

The XX International Conference DAMDID/RCDL 2018
(<http://damdid2018.frccsc.ru>)

Workshop on
Invoking the Global Discussion on FAIR Data and EOSC

Venue: Lomonosov Moscow State University
October 9, 2018

The acronyms FAIR Data and EOSC, emerging within the Europe Union, have achieved a global awareness. Now it is time to invoke a global discussion and to integrate views and experience beyond disciplinary and geographic boundaries. The FAIR principles on data require data to be Findable, Accessible, Interoperable and Re-usable. While their seem to be practicable solutions to make digital objects (files, cloud objects, structures in SQL or No-SQL databases, spreadsheets, etc.) findable and accessible by assigning persistent identifiers and rich metadata to them, ways to increase interoperability and improving re-usability are not as evident. Different data organisations, data models and semantic spaces need to be explicitly described and mapped which is known to be very time consuming at this moment in particular since the data universe is changing rapidly. Conditions for re-usage need to be made explicit which are also difficult to formulate. And, all specifications need to be available in machine actionable forms to enable automatic processing which is the only way to scale up processing in data rich science in the future. A new culture of exchanging data and new methods are required to make FAIR data reality.

In addition, the European Commission started the European Open Science Cloud (EOSC) initiative to develop the eco-system of infrastructures that will help to realise a FAIR data domain for data intensive sciences. It needs to include salient infrastructure components and integrate different services as they are being developed by research disciplines, by existing infrastructure initiatives as well as by computer science driven initiatives in a highly interoperable manner. Obviously, EOSC is describing a process that is agile enough to adapt to the dynamic changes in our data universe and not a blueprint for a top-down designed infrastructure.

This workshop will bring a number of different key experts who were actively leading the agendas with respect to FAIR and EOSC and who come from different research disciplines including computer science with their views and expectations about FAIR and EOSC. In particular, ways to achieve a higher degree of interoperability and re-usability will be discussed. The workshop will also be a moment where experts from different regions will take the chance to discuss state and future of the FAIR action plan and the EOSC initiative, bring in their views and enrichments, and discuss active participation.

Therefore, this workshop will be a milestone not only for creating more awareness about FAIR and EOSC, but also in refining the current plans. In particular, we hope to get clarification of what Open Science means for the countries outside of the EU as well as Perspectives of the EC FAIR/EOSC initiative for involvement of all interested Stakeholders beyond EU.

For further information on FAIR and EOSC we refer to the following documents:

- Force11 FAIR Principles: <https://www.force11.org/group/fairgroup/fairprinciples>
- FAIR Nature Paper: <https://www.nature.com/articles/sdata201618>
- FAIR Expert Group Intermediate Report: <http://doi.org/10.5281/zenodo.1285272>
- European Open Science Cloud (EOSC):
<https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud>
- EOSC Declaration:
https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf#view=fit&pagemode=none

Tentative Program

Organizers: Peter Wittenburg, Leonid Kalinichenko

| | | |
|-------|---|---|
| 9.30 | Welcome and Reception | Academician of RAS Igor A. Sokolov |
| 9.45 | Relevance of EOSC and FAIR in the realm of Open Science and phases of implementing the EOSC | Dr. Michel Schouppe, European Open Science Cloud |
| 10.30 | FAIR Data, FAIR Services and the FAIR Data Action Plan | Dr. Simon Hodson, CODATA |
| 11.00 | Coffee | |
| 11.30 | Building FAIR environmental services platforms in Europe | Dr. Ari Asmi, ICOS Project |
| 12.00 | FAIR data based on extensible unifying data model development | Dr. Sergey Stupnikov; Dr. of Sci, Prof. Leonid Kalinichenko (RAS) |
| 12.30 | Meaningful data interoperability and reuse among heterogeneous scientific communities | Nikolay Skvortsov (RAS) |
| 13.00 | lunch | |
| 14.00 | Digital Objects as a concept to help implementing FAIR and EOSC | Dr. Peter Wittenburg (MPS) |
| 14.30 | Ready, Set, GO FAIR | Dr. Erik Schultes (GO FAIR) |
| 15.00 | How FAIR is EOSC? An overview of FAIR activities within the various EOSC funded initiatives | Dr. Damien Lecarpentier (CSC-EUDAT) |
| 15.30 | Panel: Data Access Challenges for Data Intensive Research in Russian Fed. and EOSC | Alexander Fazliev (Atmospheric optics), Eugene Gordov (Climate and Ecology), Nadezhda Kiselyova (Materials Science), Oleg Malkov (Astronomy), Natalia Ponomareva (Neurology), Alexey Pozanenko (Astrophysics), Leonid Kalinichenko (Computer Science) |
| 16.00 | Coffee | |
| 16.30 | Discussion | |
| 18.00 | End | |

DAMDID Workshop Speaker Information and Presentation Abstracts

Michel Schouppe

Title

Relevance of EOSC and FAIR in the realm of Open Science and phases of implementing the EOSC

Abstract

In the realm of its Open Science strategy, the European Commission (EC) has announced its intentions to promote open access to research data resulting from publically funded research under Horizon 2020 to increase the circulation and exploitation of knowledge. Openness should become the default case respecting constraints to privacy, national security and intellectual property rights. For making open science happen, two major pillars need to be addressed: 1) Data needs to be FAIR (Findable, Accessible, Interoperable, Reusable) to overcome the fragmentation and inefficiencies and thus to increase data reuse effectively. 2) Commissioner Moedas made clear that "Europe's final transition must be one from fragmented data sets to an integrated European Open Science Cloud (EOSC). By 2020, we want all European researchers to be able to deposit, access and analyse European scientific data through an EOSC." EOSC and FAIR research data are closely related and the EC is working out phases for implementing EOSC and action plans to make data FAIR.

Affiliation

Senior expert, European Open Science Cloud

Short Bio

To be provided.

Simon Hodson

Title

FAIR Data, FAIR Services and the FAIR Data Action Plan

Abstract

Building on previous summaries of the attributes that make research data most usable by other scientists (for replication, reanalysis and reuse), the framework provided by the FAIR data principles has gained support from policy makers, funders and researchers and has helped progress towards the vision of Open Science. Combined with the principle that research data should be 'as open as possible, as closed as necessary', FAIR data has allowed important principles to be communicated in a more accessible and persuasive way.

This talk will first explore how to implement the FAIR data principles and related and enabling practices, with reference to the work of the European Commission's Expert Group on FAIR data which is developing a report and Action Plan on how to make FAIR data a reality. Second, the talk will consider the challenges of interoperability and reusability in relation to data intensive and multi-disciplinary research areas that seek to address the most important global, environmental and societal challenges. Encouraging the development and growing maturity of vocabularies and standards required for interoperability and the integration of data from diverse disciplines will be a major challenge for 21st century research.

Affiliation

Executive Director

CODATA, the Committee on Data of the International Science Council

Short Bio

Simon Hodson has been Executive Director of CODATA since August 2013. Simon is an expert on data policy issues and research data management. He has contributed to influential reports on Current Best Practice for Research Data Management Policies and to the Science International Accord on Open Data in a Big Data World. As part of his CODATA role, Simon leads or participates in numerous projects, Working Groups and Steering Groups. Most recently he has served as Chair of the European Commission's Expert Group on FAIR Data; co-chair of the OECD Global Science Forum Report on Business Models for Sustainable Digital Repositories; as well as directing the African Open Science Platform Project.

Ari Asmi

Title

Building FAIR environmental services platforms in Europe

Abstract

Research Infrastructures are one of the main elements in the European research landscape, providing the services and data needed for many ambitious scientific projects, as well as being major investments for future and innovation. Thus, they are a natural firstcomers to develop and implement new data and service initiatives. The European cluster of environmental research infrastructures ENVRI has been working on interoperable data for almost a decade, and development of actual policies, (small s) standards and practices for FAIR data and FAIR services has been significant. This talk gives you the overview of particular challenges, solutions and look for future on this crucial field.

Affiliation

ENVRIplus director

Integrated Carbon Observation System ERIC

Short Bio

Originally a climate scientist, Ari Asmi got tired on the difficulty of finding, using, and accessing observational data. He has worked since in many research data and particularly research infrastructure development projects on environmental sciences. Currently he is the director of ENVRIplus environmental research infrastructure project and the coordinator RISCAP project analysing the global research infrastructure landscape, as well as active member of RDA, EGU and many other scientific associations.

Sergey Stupnikov, Leonid Kalinichenko

Title

FAIR data based on extensible unifying data model development

Abstract

Data sources nowadays are quite heterogeneous: they are represented using very different data models. Variety of data models includes traditional relational model and its object-relational extensions, array and graph-based models, semantic models like RDF and OWL, models for semi-structured data like NoSQL, XML, JSON and so on. These models provide also very different data manipulation and query languages for accessing and modifying data.

A prerequisite for (meta)data interoperability, integration and reuse within some data infrastructure is unification of source data models and their data manipulation languages. A unifying data model (called *canonical*) has to be chosen for the data infrastructure. The canonical data model serves as the language for knowledge representation mentioned in FAIR 11. Every source data model has to be mapped into the canonical model. Mapping can be accompanied with the extension of the canonical model if required. A mapping should be formalized and verified: a formal proof that the mapping

preserves semantics of data structures and data manipulation operations of the source data model should be provided.

As the *core* of the canonical model some concrete data model like SQL (conforming to ISO/ANSI SQL standard of 2011 or later) is used. To cover features of various source data models the canonical model has to be *extensible*. Examples of extensions are specific data structures (data types), compound operations or restrictions (dependencies). An extension is constructed for every source data model. Canonical model is formed as the union of the core data model and all extensions.

Data unification techniques were extensively studied at FRC CSC RAS. As the core of the canonical model specific object-frame language with broad range of modeling facilities was used. Approaches for mapping of different classes of source data models were developed: process models, semantic models, array and graph-based models, some other kinds of NoSQL models. Techniques for verification of mappings applying a formal language based on the first order logic and set theory and supported by automatic and interactive provers were developed.

Techniques mentioned are proposed as a formal basis for (meta)data interoperability, integration and reuse within FAIR data infrastructures. Such infrastructures may combine virtual integration facilities (subject mediators) as well as data warehouses to integrate heterogeneous data sources in an interoperable way.

Sergey Stupnikov

Affiliation

Senior research scientist; Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences (FRC CSC RAS), Moscow

Lecturer; Faculty of Computational Mathematics and Cybernetics, the Lomonosov Moscow State University.

Short Bio

Sergey Stupnikov is a computer scientist concentrating on heterogeneous data integration; data model unification, mapping, formal semantics and verification. Sergey is affiliated with the Institute of Informatics Problems at FRC CSC RAS. He is also a lecturer on data integration and scientific advisor within a master program entitled “Big data: infrastructures and methods for problem solving” at the Lomonosov Moscow State University. Main publications of Sergey can be found here: <http://synthesis.ipi.ac.ru/synthesis/staff/ssa.html>

Leonid A. Kalinichenko

Affiliation

Chief Research Scientist at the Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences (FRC CSC RAS), Moscow. Professor of the Department of Computational Mathematics and Cybernetics at the Lomonosov Moscow State University.

Short Bio

Leonid A. Kalinichenko received Ph.D. degree at the Institute of Cybernetics in Kiev (USSR), main results were devoted to discrete events systems simulation (languages, tools, applications) and later, in 1985, the degree of Doctor of Sciences from the Moscow State University devoted to methods and tools of heterogeneous databases integration. Both degrees were in Computer Science, obtained while USSR still existed.

He is an author of ten books, more than 200 research papers in journals and conference proceedings. His current research interests include: interoperable heterogeneous information resource mediation, heterogeneous information resource integration, semantic interoperability, compositional development of information systems, middleware architectures, digital libraries. In 2015 he organized a master program entitled “Big data: infrastructures and methods for problem solving” at the Lomonosov Moscow State University.

During the past period Leonid A. Kalinichenko has successfully formed several international conferences (including the European Conference "Advances in Databases and Information Systems"

(ADBIS), Russian Conference on Digital Libraries (RCDL) that after 16 years of successful work the latter was transformed into “Data Analytics and Management of Data Intensive Domains” (DAMDID/RCDL). Leonid A. Kalinichenko acted as a chair (co-chair) of the Program or Steering committees of the conferences. Main publications of Leonid can be found here:

<http://synthesis.ipi.ac.ru/synthesis/staff/lak.html>

Nikolay A. Skvortsov

Title

Meaningful data interoperability and reuse among heterogeneous scientific communities

Abstract

FAIR data principles declare data interoperability and reuse through the use of machine and human readable specifications. For this purpose data should be well identified, specified with ontologies, accompanied by provenance, and complied with known to the canonical data models, or have known mapping to them. Requirements for data infrastructures are subsequent upon FAIR principles. Some of them relevant to data semantics problems and data lifecycle with respect to heterogeneous research communities are discussed in this talk.

Meaningful data exchange and reuse by machines (helpful for humans too) requires quite formal specifications of subject domains allowing automatic inference. Domain knowledges define restrictions and permissible states of data from the view of a specific domain. 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 can provide establishing relations and semantic interoperability between data belonging to different domains.

Semantics based approaches to research objects should be provided by inseparable links of data and well defined methods related to objects of research. It means that method collections are considered as a specific data kinds. Specific methods should be conceptually specified and collected in any domain in addition to general purpose methods such as multidimensional 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.

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 develop conceptual specifications related to their research and approaches and maintain strong commitment to them. 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.

Affiliation

Senior research scientist; Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences (FRC CSC RAS), Moscow

Short Bio

General research interests of Nikolay A. Skvortsov are ontological and conceptual modeling of research domains and data semantic interoperability issues. He has been affiliated with the Institute of Informatics Problems, Federal Research Center "Computer Science and Control", Russian Academy of Sciences, Moscow, Russia. Recent years Nikolay Skvortsov investigated requirements for reuse of data and implementations of scientific methods in research communities primarily using examples problem development and solving in astronomical research domains. Main publications of Nikolay can be found here: <http://synthesis.ipi.ac.ru/synthesis/staff/nskv.html>

Peter Wittenburg

Title

Digital Objects as a concept to help implementing FAIR and EOSC

Abstract

The FAIR principles are excellent guidelines for improving data practices, however, they are not meant as blueprint to build efficient data infrastructures. EOSC is a fantastic framework to create a momentum towards an eco-system of data infrastructures, but currently it lacks a technical concept. Digital Objects are not just a technical term that have already shown their relevance for data for example in cloud systems, but they are a way to structure data and tools in a way that implements FAIRness and opens a way to implement EOSC.

See also: <http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0>

Affiliation

Senior Data Systems Advisor

Max Planck Computing and Data Facility

Short Bio

Being educated as electrical engineer I worked from 1976 until 2013 as technical director at a new institute of the Max Planck Society studying how our brain processes languages. Then I joined the newly founded Max Planck Computing & Data Facility. I also took responsibilities in various European and international projects centred around research and data infrastructures (CLARIN, EUDAT, RDA).

More information: www.mpcdf.mpg.de/~pewi

Eric Schultess

Title

Ready, Set, GO FAIR

Abstract

The 15 FAIR Principles have found unusually rapid uptake among a broad spectrum of stakeholders, from research scientists who make data, to publishers who distribute data, to science funders who track impact of data. Erik will describe the FAIR Principles and show examples of how they have been implemented. Erik will also present a set of core FAIR Metrics that can help gauge the level of FAIRness of any digital resource. Of particular interest to DAMDID, Erik will demonstrate how additional FAIR Metrics can (and should) be defined by particular Data Intensive Domains in order to address community-specific data structures and analytic requirements. This discussion, and these examples will be presented in the context of the International GO FAIR Initiative. GO FAIR is a voluntary community of stakeholders devoted finding consensus on standards and solutions that comprise an emerging Internet of FAIR Data and Services.

Affiliation

International Science Coordinator & Interim GO TRAIN Coordinator

GO FAIR International Support and Coordination Office, Leiden

<https://www.go-fair.org>

Short Bio

Erik Schultes is an evolutionary biologist with a long standing interest in data-intensive research questions. Erik has been affiliated with the University of California, Los Angeles; The Whitehead Institute for Biomedical Research (MIT); The Santa Fe Institute; Duke University Medical Centre and Department of Computer Science. Erik currently holds a visiting researcher appointment in the Human Genetics Department, Leiden University Medical Center. In recent years, Erik has been working with a diverse community of stakeholders to develop FAIR data and services both within, and outside, the life sciences.

Damien Lecarpentier

Title

How FAIR is EOSC? An overview of FAIR activities within the various EOSC funded initiatives

Abstract

EOSC is transforming the way data infrastructures are collaborating and providing research data services in Europe. FAIR is becoming widely adopted as a concept to manage research data and is often presented as cornerstone of EOSC implementation. But how does EOSC actually support FAIR? This presentation will provide an overview of the current activities being planned around FAIR in the various EOSC projects (EOSC pilot, EOSC-Hub, new ESFRI clusters, etc.) and their expected impact, identifies possible shortcomings and shed light on some of the recent initiatives being prepared to make FAIR the foundational principles of EOSC.

Affiliation

Programme Director, Research Data

CSC – IT Center for Science

Short Bio

Dr. Damien Lecarpentier (M) is a Programme Director at CSC for research data. He has been managing the EUDAT initiative since its inception in 2011 and has been involved as work package and task leader in several EC-funded projects in the areas of grids, HPC, and e-Infrastructure policy before taking responsibility of EUDAT. Damien has also contributed to the establishment of the Research Data Alliance (RDA) in Europe, through the iCORDI and the RDA support projects, and is currently actively involved in activities related to the European Open Science Cloud via the EOSCpilot and the EOSC-hub in which he leads the contribution of the EUDAT CDI.