

Towards a Federation of Government Metadata Repositories

Gofran Shukair¹, Nikos Loutas^{1,4}, Vassilios Peristeras², Klaus Reichling³, Fadi Maali¹,
Konstantinos Tarabanis⁴

¹Digital Enterprise Research Institute, NUI Galway

<mailto:{firstname.lastname}@deri.org>

²EC, DG for Informatics, Interoperability Solutions for European Public Administrations

vassilios.peristeras@ec.europa.eu

³init[AG für Digitale Kommunikation, Berlin, Germany

klaus.reichling@init.de

⁴Information Systems Lab, University of Macedonia, Thessaloniki, Greece

kat@uom.gr

Abstract. Data models, taxonomies, ontologies, code lists and semantic data exchange formats are the key resources for achieving data interoperability. These resources exist in several national repositories that differ both in scope and in the target groups they address. They are implemented using different technologies and expose different interfaces to the end user. However the semantic content they include can often be reused even bypass the domain they were originally designed for. A standardized model that allows precise and rich description of repositories content is a key enabler of seamless data exchange. Such a model enables building a federation of repositories where users can search for resources or assets across all the available repositories using a unified user interface.

Keywords: E-government, Semantic Interoperability, Data Exchange, federation, Cross-repository Querying.

1 Introduction

Resources, such as data models, schemata, taxonomies, ontologies and code lists are the means for seamless data exchange. In the context of this work, we use the term *asset* to refer to these types of resources. Hence, an *asset* is a container dedicated to group artifact types. It can be considered as collection of data or a dataset. Currently assets exist in isolated national e-government metadata repositories. These repositories differ *(i)* in scope, *(ii)* in target group they address, *(iii)* in their underlying implementation technologies *(iv)* in interfaces they expose to the end user. The semantic content they include can often be reused even bypass the domain they were originally designed for. But the physical isolation of these repositories and the heterogeneity of the assets hamper the reusability of common concepts and cross-repository search.

The creation of a common metadata schema for these assets allows smooth data exchange and integration between different repositories. This enhances the

discoverability and reusability of the assets and results in consistent and semantically interoperable instance data.

Towards this direction, the European Commission has launched the SEMantic Interoperability Centre Europe (SEMIC) initiative led by the ISA7 program [1]. SEMIC fosters the reuse of syntactic and semantic assets and it provides one kind of governments metadata repositories platforms. The directions discussed in this position paper are inspired by, build upon and extend further the work of SEMIC. We aim to semantically interlink SEMIC with national asset repositories across the EU, thus creating a flexible federation which will facilitate cross-repository search and will consequently boost asset reusability.

Hence, this position paper aims to (a) define a common data model to describe the assets of national e-government metadata repositories and (b) to develop a proof-of-concept prototype showcasing the added-value of federation through cross-repository asset discovery and retrieval using a single point of access.

The remainder of this position paper is structured as follows. Section 2 introduces the common asset model. Section 4 introduces a proof-of-concept architecture for the federation. Finally, conclusion and future directions are briefly discussed in section 4.

2 Common Model Development

Semantic interoperability requires a mutual agreement on the meaning of concepts. Without such an agreement, interoperability conflicts can arise of which Peristeras et al. [2] provide a conceptual analysis and classification. Toward this direction, this section introduces the Asset Description Metadata Schema (ADMS) and discusses its modeling and implementation using Linked Data technologies. Hence, ADMS is implemented as an RDF Schema vocabulary which includes the core elements and the attributes needed to model an asset.

The Linked Data community has developed a number of vocabularies for the description of datasets. Such efforts were reviewed as an asset is a dataset per se. For example Void [3] was proposed for describing RDF datasets. But given the asset definition provided in section 1, an asset is not necessarily an RDF dataset. It might be a csv file, a text file or an XML file.

dcat [4] is a vocabulary to describe government catalogues and datasets. It is a project¹ of W3C's eGov Interest Group. dcat introduces a dataset class as a collection of data, published or curated by a single source, and available for access or download in one or more formats. dcat deals with general government datasets which can be a CSV file or a geographic Shape file. ADMS assets span beyond these two types of files. Hence, special descriptive properties are required. Moreover, ADMS assets usually go through a different lifecycle that needs to be captured and formally described.

The two main classes defined in the ADMS model are *adms:Asset* and *adms:Release*. The current view of the ADMS model is illustrated in Fig 1.

adms:Asset: Represents a reusable dataset, such as an ontology, code list or a taxonomy. It also serves as a container to group various versions of the files of the

¹ http://www.w3.org/egov/wiki/Data_Catalog_Vocabulary

dataset. It is defined as a subclass `dcat:Dataset` and extends it by (i) defining ADMS properties to capture domain-specific information, and (ii) reusing properties from existing linked data vocabularies (e.g. `dcat`, `dc`), e.g. `dc:publisher` and `dc:spatial`.

adms:Release: Represents an actual version of an asset. It can be thought of as a set of files that form a certain version of an asset. *adms:release* is a property linking each asset to its latest release.

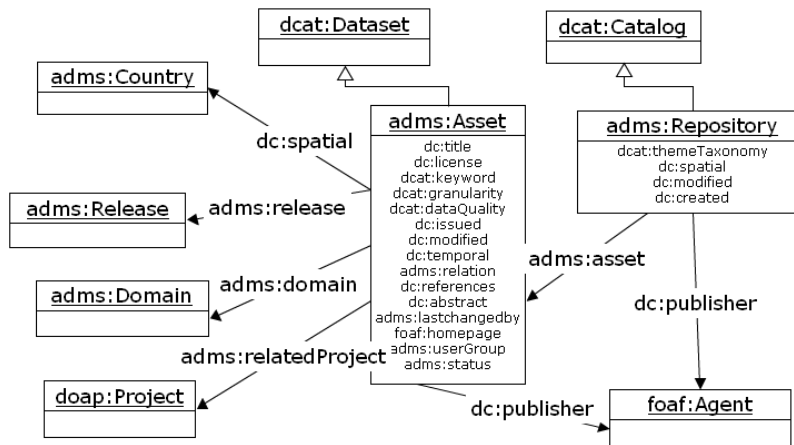


Fig. 1. ADMS Model in RDF

Assets are usually defined to describe a specific domain. Capturing contextual information is essential in facilitating searching and querying functionalities. In the presented model context is described using the following classes and properties:

- *adms:Domain*: A subclass of *skos:Concept* describes the main topic of the Asset. It is assigned to an asset using the *adms:domain* property where each asset covers one or more domain.
- *adms:relatedProject*: A property that links each asset with one or more projects in which the asset was developed, used or somehow related.
- *adms:Country*: Represents countries involved in the development and/or countries in the scope of the asset. It is assigned to an asset using the *dc:spatial* property.

Different classes are explicitly defined for each asset's content type (*adms:Release*), also instances of these classes are created for each commonly used type as the following (Fig. 2):

- *adms:SemanticContent*: Represents the semantic specification of an asset. Example instances of this class are code-list, mapping, taxonomy and ontology.
- *adms:SyntaxContent*: Represents the syntactical specification of an asset. Example instances of this class are XML Schema, WSDL and csv.

² <http://dublincore.org/>

- *adms:ModelContent*: Represents the models that may be in an asset. Example instances of this class are UML and UMM. Each of the previous classes is related to a release of an asset using the properties *adms:semantic*, *adms:syntax* and *adms:model* respectively.

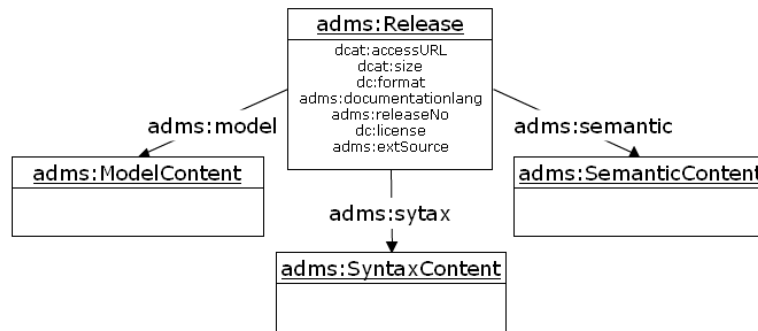


Fig. 2. Asset Content types

An asset goes through many successive stages of quality assessment, development and feedback in order to mature and be a good candidate for reuse. Therefore we express this in ADMS using different instances of *adms:Release*. Each asset is related to its latest release using the property *adms:release*. Each *adms:Release* is related to its previous one using *dc:hasVersion*.

3 Proof-of-concept Architecture

A proof-of-concept architecture for the federation is illustrated in Fig. 3. The architecture is model-driven (based on the ADMS model) and allows assets housed in different national repositories to be queried, discovered and retrieved through a single point of access (implemented as a SPARQL endpoint).

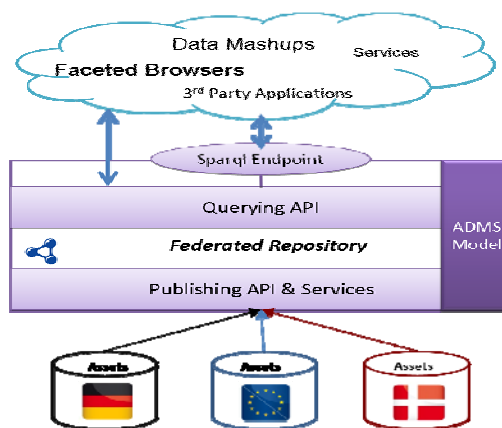


Fig. 3. Proof-of-concept prototype

The ADMS model defines the administrative and descriptive metadata of the assets housed in the repositories. Repositories can publish their assets' metadata using the Publishing API and services. Details on the querying and publishing APIs are beyond the scope of this position paper. Finally, services, such as data mashups and faceted browsers, and third party applications can be built based on this architecture, thus facilitating the access to, the discoverability and the easy reusability of the underlying assets.

4 Conclusion and Future Work

This position paper proposed a conceptual model and the RDF implementation for the ADMS model. The ADMS model acts as an interchange format that enables the standardized description of semantically interoperable metadata of assets housed in the government repositories.

The purpose of the ADMS model is to serve as the common denominator between the metadata models of national asset repositories. Having said that, we plan go on with the detailed analysis of the available metadata models in order to refine ADMS classes and properties. Through this position paper, we provided a first approach for triggering discussion which is always a prerequisite to attract and create real possibilities for future adoption and take up.

In parallel, we will continue with the development of the proof-of-concept for the federation, thus using it a means to provide specific evidence about the expected benefits and the impact of our collaborative effort.

Acknowledgments. This work is a joint effort between Jinit[, EC ISA, NUI Galway and the University of Macedonia. It is funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

1. Reichling, K., Luts, M., Fahl-Spiewack, R.: A pan-European repository: SEMIC.EU as the point of reference for eGovernment ontologies. In: Proceedings of the 1st Workshop on Ontology Repositories and Editors for the Semantic Web. Crete, Greece (2010)
2. Peristeras, V., Loutas, N., Goudos, S. K., Tarabanis, K. A.: A conceptual analysis of semantic conflicts in pan-European e-government services. In: J. Information Science, 34 (2), 877-891, 2008.
3. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets. In: WWW 2009 Workshop: Linked Data on the Web (LDOW2009). Madrid, Spain (2009)
4. Maali, F., Cyganiak, R., Peristeras, V.: Enabling interoperability of government data catalogues. In Electronic Government, Lecture Notes in Computer Science, pages 339-350. Springer Berlin / Heidelberg, 2010.