and with other entities (e.g. persons, organisations, documents, activities, funding sources) as well as data set classifications. This enables the representation and reuse of semantically well-defined information about a data set's provenance, purpose, coverage, etc. This level of metadata allows functions and services provided over data sets such as search and discovery, visualisation, navigation and browsing, mining, analytics and reporting to be available for more detailed analysis by end users. A formal conceptual model is needed for this level, able to represent the concepts and relationships of interest to applications, so that integration does not lead to loss of information and semantic ambiguity.

Level-3. **Domain Metadata.** Detailed metadata standards for data sets of particular types or domains (e.g. CSMD for scientific data sets (Matthews et. al. 2009), SDMX for statistical data, (SDMX 2011), INSPIRE for geospatial data[2], the Data Documentation Initiative (DDI) for social science data[3]). These can be used for advanced domain-specific services and tools that can be provided for particular categories of data sets.

[1] CERIF releases are available at http://www.eurocris.org/Index.php?page=CERIFreleases&t=1.

[2] http://inspire.jrc.ec.europa.eu/

[3] Data Documentation Initiative: http://www.ddialliance.org/

Houssos, Nikos; Jörg, Brigitte; Matthews, Brian: *A multi-level metadata approach for a Public Sector Information data infrastructure*. In: Jeffery, Keith G; Dvořák, Jan (eds.): E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production: Proceedings of the 11th International Conference on Current Research Information Systems (June 6-9, 2012, Prague, Czech Republic). Pp. 19-31. ISBN 978-80-86742-33-5. Available from: www.eurocris.org.

A crucial aspect of defining the metadata architecture of a platform aggregating PSI datasets is the selection of the appropriate conceptual model for Level-2, usage and contextual metadata. This model should be able to express the metadata for datasets originating from the wide range of sources that will be used to feed the ENGAGE system. Initial experiments have shown that CERIF is an appropriate data model for this purpose. Furthermore, the contextual CERIF metadata can be exposed as RDF linked data (Berners-Lee 2006, Bizer et al. 2009); this is accomplished via automatic generation of RDF from CERIF. A simplified illustration of the 3-level metadata approach is provided in Figure 1.

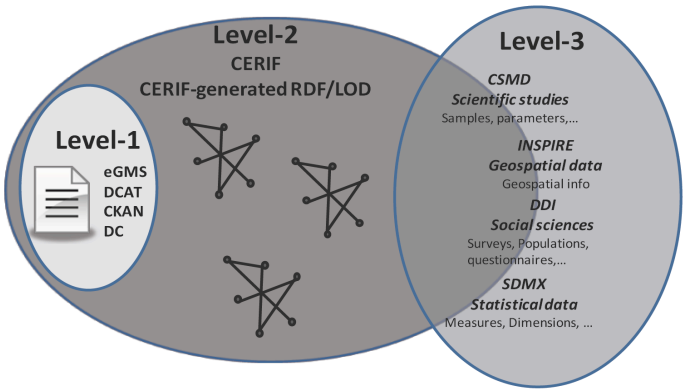

Figure 1. Overview of the 3-level Metadata Model

2.1 Rationale and design choices

The rationale for this approach to metadata is based on the observation that this case is a typical information integration system with a large and diverse set of source systems and schemata. The design of such a system is a complex task, involving issues such as the selection of a common data model, the approach for the specification of mappings (e.g. Local-as-View or Global-as-View), way of integration (e.g. virtual or materialized), data cleaning and de-duplication, query language and query processing. The present paper concentrates on the problem of identifying how metadata is represented and managed, a challenging undertaking given the heterogeneity of the datasets to be included in ENGAGE.

A major design decision relates to the expressiveness of the metadata representation. Two different approaches can be followed:

- A "Lowest common denominator" approach. It defines a common metadata schema that is significantly limited in terms of representation capabilities and tries to capture per dataset only a subset of the information that is available for it at its source. This approach typically relies on simple, flat schemata like Dublin Core and in many cases is adequate for supporting basic discovery functions for end-users. The main advantage is the ease of initial implementation and low cost for the introduction of each new source (e.g. in terms of required mapping efforts). The major disadvantage is that information is lost; only a part of the metadata (and associated semantics) of the dataset survives the integration into the aggregating system.

  Examples of information that are typically not represented in the common schema under such an approach are the semantic roles in relationships among entities and associated timestamps. For instance, if three organizations are related to a dataset in the integrated database, neither

the role of each (e.g. creator, owner, maintainer) is available nor the temporal dimension of the relationship (e.g. an organisation maintained a dataset from 01-01-2009 until 31-Dec-2011). Furthermore, structured, normalized data about different entities can be lost due to the "flattening" of the central metadata schema – for example data about an organisation might need to fit into a single field, so it might not be possible for data elements about the organisation like URL, postal address and contact details to be included.

- The conceptual model approach. It involves the definition of a core conceptual model - sometimes termed "property-centric ontology" (Doerr, 2003) or "enterprise model" (Wiederhold, 1992) - that is able to represent the concepts of interest to data sets applications and, importantly, their relationships in a semantically clear way (Calvanese, 2009, Doerr, 2003). Every data source is typically expressed in terms of the canonical conceptual model (in a way somewhat analogous to the Local-as-View approach to mapping specification) so that any query formulated with the common ontology can be answered by all sources, with the replies being also represented in terms of the canonical schema. This way, information meaning is not lost for end users despite the heterogeneity of the sources, at the cost of a detailed mapping of the data source schema to the core model.

The method proposed by the ENGAGE project is the conceptual model approach, which enables this pan-European PSI infrastructure to include rich metadata and thus increase the potential to support sophisticated value-added services to end-users. However, flat metadata for basic discovery services (Level-1) will also be generated and be provided by the platform.

Based on the above discussion, it becomes obvious that the appropriate level for the integration of datasets metadata is Level-2. Level-1 provides only simple, flat metadata, while Level-3 provides subject- or type-specific information for which horizontal, domain-independent integration across disciplines is extremely difficult, if not impossible. Thus Level-3 metadata is reserved for more detailed analysis and exploration once suitable datasets have been identified. The common ENGAGE platform should allow access to such metadata, but once detailed analysis is started, the user should be handed on to domain specific tools.

2.2 Selection of a global conceptual model

A key aspect of defining the metadata architecture of a data integration platform for PSI is the selection of the appropriate conceptual model for Level-2 metadata. An effort was made to identify existing specifications within the PSI domain that could be re-used. Unfortunately, current schemata already available in the vertical domain of governmental datasets are simple, non-normalized, without sophisticated capabilities for capturing semantic relationships among entities. Thus, they cover only Level-1 in our metadata architecture and are not adequate for implementing the proposed approach. Therefore, two other candidates were identified:

a) The EU recommended CERIF data model (Jörg, 2010), developed by euroCRIS. CERIF has many features that make it attractive for serving as a conceptual modelling vehicle for contextual metadata about PSI datasets. CERIF is a normalized, graph based data model. It is defined as a conceptual model using Entity-Relationship modelling. Logical / physical representations are provided for common relational database systems. A dataset in CERIF is naturally represented as a ResultProduct, while entities that need to be connected with datasets are also modelled in CERIF; these include persons, organizations, projects (covering also the concept of activity) and funding. CERIF has built-in support for multi-lingual metadata values and associating entities with spatial and geolocation information. More importantly, CERIF has an

explicit, generic mechanism, called the CERIF Semantic Layer, which allows for the representation of relationship kinds (Storey 1993, Wand et. al. 1999), application views or domain-centred classification schemes. In CERIF, the so-called link entities represent relationships, for example among person-dataset, organization-dataset, person-organisation. The semantics of each relationship are captured by a so-called class term (essentially, a label with declared semantics) that belongs to a classification scheme (e.g. vocabulary, thesaurus, taxonomy). Each relationship instance additionally contains temporal information (e.g. a time interval specification). Furthermore, any classification scheme (flat, hierarchical or graph based) can be imported into a CERIF based repository or information system.

b) A newly built model based on Semantic Web languages, either a vocabulary defined with the RDF Schema or an OWL ontology that makes use of the advanced features of these modelling technologies. RDF statements constitute directed labelled graphs that can represent any information. It provides an open-ended view of the world and allows inference of new information from existing facts. OWL is a richer language, used for defining ontologies. It is based on description logics and builds on RDF to provide significantly more facilities for modelling, albeit at the cost of computational complexity and non-decidability when the more advanced features of OWL are utilised.

The following comparison points resulted from the evaluation of the two candidate technologies:

- CERIF is an existing, established specification and it is being used by hundreds of CRIS systems internationally. OWL is a generic technology that can be used to define any data model/ontology; however, for PSI there is no internationally adopted ontology in OWL, therefore a new one would need to be built from scratch if OWL is selected.
- An important strength of CERIF is the ability to easily specify properties on relationships. A example of this is the attachment of temporal information on semantic relationships. For instance, it is possible in CERIF to state that an organisation was the maintainer of dataset from 01-Dec-2005 to 15-Jan-2010 or that the creation of a dataset was funded by a project within a specific time interval. These features allow temporal queries on the dataset metadata using technology which is built into CERIF and used in implementations. This information can be also represented in RDF (Gutierrez, 2007), although in non-trivial ways, potentially leading to a proliferation of complex statements that make it hard for humans to understand and write software programs that process them. Moreover, temporal features of RDF and relevant extensions to OWL and the SPARQL query language are still a research topic (Perry, 2011) that has been adopted in W3C recommendations and implementations of semantic web platforms.
- A major strength of using Semantic Web technologies is the potential of providing PSI as Linked Open Data (LOD).. By publishing data on the Web according to the Linked Data principles, data providers add their information to a global data space, which allows data to be discovered, navigated and re-used by human users and by software applications. Publishing Linked Open Data is a principal requirement for a system aggregating PSI and providing value-added services such as the ENGAGE platform.
- CERIF-based systems typically run on relational databases which are the most common data management technology. Consequently, a wealth of extremely mature and robust software tools are available for the creation and maintenance of systems with RDBMS back-ends, enabling rapid creation and easy maintenance of applications. Furthermore, vast experience and numerous techniques for ensuring proper support for non-functional requirements (e.g. performance/scalability, security) exist in the industry and academia. Tools and platforms for

developing applications with Semantic Web technologies using triple stores and SPARQL endpoints for querying them are constantly improving; however, their maturity cannot be compared with their counterparts in relational databases.

This comparison and analysis led to the approach described in Section 2.1 for Level-2 metadata, which essentially uses CERIF as the global conceptual model of the PSI platform, but also specifies that the metadata in the platform is published as LOD and exposed via a SPARQL endpoint.

3 Mappings to CERIF from current PSI metadata schemata

A detailed exercise was undertaken to validate this approach, Major data schemata in the PSI domain were mapped to CERIF to ensure that the latter has the necessary expressiveness to cover the governmental datasets metadata. The following PSI data schemata were mapped to CERIF:

- The **e-Government Metadata Standard (eGMS[4])** is the UK e-Government Metadata Standard. It defines how UK public sector bodies should label content such as web pages and documents in order to make such information more easily managed, found and shared. eGMS is an application profile of the Dublin Core Metadata Element Set and consists of mandatory, recommended and optional metadata elements.
- The **Data Catalog Vocabulary (DCAT[5])** is an RDF Schema vocabulary for representing PSI data catalogues, currently being developed within the W3C Government Linked Data Working Group,. DCAT has a structure that is to a limited extent normalised (e.g. persons and organisations are modelled as structured entities separate from the dataset). However it is not able to capture different roles/semantics in the relationships among entities. For example, an organisation or person can only be the publisher of the dataset.
- The **CKAN Domain Model[6]** is used in the popular CKAN software platform, which enables easy creation of PSI portals for publishing datasets. 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.

Overall, this mapping exercise showed that the information in these data models can be represented in CERIF in a straightforward way, without loss of the semantics. This is mainly due to the capability of CERIF to represent relationships among entities with declared semantics and temporal information, while maintaining a high degree of extensibility through the Semantic Layer. This enables CERIF to represent rich sets of metadata as relationships or classifications, without the need of explicit addition of new fields in the model.

The detailed mappings of these schemata to CERIF are shown in the tables given in the Appendix. The names of the CERIF entities in the tables follow the more concise names used at the physical CERIF level for brevity. Due also to the economy of presentation, the mapping from eGMS to CERIF is given only for the top-level eGMS elements (not their refinements).

We discuss in the rest of this paragraph some important aspects of the mappings. Datasets are modelled using the CERIF entity cfResultProduct (cfResProd). Individual digital resources (e.g.

[4] e-Government Metadata Standard: http://www.esd.org.uk/standards/egms/

[5] http://www.w3.org/TR/vocab-dcat/

[6] http://docs.ckan.org/en/latest/domain-model.html

files) are represented using the cfMedium entity of CERIF. Entities such as APIs, web services and feeds are mapped to the Service entity (cfSrv). Data about organisations and persons is naturally represented by the respective CERIF entities (cfOrgUnit and cfPers). Categorical data fields, which take values from a controlled vocabulary, are modelled using the CERIF Semantic Layer: each vocabulary is a classification scheme containing classification terms (instances of the CERIF entity Class). For example, the Status field of eGMS is represented by classification terms (cfResProd_Class); each possible status value is a separate classification term that belongs to the Status classification scheme.

Many data fields in current Level-1 PSI schemata can be modelled in CERIF by relationships with declared semantics. For example fields like author, maintainer, and publisher are mapped in CERIF to relationships among datasets and persons or organisations. Notably, the normalised structure of CERIF avoids the inelegant approach of CKAN for capturing other information (e.g. email) about persons. In particular, in CKAN there are separate fields within the Dataset entity for author_email and maintainer_email. In CERIF theses fields belong to the respective entities (cfPers and cfOrgUnit), connected with datasets via CERIF link entities with appropriate classification terms (e.g. author, maintainer, etc.) This also holds for certain dates; for example the release date in DCAT is modelled as a time-stamped relationship between the dataset and the "agent" (organisation or person) performing the release.

4 Developing a Metadata Infrastructure for PSI

In this section we briefly outline an architecture using this approach within a common portal infrastructure. The realisation of this architecture is the subject of ongoing work within ENGAGE.

4.1 An architecture for integrating PSI metadata

The proposed architecture is outlined in Figure 2. The CERIF-based Level 2 metadata forms the core model, each of the external PSI data sources are analysed, and mapped into this core model. Data collection from sources will be performed in a variety of ways and workflows (e.g. harvesting, submission) and may be followed by curation within the ENGAGE system; these aspects are currently under investigation and beyond the scope of the present article. Metadata that could be available from sources may include schemata like eGMS, DCAT, CKAN, Dublin Core or custom data representations. In certain cases Level-3 metadata might be provided by sources, probably containing some contextual, domain-independent information that can be mapped to CERIF.

The CERIF model forms the basis of searches and explorations of the data. Level 1 metadata can be generated from CERIF in common standards such as Dublin Core or eGMS, to allow simple searches and explorations. The core metadata can also be exposed as Linked Open Data to allow access via SPARQL end points, allowing third party tools to explore the data. This core metadata can also be mapped into parts of Level 3 metadata to allow deeper exploration of the data with domain-specific tools.

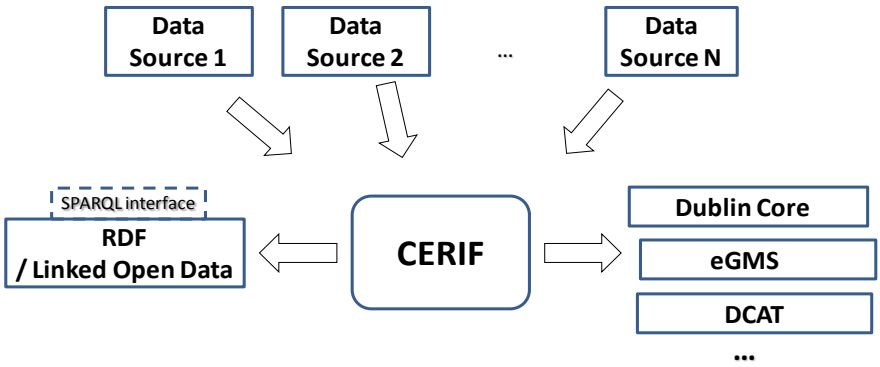

Figure 2. Overview of metadata representation and management approach

4.2 Publishing as Linked Open Data

Linked open data is becoming a widely used to publish and access data and metadata on the web using semantic web technologies. LOD is based on some simple principles (Berners-Lee, 2006, Bizer et al. 2009): assigning dereferenceable stable URIs to data sources, publishing metadata in RDF; and providing SPARQL endpoints for access. In order to support the use of LOD within the common infrastructure, we need to: use a stable URI scheme to identify entities within the CERIF model; and develop an exchange layer on the RDBMS implementation of CERIF to allow querying of the metadata using SPARQL and the delivery of the metadata in RDF.

CERIF is compatible with Linked Open Data, since the structure of the link entities and the semantic layer as well as the URI identifier with every research entity enables a straightforward publishing of data from a CERIF database according to the LOD principles. A well-known case where this has been achieved is the VOA3R project, where a CERIF back-end has been exposed as LOD (Jörg et al., 2012). A new task group within euroCRIS has been established, investigating a standard way of providing CERIF metadata as LOD.

Tools are available to support the publishing of data from RDBMSs as LOD or provide virtual representations. For example, tools such as Virtuoso[7] provides a data server to support publishing of relational and other data. The specification of a standard language (R2RML[8]) for defining mappings from relational databases to RDF datasets is being developed by the W3C. Most tools for publishing RDBMSs as LOD are already supporting R2RML or have plans to do so.

4.3 Mapping to domain specific metadata schema

As discussed above, specific data sets will typically have detailed metadata in domain specific formats. Thus, in order to allow detailed exploration of the data, there needs to be a hand over between "level-2" metadata and "level-3" metadata in the above architecture, and therefore there needs to be an exchange of metadata between the common CERIF based system to the domain specific system of the metadata for the identified sets. The Level-3 schema will have more de-

[7] http://virtuoso.openlinksw.com/

[8] http://www.w3.org/TR/r2rml/ . A candidate recommendation was released in February 2012

tailed information than CERIF; for example CSMD has details on instruments, samples and parameters used within an experiment, DDI about study populations and questionnaires, and INSPIRE about geographical features and their physical parameters. These are out of scope for CERIF. However, all these metadata formats will typically have core components, describing the organisational context of the data (e.g. people-projects-organisations). Thus a partial mapping can be undertaken between CERIF and these core components, to enable the transfer between the common and the domain specific tools.

5 Summary

The representation and management of metadata is a particularly challenging issue when building a data infrastructure for Public Sector Information, aiming to aggregate datasets from a wide range of heterogeneous sources and build value-added services upon them. Given the inherent complexity of specifying and providing metadata about datasets and potentially combining it with the actual data, an approach based on a single metadata representation model is not adequate: a multi-level approach has therefore been proposed, to support different requirements and use cases. A key design choice is the horizontal, domain-independent data integration using CERIF as the *canonical* model for contextual metadata, combined with the generation of Linked Data from CERIF and the linkage with more detailed, specialized and probably domain-specific data model and tools. The suitability of CERIF for representing PSI metadata has been demonstrated through detailed mappings from major formats that are currently in use in the governmental datasets area; it has thus been described in the role of the current approach as a metadata infrastructure for PSI. In the next phase of the ENGAGE project, this infrastructure will be realised and evaluated.

Acknowledgements

The work presented in this paper has been partly supported by the ENGAGE Integrated Project (Ref No: 283700) of the EU-funded FP7-INFRASTRUCTURES Programme. The authors wish to thank Keith Jeffery, STFC, with whom they cooperated within ENGAGE on defining the 3-level metadata approach and the project partners for fruitful discussions regarding technical issues relevant to the content of the article.

References

Baker, T. (2000). A Grammar of Dublin Core. D-Lib Magazine 6(10): 3.

Berners-Lee, T. (2006). Linked Data - Design Issues. Retrieved 10 April 2012, http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C.,Heath, T., and Berners-Lee, T. (2009). Linked data – the story so far, International Journal on Semantic Web and Information Systems. 5(3), 1–22.

Calvanese, D., G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati (2009). Conceptual modeling for data integration. In Conceptual Modeling: Foundations and Applications, LNCS, Springer.

Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI Mag. 24 (3), 75-92.

ENGAGE Project (2011), www.engage-project.eu.

Gutierrez, C., C. Hurtado, and A. Vaisman (2007). Introducing time into RDF. IEEE Transactions on Knowledge and Data Engineering 19(2), 207 - 218.

Jörg, B. (2010). CERIF: The Common European Research Information Format Model. Data Science Journal, Volume 9, 24-31.

Jörg, B., I. Ruiz-Rube, M. Sicilia, J. Dvorak, K. Jeffery, T. Höllrigl, H. S. Rasmussen, T. Vestdam, E. G. Barriocanal (2012). Connecting Closed World Research Information Systems through the Linked Open Data Web**.** International Journal of Software Engineering and Knowledge Engineering, Volume 22. World Scientific.

Matthews, B., Sufi, S., Flannery, D., Lerusse, L., Griffin, T., Gleaves, M., Kleese, K. (2009). Using a Core Scientific Metadata Model in Large-Scale Facilities. 5th International Digital Curation Conference (IDCC 2009), London, UK, 02-04 Dec 2009

Perry, M., P. Jain, and A. P. Sheth (2011). SPARQL-ST: Extending SPARQL to support spatio-temporal queries geospatial semantics and the semantic web. Volume 12 of Semantic Web and Beyond, Chapter 3, pp. 61-86. Boston, MA: Springer US.

SDMX (2011). Statistical Data Metadata Exchange, v.2.1 (April 2011), http://sdmx.org/.

Storey, V.C. (1993) Understanding Semantic Relationships. The International Journal on Very Large Databases (VLDB). Volume 2, Number 4, October 1993, pages 458-488, Springer Berlin-Heidelberg.

Wang, R.Y., Storey, V.C., & Weber R. (1999) An ontological analysis of the relationship construct in conceptual modeling. ACM Transactions on Database Systems (TODS) Journal, Vol. 24, Issue 4, December 1999, pp. 494-528.

Vickery, G. (2011): Review of recent studies on PSI re-use and related market developments.

Wiederhold, G. (1992). Mediators in the architecture of future information systems. Computer 25 (3), 38-49.

Contact Information

Nikos Houssos
National Documentation Centre / National Hellenic Research Foundation (EKT/NHRF)
48 Vas. Kon/nou Av
116 35 Athens
Greece
nhoussos@ekt.gr

Brigitte Jörg
German Research Center for Artificial Intelligence (DFKI GmbH)
Alt-Moabit 91c
D-10559 Berlin
Germany
brigitte.joerg@dfki.de

Brian Matthews
STFC Rutherford Appleton Laboratory
Harwell Oxford
Chilton Didcot
Oxfordshire
OX11 0QX
United Kingdom
brian.matthews@stfc.ac.uk

Appendix: Tables Mapping PSI Metadata into CERIF

e-Government Metadata Standard (eGMS)	CERIF
Accessibility	cfResProd_Class, cfPers_ResProd, cfOrgUnit_ResProd, etc.
Addressee	cfPers_ResProd with appopriate label
Aggregation	cfResProd_ResProd
Audience	cfResProd_Class or cfOrgUnit_ResProd with Audience label
Contributor	cfPers_ResProd with appopriate label
Coverage	For spatial coverage: cfResProd_GeoBBox (upcoming, detailed geodata) or cfResProd_Class (for controlled location lists) For temporal coverage (date intervals) cfOrgUnit_ResProd or cfPers_ResProd with timestamps
Creator	cfPers_ResProd with appopriate label
Date	cfPers_ResProd with Creator label and timestamp
Description	cfResProdDescr
DigitalSignature	cfResProdInternId (to be re-investigated in upcoming release of CERIF having new mechanisms for representing identifiers)
Disposal	cfResProd_Class, cfResProd_Srv, cfPers_ResProd, cfOrgUnit_ResProd, etc.
Format	cfResProd_Class
Identifier	cfResProd.cfResProdId or cfResProd.cfURI
Language	cfResProd_Lang (considered for upcoming CERIF release)
Location	cfResProd_PAddr
Mandate	cfResProd_ResPubl
Preservation	cfResProd_Class
Publisher	cfOrgUnit_ResProd or cfPers_ResProd with appropriate label
Relation	cfResProd_ResProd
Rights	cfResProd_ResPubl, cfPers_ResProd, cfOrgUnit_ResProd, etc.
Source	cfResProd_ResProd, cfResProd_ResPubl, etc.
Status	cfResProd_Class
Subject	cfResProd_Class, cfResProdKeyw
Title	cfResProdName
Type	cfResProd_Class

Table 1. Mapping of eGMS to CERIF.

Data Catalog Vocabulary (DCAT)	CERIF
dcat:Dataset	***cfResultProduct***
Dataset.modificationDate	cfPers_ResProd or cfOrgUnit_ResProd (lastModifiedBy class with timestamp)
Dataset.title	cfResProdName
Dataset.description	cfResProdDescr
Dataset.publisher	cfPers_ResProd or cfOrgUnit_ResProd (role "publisher")
Dataset.releaseDate	cfPers_ResProd or cfOrgUnit_ResProd (releasedBy class with timestamp)
Dataset.frequency	cfResProd_Class with appropriate controlled vocabulary
Dataset.Identifier	cfResProd.cfResProdId
Dataset.spatialCoverage	cfResProd_Class with appropriate controlled vocabulary or cfResProd_GeoBBox for more precise spatial specification
Dataset.temporalCoverage	cfOrgUnit_ResProd or cfPers_ResProd with timestamps
Dataset.language	cfResProd_Lang (considered for upcoming CERIF release)
Dataset.license	cfResPubl_ResProd (link to license document)
Dataset.granularity	cfResProd_Class with appropriate controlled vocabulary
Dataset.dataDictionary	cfResPubl_ResProd (it is a document)
Dataset.dataQuality	cfResProd with CERIF Indicator and Measurement Entities
Dataset.category	cfResProd_Class with appropriate controlled vocabulary
Dataset.keyword	cfResProdKeyw
Dataset.relatedDocuments	cfResPubl_ResProd
Dataset.datasetDistribution	cfResProd_Srv
dcat:CatalogRecord	***cfResultProduct***
CatalogRecord.listingDate	cfPers_ResProd or cfOrgUnit_ResProd (listedBy class with timestamp)
CatalogRecord.modificationDate	cfPers_ResProd or cfOrgUnit_ResProd (lastModifiedBy class with timestamp)
CatalogRecord.dataset	cfResProd_ResProd
dcat:Catalog	***cfResultProduct***
Catalog.homepage	cfResProd.cfURI
Catalog.publisher	cfPers_ResProd or cfOrgUnit_ResProd (lastModifiedBy class with timestamp)
Catalog.spatialCoverage	cfResProd_GeoBBox (upcoming, detailed geodata) or cfResProd_Class (for controlled location lists)
Catalog.themes	cfResProd_Class with appropriate controlled vocabulary
Catalog.title	cfResProdName
Catalog.description	cfResProdDescr
Catalog.language	cfResProd_Lang (considered for upcoming CERIF release)
Catalog.license	cfResPubl_ResProd (link to license document)
Catalog.dataset	cfResProd_ResProd
Catalog.catalogRecord	cfResProd_ResProd
dcat:Distribution	***cfMedium***
Distribution.accessURL	cfMedium.cfURI
Distribution.size	cfMedium.cfSize
Distribution.format	cfMedium.cfMimeType
dcat:Download	***cfMedium***
Download.accessURL	cfMedium.cfURI
dcat:Feed	***cfSrv, cfSrv_Medium***
dcat:WebService	***cfSrv, cfSrv_Medium***
dcat:Category and category scheme	***cfClass, cfClassScheme***
dcat:Organisation	***cfOrgUnit***
dcat:Person	***cfPers***

Table 2. Mapping of DCAT to CERIF.

CKAN domain model		CERIF
Dataset		***cfResultProduct***
	id: unique id	cfResProd.cfResProdId
	name: unique name that is used in urls and for identification	cfResProdInternId
	title (dc:title): short title for dataset	cfResProdName
	url (home page): home page for this dataset	cfResProd.cfURI
	author (dc:creator): original creator of the dataset	cfPers_ResProd or cfOrgUnit_ResProd with author class label
	author_email:	cfPers_ResProd, cfPers_Eaddr or cfOrgUnit_ResProd, cfOrgUnit_Eaddr
	maintainer: current maintainer or publisher of the dataset	cfPers_ResProd or cfOrgUnit_ResProd with maintainer class label
	maintainer_email:	cfPers_ResProd, cfPers_Eaddr or cfOrgUnit_ResProd, cfOrgUnit_Eaddr
	license (dc:rights): license under which the dataset is made available	cfResPubl_ResProd (link to license document)
	version: dataset version	cfResProdVersionInfo (in upcoming CERIF release) or cfPers_ResProd / cfOrgUnit_ResProd / cfResProd_ResProd with appropriate semantic label.
	notes (description) (dc:description): description and other information about the dataset	cfResProdDescr
	tags: arbitrary textual tags for the dataset	cfResProd_Class with appropriate controlled vocabulary
	state: state of dataset in CKAN system (active, deleted, pending)	cfResProd_Class with appropriate controlled vocabulary
	resources: list of [[Domain Model/Resource|Resources]]	cfResProd_Medium
	groups: list of [[Domain Model/Group|Groups]] this dataset is a member of	cfResProd_ResProd
	"extras" - arbitrary, unlimited additional key/value fields	Many extra data elements can be modelled via the CERIF Semantic Layer as semantic relationships or classifications. In extreme cases, additional fields might be added to CERIF entities, on the condition of respecting the rules on multi-linguality support.
Resource (corresponds to an online resource)		***cfMedium*** (file), ***cfSrv*** (API), ***cfResProd or cfSrv*** (visualization), ***cfResProd*** (code), ***cfResPubl*** (documentation)
	url: the key attribute of a resource (and the only required attribute). The url points to the location online where the content of that resource can be found.	{cfMedium | cfSrv | cfResProd | cfResPubl}.cfURI
	name: a name for this resource (could be used in a ckan url)	cfMediumTitle or cfSrvName or cfResProdName or cfResPublTitle
	description: A brief description (one sentence) of the Resource. Longer descriptions can go in notes field of the associated Data Package.	cfMediumDescr or cfSrvDescr or cfResProdDescr or cfResPublAbstr
	type: the type of the resource. One of: file | file.upload | api | visualization | code | documentation	{cfMedium | cfSrv | cfResProd | cfResPubl}_Class with appropriate controlled vocabulary
	format: human created format string with possible nesting e.g. zip:csv. See below for details of the format field.	{cfSrv | cfResProd | cfResPubl}_Class with appropriate controlled vocabulary or cfMedium.cfMimeType
	size: size of the resource (content length). Usually only relevant for resources of type file.	cfMedium.cfSize
	last_modified: the date when this resource's data was last modified (NB: not the date when the metadata was modified).	cfPers_Medium or cfOrgUnit_Medium with appropriate class term and timestamp.
	hash: md5 or sha-1 hash	cfResProdInternId, treat as identifiers for all other entities
Group		***cfResProd_ResProd*** (using the recursive relationships with appropriate class labels)
Dataset Relationship		***cfResProd_ResProd*** (with appropriate class labels and timestamps)
Tag		***cfResProdKeyw***
Vocabulary		***Classification schemes in the CERIF Semantic Layer***

Table 3. Mapping of CKAN to CERIF.