Methods of Semantic Integration in Distributed Information Systems: Challenges of Application

U.M. Akatkin[1], E.D. Yasinovskaya[2], M.G. Bich[3], A.V.Shilin[4]

Introduction

Semantic assets are the basis for data collection, search, analysis and data visualization using semantic properties, and enable semantic interoperability of distributed information systems as a whole. At this time interoperability in distributed systems is mostly supported on technical and organizational levels, however semantic interoperability is quite essential for heterogeneous environment of distributed systems. The ability of information systems to interact on a semantic level can be achieved by joining the efforts of IT-specialists and domain experts. The goal of this joined effort is to transfer the knowledge of the experts about the domain from paper documents into machine-readable representation (ontologies, thesauri, and glossaries). Without this step, the dissemination of knowledge outside of a particular information system is difficult and insufficient for the "understanding" and the (re)use by other interacting distributed systems.

Semantic integration of data

"Semantic modeling has been a subject of research ever since the late 1970s" [1], however, the object-oriented approach still dominates. "Technologies already exist to overcome the heterogeneity in hardware, software, and syntax that is used in different systems (e.g., the ODBC standard, XML based standards, web services and Service Oriented Architectures). While these capabilities are essential to information integration, they do not address the issue of heterogeneous data semantics that exist both within and across enterprises" [2]. The problem occurs when data sources and receivers use different contexts (assumptions); a user submits a query and interprets the results in a certain context, which is completely different from contexts received from sources. Implicit assumptions made in each source need to be explicitly described and used to reconcile conflicts when data from these systems are combined [3].

Distributed information systems as a rule combine heterogeneous sources of data, so data integration is necessary and can be provided at the physical, logical and semantic level. Physical and logical levels are the most used levels of data integration, but they do not take into account the semantic properties of data and therefore are not able to resolve semantic conflicts between these heterogeneous data sources. Data integration on the semantic level supports a unified view of data based on semantic properties of the data in the context of a single domain ontology [4] connected with the other semantic assets that they rely on. For instance, thesauri and ontologies are created using dictionaries and taxonomies, which, in their turn, are built on glossaries.

Table 1 compares the properties of traditional and semantic data integration [5]:

Table 1 - Comparison of traditional and semantic data integration

	Traditional	Semantic
Data structure	Mostly relational, focused on consistent data sets	Focused on the relationship between the data units, regardless of their similarity
Method of integration	Data is selected from sources, converted in accordance with predetermined requirements, and loaded into storages	Establishes a relationship between data units in accordance with the definitions in their common ontologies
Scalability	The cost grows exponentially with each new data source	Increasing the number of sources has an insignificant affect on the cost
Origin of sources	Internal	Internal and external
Adherence to standards	Strict adherence to standards, the violation leads to the loss of context	Flexible adherence to standards, the context is preserved under all conditions

Using methods of semantic integration in a heterogeneous information environment leads to the achievement of semantic interoperability. Multiple researchers and authors, including us, agree with Wikipedia’s definition of semantic interoperability as “the ability of computer systems to exchange data with unambiguous, shared meaning. Semantic interoperability is a requirement to enable machine computable logic, inferencing, knowledge discovery, and data federation between information systems” [6]. By thinking of semantic interoperability in terms of collaboration the European Commission identified its function as follows “Semantic interoperability enables organizations to process information from external sources in a meaningful manner. It ensures that the precise meaning of exchanged information is understood and preserved throughout exchanges between parties” [7]. Worldwide research and practice show that this is especially important for data exchange within large-scale distributed systems such as e-government, systems of systems (SoS), cross-border and cross-sector information sharing.

To achieve Semantic Interoperability, the distributed systems “must refer to an agreed authority, typically a terminology that clearly defines the meanings of the items carrying the information. The use of controlled terminologies, and controlled mapping tables and mapping rules for any transformation promises sufficient reliability. These controlled terminologies and mapping tables, also in their representations as taxonomies, ontologies, thesauri are called semantic interoperability assets” [8].

The (re)use of semantic assets is important for both enabling data exchange in large information systems and for the transformation of open data into Linked Open Data (LOD). LOD makes it possible to harvest, compare, and visualize data from heterogeneous resources – Data Portals, Data Hubs, distributed data warehouses.

Challenges of semantic integration

The problem of semantic interoperability is not new, and people have tried to achieve semantic interoperability in the past using various approaches. The experts agree that existing ontology-based approaches for semantic interoperability have not been sufficiently effective [9].

It is important to note that semantic assets are developed by teams consisting of experts in various fields. It is during their collaboration at the stage of formalization and modeling using semantic integration methods that a number of challenges arise:

- The gap between domain experts and IT professionals: modeling the information system and information exchange processes in a heterogeneous environment should be based on knowledge of the domain, its objects, and relationships between these objects. This process must be accessible and understandable to all the participants, regardless of their competence;
- Determination of the depth of domain formalization: the model may be insufficient, or conversely, excessive for the particular information system, however in most cases neither developers nor experts can determine that before the operation stage;
- The lack of a unified and accessible conceptual framework: It often leads to the creation of new semantic assets instead of reusing existing ones. Resulting in unnecessary duplication or incorrect use of terms;
- Difficulty to visualize, validate and interpret the results of semantic integration: it is hard to understand what information will be available to other participants, how it can be represented and how it can be used.

Overcoming the challenges shown above requires a common methodology supported by tools of collaboration, which will simplify the application of semantic integration methods during the design of semantic assets. However, there are no collaboration tools that support domain formalization and the design of models, which when completed can be used to provide information sharing in distributed information systems and to achieve semantic interoperability.

Currently Available Solutions

Over the last 15 years international projects such as NIEM[1], JOINUP[2] and SEMIC[3], Linked Open Vocabularies[4] etc. have provided an important experience and made a large step forward from theoretical research to the practical application of semantic interoperability, and data search and analysis based on semantic properties in a

[1] https://www.niem.gov/
[2] https://joinup.ec.europa.eu/
[3] https://joinup.ec.europa.eu/community/semic/description
[4] http://lov.okfn.org/dataset/lov/

heterogeneous information environment.

Despite the significant differences in the implementation of these projects, they all share the following basic principles:

1. Consolidation and (re)use of semantic assets, including the provision of core vocabularies and data models for information exchange.

2. Collaboration of domain experts and IT specialists using a community-driven, standards-based approach. To achieve this, various tools are used for creation and (re)use of semantic assets while modeling information systems and information exchange services, as well as, transforming open data to linked open data, followed by data harvesting, analyzing and visualizing.

These basic principles allow the achievement unambiguous meaningful interpretation of data for all the participants of information exchange in distributed information systems. However, currently available tools require programming knowledge and are difficult for domain experts to use without the help of IT specialists. This interaction is time-intensive and leaves additional chances for errors and misunderstanding.

Center of Semantic Integration

In 2016, the Plekhanov Russian University of Economics issued an order to start the research and development for project Center of Semantic Integration. It will be a collaboration platform in the sphere of semantic interoperability that provides the expert community with information, methodology, special tools, and necessary organizational and regulatory support. The Center is located on the JINR cloud infrastructure and after completion will unite the working groups, scientific teams, developers, providers of information and information exchange specialists. The project aims to achieve the following objectives:

- Research and approbation of modern approaches to (1) management of semantic assets, (2) creation and dissemination of standards, (3) development of methods and tools based on Model Driven Architecture (MDA[5]) principles to consolidate, integrate, create and (re)use semantic assets.
- Develop the beta version of Center of Semantic Integration.

"Based on OMG's established standards, the MDA separates business and application logic from underlying platform technology" [10] therefore, it can be used to radically reduce the need for programming and interaction between IT specialists and domain experts while designing semantic assets.

The Center of Semantic Integration could be used for:

- E-Government: the collaboration of experts in various fields working on the

[5] http://www.omg.org/mda/

development of semantic models for public services, cross-sector cooperation, and information sharing in order to achieve semantic interoperability of Russian e-Government systems.

- Research: involvement of young scientists and students as well as other research teams in (1) semantic modeling of various domains, (2) creating glossaries, taxonomies, thesauri and ontologies for use in research projects, (3) organizing scientific materials, and (4) improving search procedures for scientific and technical information.
- Data analysis: (1) the ability to normalize public open data, (2) transform it into linked open data reusing semantic assets and search for related data from other sources, (3) providing meaningful interpretation, (4) visualization and combined analysis of data from different sources.
- Semantic integration of information systems: stakeholders, experts and development teams get the opportunity to use the tools for domain modeling of distributed information systems, constructing information exchange schemas using the federated data model.

Conclusion

The above outlined approach used by the Center of Semantic Integration will bridge the gap between IT specialists and domain experts during collaboration by consolidating management methodology of semantic assets and the tools for their development, reuse, and support of the full life cycle.

The beta version of the Center of Semantic Integration will allow implementation and testing of this approach in order to develop interfaces that will be intuitively used for collaboration by experts in various fields.

References

[1] C. J. Date, “An Introduction to Database Systems (8th Edition)”. Pearson Education Inc., 2004, p. 1024, ISBN 0-321-18956-6

[2] Madnick S., Gannon T., Zhu, H., Siegel M., Moulton A., Sabbouh M. “Framework for the Analysis of the Adaptability, Extensibility, and Scalability of Semantic Information Integration and the Context Mediation Approach”, Massachusetts Institute of Technology Cambridge, MA, USA, 2009

[3] Madnick, S.E., & Zhu, H. (2006) “Improving data quality through effective use of data semantics”, Data and Knowledge Engineering, 59(2), 460-475

[4] Serebryakov, V.A. (2012) “Semantic integration of data”, presentation, http://sp.cmc.msu.ru/proseminar/2012/serebryakov.2012.04.20.pdf

[5] L. Chernyak (2009) “Data Integration: Semantics and Syntax”, Open Systems #10, 2009, http://www.osp.ru/os/2009/10/11170978/

[6] Wikipedia. Semantic interoperability https://en.wikipedia.org/wiki/Semantic_interoperability

[7] Annex 2 to the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions 'Towards interoperability for European public services' EUROPEAN COMMISSION Bruxelles, le 16.12.2010 COM(2010) 744 final http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf

[8] Joinup, European Commission, https://joinup.ec.europa.eu/asset/page/practice_aids/what-semantic-interoperability

[9] Walaa S. Ismail, Mona M. Nasr, Torky I. Sultan, Ayman E. Khedr (2013) “Semantic Conflicts Reconciliation as a Viable Solution for Semantic Heterogeneity Problems”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013

[10] OMG Model Driven Architecture http://www.omg.org/mda/