

Facilitating the discovery and identification of EOSC resources

Mike Bennett, Matthew Buys, Stavroula Eleftherakis, George Katsogiannis-Meimarakis, Paolo Manghi, Thanasis Vergoulis

Topics to be covered

Background. The European Open Science Cloud (EOSC) is a pan-European virtual environment for sharing and accessing research resources across borders and scientific disciplines. It provides scientists with a vast array of research services and products. However, the extensive array of content poses challenges for researchers in discovering and identifying resources of interest.

FAIRCORE4EOSC is a project¹ that aims to develop a set of core components for EOSC, focusing on supporting data FAIRness and addressing gaps identified in the EOSC Strategic Research and Innovation Agenda (SRIA). Among others, FAIRCORE4EOSC will deliver RDGraph and PID Graph, two components that can facilitate the discovery and identification of EOSC resources.

RDGraph builds upon the EOSC catalogue's content, extending it with additional entities like the Research Activity Identifiers (RAIDs) and offering a variety of advanced discovery functionalities, including natural language search, ranking of results based on impact indicators, community-based recommendations, and research activity inference. All RDGraph functionalities are accessible via UI/UX tools, regular data dumps and APIs.

PID Graph, is a PID-focused graph building upon the contents of the EOSC catalogue and the FREYA PID Graph. It offers data about, and connections between, a variety of persistent identifiers, including DOIs (for scholarly outputs), ORCID IDs (for researchers), ROR IDs (for institutions) and others. This information is made available via REST, GraphQL and OAI-PMH APIs and data dumps that expose DOI metadata (integrated with data usage statistics) from DataCite for streamlined ingestion by EOSC and related services. The PID Graph service also ingests additional links between PID entities from other EOSC sources and integrates these into the graph.

Tutorial topics & format. In this tutorial, the instructors will present the main functionalities of both components and explain the underlying technologies. The functionalities will be presented using live demonstrators offering also the option for a hands-on experience by the audience. The estimated duration of the tutorial will be 2 hours. The topics to be presented during the tutorial will be the following:

¹ FAIRCORE4EOSC: <https://faircore4eosc.eu/>

- *Part A: Introduction (15 min)*
 - Quick introduction to EOSC and FAIRCORE4EOSC
 - Background & motivation
- *Part B: RDGraph (45 min)*
 - Introduction & main functionalities
 - Live demonstration
- *Part C: PID Graph (30 min)*
 - Introduction & main functionalities
 - Live demonstration
- *Part D: Open discussion (30 min)*
 - Q&A
 - Feedback from the audience

Expected outcome. The members of the audience are expected to gain practical experience on the RDGraph and PID Graph components and how they can be useful for discovering and identifying EOSC resources. They are also expected to improve their knowledge on scholarly knowledge discovery and the technologies used to support the two components.

Setting

The tutorial will include both slide presentations and live demos, showcasing software deployed on remote servers. Consequently, a projector and a stable internet connection will be necessary in the room.

Prerequisites for attendants

Having a foundational understanding of EOSC and scholarly knowledge management terminology will be beneficial. Additionally, possessing basic programming skills or familiarity with using APIs and understanding JSON file syntax will also be advantageous.

Instructors

The instructors of this tutorial, have extensive expertise in the respective field and they are directly involved in the design and development of the RDGraph and PIDGraph components. Short bios of the instructors (ordered alphabetically based on last name) follow.

Mike Bennett is an application developer at DataCite where he is the technical lead on the development of the FAIRCORE4EOSC PID Graph. His focus is on building open and interoperable systems to enhance the availability and usability of scholarly knowledge. Outside of DataCite, he sits as a co-chair of a Technical Specification Group and on the Technical Review and Coordinating Committees for the International Image Interoperability Framework (IIIF) and previously he worked at the University of Edinburgh and the Bodleian Library at

Oxford University where he was involved in a variety of projects across subjects as diverse as Broadside Ballads, Epidemiology, and the legal history of the Scottish Court of Session.

Matthew Buys is the Executive Director of DataCite where he leads a passionate and committed team who provides the means to create, find, cite, connect, and use research globally. Matt's work is focused on building DataCite as a sustainable global community, through efforts that support and advance community-driven practices.

Stavroula Eleftherakis is currently pursuing her PhD at the University of Grenoble, under a co-supervision arrangement with the Athena Research Center in Greece, where her research is centered on enhancing the fairness of recommendation systems. She earned her BSc degree from the Department of Actuarial and Financial Mathematics at the University of the Aegean. Subsequently, she obtained her MSc degree from the Data Science and Information Technologies (DSIT) master's program at the National and Kapodistrian University of Athens.

George Katsogiannis-Meimarakis is a PhD student at the University of Grenoble in co-supervision with the Athena Research Center in Greece. His work focuses on data accessibility through natural language. He is a graduate of the Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens, where he obtained a BSc and MSc degree in Computer and Data Science, respectively. In the past, he has presented tutorials on deep learning for Text-to-SQL and Natural Language Interfaces at various conferences [EDBT'21, SIGMOD'21, TWC'22, WSDM'23, VLDB'23].

Dr. Paolo Manghi (M) Researcher in computer science at Istituto di Scienza e Tecnologie dell'Informazione (ISTI) of Consiglio Nazionale delle Ricerche (CNR), in Pisa, Italy. His research areas of interest are today data e-infrastructures for science and scholarly communication infrastructures, with a focus on technologies supporting open science publishing within and across different disciplines, i.e. computational reproducibility and transparent evaluation of science. Since April 2021 he is employed as CTO of the OpenAIRE AMKE (NPP company) and acts as coordinator of the H2020 OpenAIRE-Nexus project.

Dr. Thanasis Vergoulis (M) is the Development and Operation Director of OpenAIRE AMKE and a scientific associate at IMSI, "Athena" RC in Greece. He received his Ph.D. in Computer Science from NTU of Athens in 2014. Since then, he has been involved in several EU and national ICT projects involving big data management, scholarly knowledge representation and management, scientometrics, research analytics, and bioinformatics. He has served as a member of the program or organising committee for several CS conferences and workshops, and as a guest editor in special issues of well-established journals, while he has co-edited a book titled "Predicting the Dynamics of Research Impact" (Springer, 2021). He has been teaching courses in undergraduate and postgraduate levels in academic institutions in Greece and Cyprus.