Yuri P. Lipuntsov
Moscow State University

THE SEMANTIC WEB SPECIFICATIONS FOR DISTRIBUTED STORAGE OF GOVERNMENT DATA

Introduction

The information assets of the public sector can give opportunity to improve the maturity level of public functions, as well as improve the quality of services for citizens and businesses. For reuse of information resources is necessary to create the environment for storage in standardized forms and infrastructure for data access of end users and systems. Now most of assets published on the web-sites, applications, and databases are not suitable for usage in a machine-readable mode. On Russia example of the informatization of public sector on agency level can be realized by deferent application, agency can choice the type of application, technology by implementation information technology (IT) projects. The consequence of this is the data model is predetermined by business-logic of specific applications and technology.

At the regional level, where the region's authorities performed the identical function but have differences in the organizational structure and sets of the functions performed by individual business units, there is a significant technological inconsistencies that do not allow present the understandable picture on the next levels of the hierarchy.

Open data dictionary DCAT

The most services of public administration have cross-agency connections, different departments operate on the same objects, performed different operation. To perform the agile analysis of public data the government sector needs linked data, organization of data records so, that any object has unique identifier.

The usage of that tool can allow us to organize a binding transaction data about the object to one data set and improve the efficiency of information exchange.

The technological base of data binding is the URI (Uniform Resource Identifier): Identifies the resource, which will provide information about the term of the controlled vocabulary or object and assigned a URI. For each base object and can be determined by the URI based on http protocol. This means that user or application received data of the object with URI, he can see the description of the term or object by URI address. This data format allows understanding the semantic of these data for the user and the software agent.

For example, if some company has a unique URI it can allow us to obtain all actions of company life cycle: registration, shareholders decisions, licenses, manufactory equipment, market activity, financial reports. Publishing this data on the open data portal allows users to receive access by API. For the publication of open data W3C has certified dictionary Data Catalog Vocabulary (DCAT), designed to facilitate interoperability between published data catalogs (http://www.w3.org/ns/dcat #) DCAT. Publishing data by DCAT dictionary improve the search of data for software agents and use this data for multiple topics. This technology can allow create distributed data catalogs and implement the federation for search system.

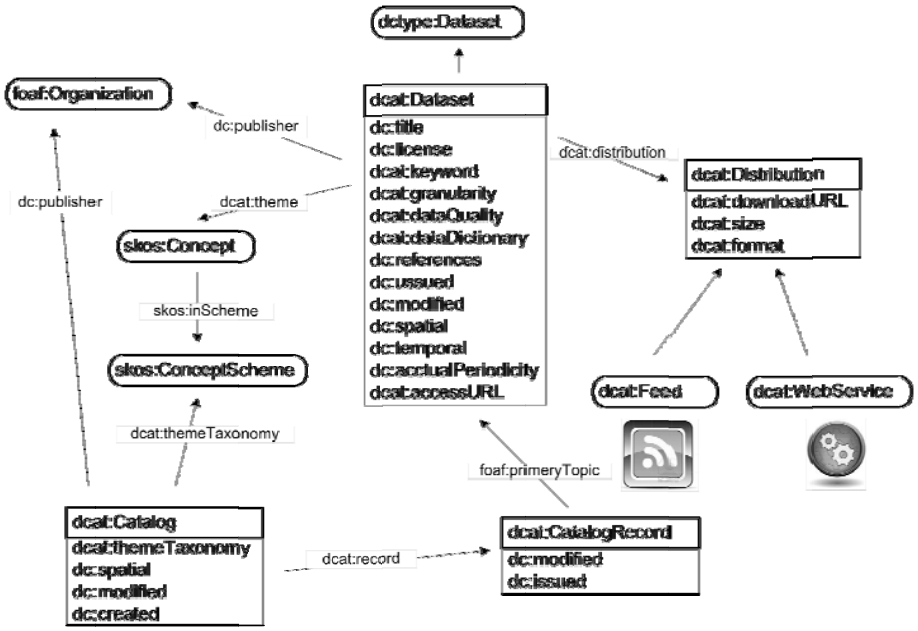

Fig. 1. Data Catalog Vocabulary
Source: [www4].

Vocabulary DCAT (Figure 1) include some terms from other vocabularies: Dublin Core (DC) (http://dublincore.org/documents/dcmi-terms/), and FOAF, SKOS (http://xmlns.com/ foaf/0.1/). New terms of vocabulary DCAT is the class Dataset (dcat: Dataset). This class include the records (class dcat:Record). The Datasets are described by classificatory in the terms vocabulary SKOS. The published open data give users free access to data without license. Access to the data is described by the class dcat: Disrtibution. Access can be provided through a subscription (dcat: Feed) or by a software interface (dcat: WebService).

A number of countries are active in the publication of open data: United States (http://data.gov), United Kingdom (http://data.gov.uk), European Union (http://lod2.eu). In January, 2013 the Government of Moscow presented the open data portal (http://data.mos.ru/). The preliminary stage of open data publication is creation of system for unique identification any base object in URI format. The UK Government, the United States performed a grate work to create standard identifiers for base entities such as schools and roads, public bodies and their functions, etc. This allow to any government agencies publish data about this objects at the federal level and at the level of cities, states, provinces, countries, etc.

The inter-agency and inter-layer information exchange

Publication of government data in open format cannot solve all problems in public sector information exchange. More important task is organizes the interoperability of government systems. One of the important areas of inter-agency and inter-layer interaction is the development of data models that are available to all interested participants. Creation and usage of standard data model will align the conceptual schema for data, including the reference data (e.g. uniform codes, identifiers, taxonomies, registers, geospatial data, licenses, etc.). Agreement on the data models will be an important step in the interoperability of information systems. Initiative for metadata management can help reduce the number of conflicts in semantic interoperability.

European Union initiative developed the theoretical models of the government data publication by standard namespace. This solution is presented as three models that describe the organization of information assets storage in repository, their metadata, and software for data access and processing (Figure 2). Interoperability of inter-agency interaction in case of ADMS is replaced by network of repositories that store information assets.

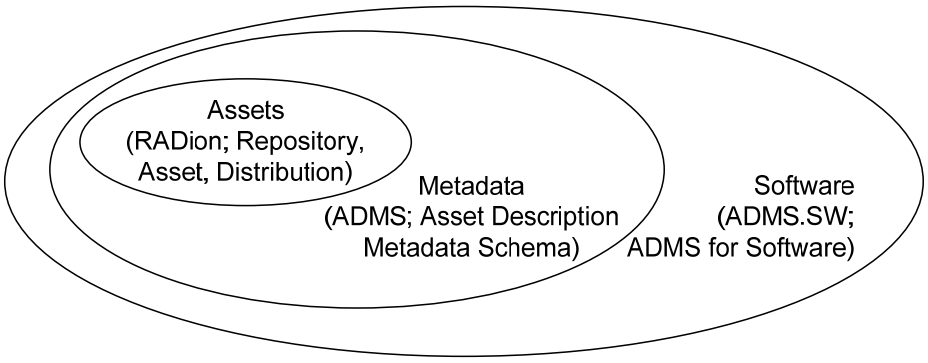

Fig. 2. The three-level models of ADMS
Source: Created by the author on base [www1].

For cross-agency connection is necessary to implement two types of models:

1. Organizational model that describes all the members of the community and their interactions.
2. The data model of the domain.

First group of models realized by RADion (Figure 3) and ADMS models. RADion describes the repository parameters that stores assets (Repository), own information assets (Assets) and distribution (Distribution).

The role of users is to manage the licenses (License: RADion), which allows for reading, editing, publication of assets. The implementation of e-Government services and information for administration involves different types of information assets: the base registries, domains, transactions, flow of documents, archival data. The management of information needs know all transaction of information asset. ADMS can provide this type of data about any asset publish in repository.

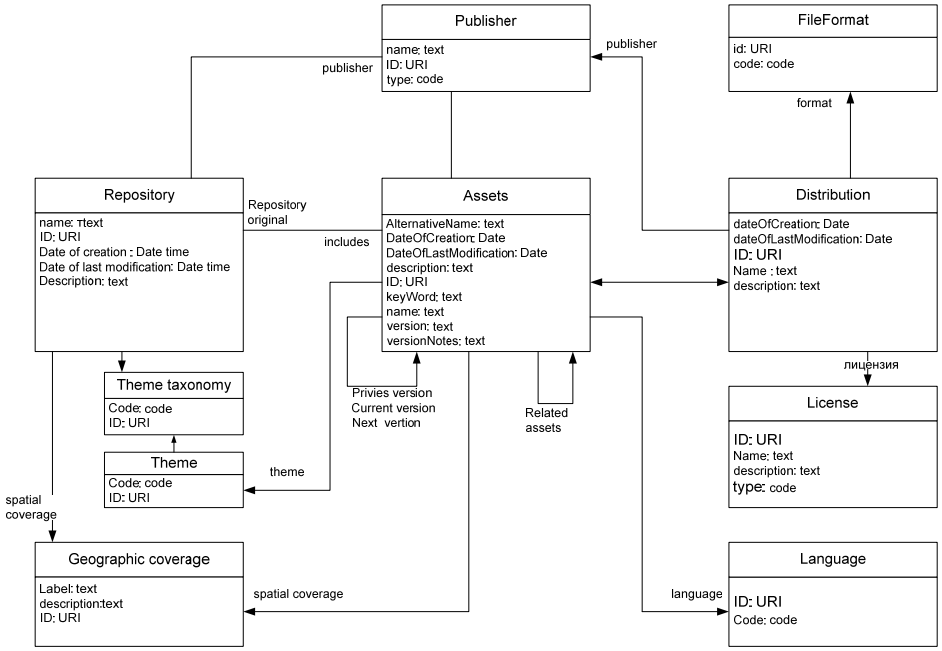

Fig. 3. RADion model for storage and distribution of information assets
Source: [www1].

Semantic of assets realized by the metadata model ADMS, where any information asset connected with detailed description: “The level of interaction”, “Documentation”, states the type of the asset, the period of time the relevance of asset performance technology and other positions.

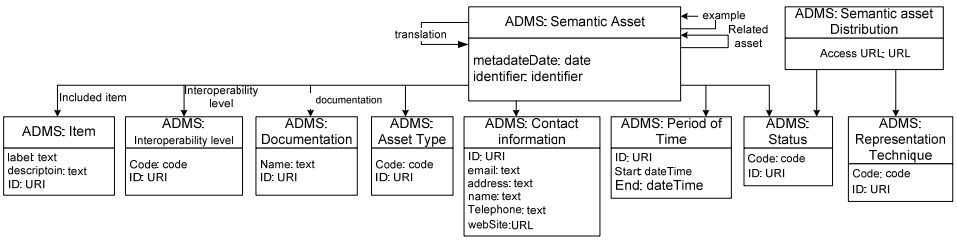

Fig. 4. Metadata description of information assets
Source: [www1].

The dictionaries of ADMS are published by SKOS specification. Now in SKOS format published six controlled vocabularies as components of the ADMS:

- the type of information asset, for example domain “model”, “scheme”, “taxonomy”,
- the level of interoperability, for example “The legal interoperability”, “organizational interoperability”, “semantic interoperability” and “technical interoperability”,
- license type: for example “attribution” BY, or “no derivative work” ND,
- type of publisher: for example “Company” and “public authority”,
- format of presentation: for example “UML”, “XML Schema”, “Schematron”, “OWL”, etc.,
- status: for example “Completed”, and “work in progress”.

The technical aspects of interaction and the sharing the components and applications developed the third model: ADMS SW (ADMS SoftWare), which describes the technical parameters of information assets. Description of components made using the product classification Trove, project SourceForge.net. This system is used for the classification in software development projects. This model is attempt to create environment and technological infrastructure for providing information for the software development and distribution. Project experience with open source license shows that this approach allows for a high rate of development and regular modification of software. The development of a component work more efficiently in a format where any participant can take part in the development or revision of the components than the licensing of development. Enthusiasts and followers of the projects give a lot of time open libraries that allow the development of this area with greater efficiency than amply funded projects.

The dictionary for publishing open data DCAT, focused on providing access to citizens and businesses, while models in the ADMS were created for the interoperability of government systems, which is most important in the execution of administrative processes that are interagency by nature.

A set of RADion, ADMS, ADMS SW models are more focused on data integration, while the task of DCAT is the description of open data.

The ISA program along with a number of previous EU initiatives are aimed at increasing democratization of the society by improving the transparency of authorities and participation of citizens, on the one hand, and at creating a framework for implementation of pan-European electronic services and provision of information support for decision making, on the other.

Information or semantic interoperability closely related with data integration. The type of integration that are used in the corporate sector (ETL, Federation for RBD, SOA) [Gior11] do not always meet the needs of government, as the corporate sector is providing more opportunities to implement a rigid (strong) integration model. Such methods can be fruitfully used for e-services, but they will be of no use in the field of public administration information support.

If there is a good solution created for the integration type by a semantic web, it will be the universal solution. This method of integration has several advantages: It can be used to integrate all types of information assets, both structured and unstructured, while the majority of the methods used in the corporate sector are focused on the structured data integration.

Another positive aspect is that any user can determine the composition of the required data themselves. The data layer and data structure are accessible to the user, not hidden, unlike when they use the web service integration. In SOA, the user can receive the ready service and get a certain data set. Service parameters often not enough to get the needed dataset, so they have to turn to service providers. This generates a large number of services that turn into a hairball [ScLy10, p. 337].

It may be not that easy to create a complete picture using the semantic web as a method of data integration, which is provided by tight integration methods (normalization, integrity), but this is an opportunity to provide informational support to public administration officers, which will adhere to the "it is better to be approximately right, than precisely wrong" principle of management accounting of the corporate sector.

In this case the interoperability as a smooth interaction of information systems belonging to different departments is replaced by interaction with the repository, in which each member of the interagency processes publishes its information assets.

Implementation of electronic public services and the provision of administrative officers with information involve different types of information assets: Data of core registers, domain registers, transactions, document flow (interactions) and archival data. The sequence of transfer of individual assets to the repository is a matter of public information governance. ADMS makes it possible for departments to publish and register these assets properly.

Information sharing environment: data integration of educational structures

New vocabularies are already being used in some areas: for example Eurostat liked statistical data [www8], the common information sharing environment for maritime domain [www9]. We have made some steps toward to create an environment of data exchange in education sector.

Education is the one main government services for many participants and categories of players of economic system. One of the problems of Education is to map the customer's expectations of a potential employer, company and the performance of educational institutions, on one side and the comparability of curricula in terms of structures, programmes and actual teaching for higher education institutions of the Bologna Process [www7], on other side.

The interaction of individual actors of Education need to create an environment to perform main steps: development, data integration, coordination, and the common concepts that meet the agreements of all categories of participants.

The aims of system interaction in the education sector to implement approach an expanding organization when information systems of all categories of participants are used to interact with government agencies, businesses, universities, alumni, to expand the scope of a single organization for the information and knowledge exchange. Such cooperation will help to raise the responsiveness of the educational environment and to coordinate the goals and activity of different organizations.

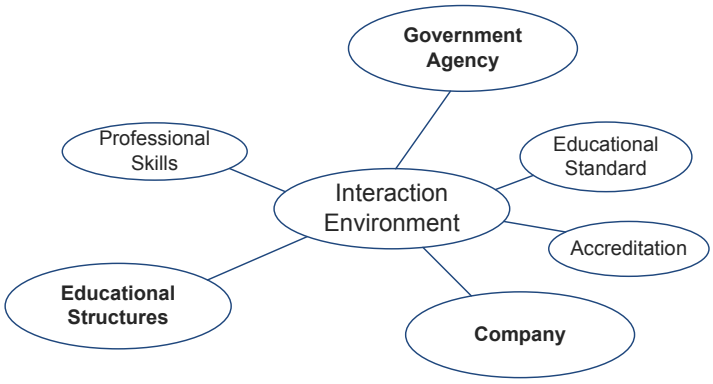

Fig. 5. Main classes of data in Education
Source: Created by the author.

Communication environment in the sector of education suggests the presence of such modules as:

1. Curriculum design for manage set of disciplines.
2. Syllabus engine in form of learning management system.
3. Competency management system as employers need and evaluation in terms of competences.
4. Competency development framework for different area, like [PMI07].

Now at the Economics Faculty of Moscow State University created the first two modules in the traditional relational model [www10], [www11]. We are working on the transition of these modules in the technology of semantic web. At the same time we are developing environment for permanent contacts on the basis of information technology with employer associations and government agencies.

One of the highlights of educational exchange data environment is the preparation of comparable syllabus of individual disciplines, the principle of what for some area will be worked out within the framework of the project Active Teaching Methods implemented by Economics Faculty of Moscow State University and University of Economics in Katowice.

Conclusion

Semantic web model are replacing an object-oriented paradigm as a method flexible sharing of information between participants. The relational technologies are suited for the data exchange within a controlled environment. In case of unknown number of participants, a variety of data and systems are needed are very simple methods of integration suitable for all participants. These technologies provide the semantic web, as approach is based on simple data organization scheme, universal for many domains. Using these schemes any participant can make information resources accessible to different users, automatically integrate into the global information space. These technologies are widely used to create multi-lingual online encyclopedia, social networks, intelligent search, creating the information space within the user community.

The new technologies for data exchange suppose two phases implementation in practice. The first phase is gain experience in the sector of open data. This area involves the development and adoption a set of technology standards. W3C with the European Union demonstrate progress in this direction, one example of which is the family of models ADMS.

The next phase involves the use of semantic web technologies for data integration within the community. The participants' interaction in this case is based on a data model of the domain. This model of information exchange makes it possible to build a semantically consistent expression by any operators.

One of the main subjects' areas for development the domain model is Education, which requires the environment for information exchange between companies, universities and government agencies. Creation the environment is a field for inter-university projects, with core in the universities as a place with the focused specialists, technology, stable relationships with other participants.

References

Books

[Gior11] Giordano A.D., Data Integration: Blueprint and Modeling Techniques for a Scalable and Sustainable Architecture, IBM Press, 2011.

[ScLy10] Schmidt J.G., Lyle D., Lean Integration: An Integration Factory Approach to Business Agility, Addison-Wesley, 2010.

[PMI07] Project Manager Competency Development (PMCD) Framework, Second edition, Project Management Institute, 2007

Periodicals

[BaKa12**]** Bauer F., Kaltenböck M., Linked Open Data: The Essentials a Quick Start Guide for Decision Makers, Retrieved from www.semantic-web.at/LOD-TheEssentials.pdf, 2012.

[JaMV12**]** Javier F., Mendez M., Vicente J., Advantages of Thesaurus Representation Using the Simple Knowledge Organization System (SKOS) Compared with Proposed Alternatives Juan-Antonio Pastor-Sanchez, Retrieved from http://informationr.net/ir/14-4/paper422, May 2012.

[1Gui86] ISO 5964:1986 Guidelines for the Establishment and Development of Multilingual Thesauri, 1985.

[2Gui86] ISO 2788:1986 Guidelines for the Establishment and Development of Monolingual Thesauri, 1986.

[Peri12] Peristeras V., ADMS, Federation and Core Vocabularies Improving Semantic Interoperability in European Public Administrations, SEMIC, Brussels 2012.

[Conf10] Lipuntsov Y., and list of Co-Authors, Conformance Guidelines – The Path to a Pan-European Asset, ISA program, EU, Brussels 2010.

Web sites

[www1] ADMS Working Group, Asset Description Metadata Schema (ADMS), Retrieved from https://joinup.ec.europa.eu/asset/adms/release/100, April 2012.

[www2] Cyganiak R., State of Play in Linked Open Data, Retrieved from http://www.slideshare.net/init_brussels/cyganiakrichardstateofplaylod, June 2011.

[www3] Semantic Catalog (RDF), Retrieved from http://www.data.gov/semantic/data/alpha, October 2011.

[www4] Data Catalog Vocabulary Project, Retieved from http://www.w3.org/egov/wiki/Data_Catalog_Vocabulary, April 2013.

[www5] Data Catalog Vocabulary Project, Retrieved from http://www.w3.org/TR/vocab-dcat/, February 2013.

[www6] Weitzner D., Kagal L., Berners-Lee T., Connolly D., Promoting Interoperability between Heterogeneous Policy Domains, Obtained from the DIG: http://dig.csail.mit.edu/2006/Talks/1017-w3cws-rein/ # (1).

[www7] Tuning Educational Structures in Europe, Retrieved from http://www.unideusto.org/tuningeu/, May 2013.

[www8] Eurostat Dashboard, Retrieved from http://epp.eurostat.ec.europa.eu/inflation_dashboard/, May 2013.

[www9] SEMIC 2013 – Semantic Interoperability Conference, Retrieved from http://joinup.ec.europa.eu/community/semic/event/semic-2013-semantic-interoperability-conference-2013, June 2013, May 2013.

[www10] Personal Cabinet Employees and Students of EF MSU, Retrieved from http://lk.econ.msu.ru, May 2013.

[www11] Resource of Disciplines of EF MSU, Retrieved from http://on.econ.msu.ru, May 2013.

SPECYFIKACJE SIECI SEMANTYCZNEJ DLA ROZPROSZONEGO PRZECHOWYWANIA DANYCH ADMINISTRACJI PUBLICZNEJ

Streszczenie

Efektywne wykorzystanie zasobów informacyjnych sektora publicznego może wpłynąć na wzrost jakości realizowanych usług i wykonywania powierzonych funkcji publicznych. Ponowne użycie zasobów informacyjnych sektora publicznego polega na wykorzystaniu organizacyjnych standardów przechowywania ich infrastrukturalnych rozwiązań w celu zapewniania dostępu użytkownikom końcowym do aktywów i systemów informatycznych.

Obecnie większość z tych witryn, aplikacji, baz danych nie jest dostępna w formie możliwej do zwyczajnego, zrozumiałego odczytania. Federalne i regionalne władze w Rosji samodzielnie określają wykorzystywane aplikacje, technologie informacyjne do ich tworzenia. Prowadzi to do tego, że modele danych aplikacji są z góry określane przez biznesłogikę funkcji automatyzujących. Jest to szczególnie istotny problem na poziomie regionalnym, ponieważ w każdym regionie organy władzy wykonują identyczne funkcje, ale z powodu różnic struktur organizacyjnych, przydziału obowiązków regionalne rozwiązania technologiczne nie są zintegrowane z innym poziomem w hierarchii. W artykule omówiono format przechowywania państwowych danych na poziomie federalnym i regionalnym.

Praktyczne zastosowanie omówionych standardów przedstawiono na przykładzie sektora edukacji. Wymiana informacji w tym sektorze ma kluczowe znaczenie przy określaniu współdziałania (interoperacyjności) pomiędzy pracodawcami, instytucjami edukacyjnymi i agencjami rządowymi. Wiodąca rola w tworzeniu środowiska dla wymiany informacji sektora edukacji powinna być przynależna uczelniom, gdzie znajdują się eksperci przedmiotowych dziedzin, technologie, istnieją silne powiązania z innymi członkami wspólnoty edukacyjnej.