Meaningful Data Interoperability and Reuse among Heterogeneous Scientific Communities

© Nikolay Skvortsov
Institute of Informatics Problems, Federal Research Center "Computer Science and Control", Russian Academy of Sciences,
Moscow, Russia
nskv@mail.ru

Abstract. FAIR data principles declare data interoperability and reuse through the use of machine and human readable specifications. Adherence to these principles has some subsequences for data infrastructures and research communities. Meaningful data exchange and reuse by humans and machines requires formal specifications of subject domains accompanying data and allowing automatic inference. Development of formal conceptual specifications in research communities might be stimulated by a necessity to reach semantic interoperability of data collections and component, reuse of data resources. Data lifecycle hence includes collecting domain knowledge specifications, classifying all data, methods and services by these specifications, collecting and sharing them for reuse. Formal inference allows meaningful search and verified reuse of data, methods and services from collections.

Keywords: FAIR data principles, conceptual modeling, research community

1 Introduction

Curation and sharing research data to make it reusable for both human and machine is a topical issue for last years. For example, WF4Ever project [1] is aimed at preserving data, workflows and research results for their sharing and reuse. Research objects are declared as containers that encapsulate data, metadata, workflows, documentation, links to external resources and share all resources related to a research for a community.

Collaborative data infrastructures support sharing of various resources such as collections, archives, databases, storage and computing capacities, and provide services to search, access and manage them. For example, EUDAT [2] is a network of numerous community specific data repositories and some of Europe's largest data centers using common data services for data and service providers and research communities. EUDAT Collaborative Data Infrastructure (CDI) is a European infrastructure of integrated data services and resources to support research. Heterogeneous research data infrastructure interact to share research data globally and make science open. EOSC [3] initiative integrates services and data from research data infrastructures, provides curation and preservation of scientific data repositories, computing capacity for research data analysis.

FAIR data principles [4] has gathered basic features used in data curation and preservation and now are being propagated in research data infrastructures and open science. These principles are aimed to provide data interoperability and reuse by machines and humans. For this purpose datag should be well identified, specified with ontologies, accompanied by provenance information, and be comply with known data models, or have known mapping to them.

FAIR data principles have been defined informally. So they rises a number of different interpretations, including application of Linked Data principles to provide FAIR ones [5], or lists of more detailed informal requirements based on FAIR ones [6], or just simplified numerical rating of conformity with FAIR principles [7]. At the same time, it seems that FAIR data principles should have some definite subsequences for requirement to research data infrastructures. Ones relevant to data semantics problems with respect to research communities are discussed in this talk.

2 Subject domain specifications

FAIR data principles declare data interoperability and reuse through the use of machine and human readable specifications. It means that data are FAIR if only there is an approach to define and clarify semantics of data in some domains of knowledge. Meaningful data exchange and reuse by machines (helpful for humans too) requires quite formal specifications of subject domains allowing automatic inference.

Similarity and machine learning approaches could be applied to help humans search and operate with data but do not define formal specifications of the used resources and evidence-based inference over metadata. Domain knowledges should define restrictions and permissible states of data from the view of specific domain. Advanced ontological and rule models should be used for metdata development.

Conceptualization and conceptual specifications are necessary not only in general domains, but in domains of interest of narrower and more specialized communities, as well as in overlapping domains, in which cooperation of research teams and reuse of specifications often occurs.

Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018

Most researches are held on intersection of several domains, so they use constraints of several domains simultaneously as points of view to specify research objects. Inference in multidomain specifications should provide establishing relations and semantic interoperability between data belonging to different domains.

3 Collections of methods and experiment specifications

For comprehensive investigations of specific real-world entities, it is important to share data, tools, research results, methods and specifications defining the semantics of entities and phenomena in the domain as well as the semantics of methods applied to them. Thus, no matter which kind of information object is used for research, it should be supplied with metadata in terms of ontologies. Those are data, metadata, publications, implementations of research methods, workflows describing the research processes. Inference over ontologies makes it possible to select them from collections and access by selected identifiers.

Semantics based approaches to research objects should be provided by inseparable linking of data and well defined methods related to objects of research. It means that method collections are considered as a specific data kinds. Methods used in any research domain should be defined, conceptually specified and collected in addition to general purpose methods such as multidimentional data analysis or machine learning. Meaningful access to known implementations of methods should be provided to humans and machines and be understandable for the both.

Experiments over data in research infrastructures are constructed using shared and interoperable data, services and workflows. Research experiments can include data analysis, modelling in accordance with hypotheses and testing models by observational data. Besides providing access to data and method implementation collections, research infrastructures should include instruments for experiment supporting, in particular, formulation and testing of hypotheses [8].

4 The role of communities

Since shared semantics of research objects are becoming increasingly important for data reuse in each specific discipline or subject domain, heterogeneous communities working in a domain should have conceptual specifications related to their research and approaches and maintain strong commitment to them.

Communities of researchers and vendors of analytical tools, research instruments and data owners are interested in the long-term shared access to heterogeneous data and method collections. So the only way of conceptualization and formal specification of a domain is development in communities stimulated by a necessity to reach a semantic interoperability of interacting components, integration of data collections, reuse of data resources and method reproducibility due to binding to semantics of the subject domains.

Community members (humans or machines) operate within the ontological commitment defined by shared ontologies, i, e. use of the concepts of the subject domain in a consistent way with respect to the theories specified by the ontologies. Ontologies are important for the automation of consistency control on any manipulations with the domain concepts. An interaction of communities in solving interdisciplinary problems requires simultaneous querying using different domain vocabularies. In that case, the researchers should commit to the specifications of several domains.

Activities of communities are defined by data lifecycle to provide their interoperability and reuse in related domains. Maintenance of shared domain specifications becomes a basis for arranging collections of data and sources, collections of specific methods, embedding research results into such collections for further research.

Acknowledgments

The work was supported by Russian Foundation for Basic Research (grant 18-07-01434).

References

[1] Belhajjame K., et al: Workflow-Centric Research Objects: A First Class Citizen in the Scholarly Discourse. In: ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica2012), pp. 1-12. Heraklion (2012).

[2] Schentz H., le Franc Y. Building a semantic repository using B2SHARE. In: EUDAT 3rd Conference (2014)

[3] EOSC Declaration. https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf

[4] Wilkinson M., et al: The FAIR Guiding Principles for scientific data management and stewardship. In: Scientific data, vol. 3 (2016)

[5] Wilkinson M.D., et al: Interoperability and FAIRness through a novel combination of Web technologies. In: PeerJ Preprints 5:e2522v2 (2017) https://doi.org/10.7287/peerj.preprints.2522v2

[6] Guidelines on FAIR Data Management in Horizon 2020. Directorate-General for Research and Innovation European Commission (2016). http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[7] Doorn P., Dillo I. FAIR Data in Trustworthy Data Repositories. DANS / EUDAT / OpenAIRE Webinar (2016). https://eudat.eu/events/webinar/fair-data-in-trustworthy-data-repositories-webinar

[8] N. Skvortsov, L. Kalinichenko, D. Kovalev. Conceptualization of Methods and Experiments in Data Intensive Research Domains // Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016). - CCIS, Vol. 706. - P. 3-17. – Springer, 2017.