Tutorial Title:

Exploring the OpenAIRE Graph on Google Big Query

Instructors' Names and Affiliations:

Dr. Andrea Mannocci, CNR-ISTI, Pisa, Italy. Contact: andrea.mannocci@isti.cnr.it
Dr. Alysson Fernandes Mazoni, University of Campinas, Brazil. Contact: afmazoni@unicamp.br

Tutorial Objectives and Learning Outcomes:

Scientometric research is traditionally anchored to proprietary data; however, more recently, the global research community involved in this discipline is increasingly committed to openness, transparency, and responsible research assessment principles. This is reflected in pivotal initiatives such as the Barcelona Declaration on Open Research Information [1], the Agreement on Responsible Research Assessment (ARRA) [2], the Coalition for Advancing Research Assessment (CoARA) [3], and the San Francisco Declaration on Research Assessment (DORA) [4].

All these frameworks advocate for using open, interoperable, and reusable research data for the benefit of Open Science practices and transparency and to ensure equitable and responsible evaluation practices. Open data facilitates reproducibility, fosters innovation, and promotes inclusivity by enabling researchers worldwide to access and analyse comprehensive datasets without restrictive barriers. By leveraging open data, the scientometric community can better align with the principles of transparency and accountability while addressing pressing global challenges in research evaluation and policy development.

To target these increasing demands, a number of sources of data that can potentially fuel scientometric inquiry are recently being developed, such as OpenAlex [5] (coming from the now defunct Microsoft Academic Graph), and OpenCitations [6-7], to name a few. This tutorial is tailored around and focuses on the OpenAIRE Graph [8] (hereafter the Graph, for brevity), which can complement this landscape by providing a thorough perspective on Open Science and by parting from a strictly literature-based representation of the global scientific record.

However, the Graph can appear impractical for most users by its sheer size as it contains nearly 240 million research outputs beyond publications, 320,000 organisations, 3 million projects, and over 5 billion relationships between these entities collected from over 130,000 data sources. Its most direct access as a whole is a dataset of about 270 GB compressed JSON files deposited on Zenodo [9], which can be a major hurdle if nothing other than processing on a local machine is available.

To address this accessibility challenge, OpenAIRE is currently committed to offering access to the Graph via the Google Cloud platform, where the Graph is shared as a public dataset hosted in the cloud, which can be explored and queried from Google Cloud solutions, such as BigQuery. That way, anyone with a Google account can explore the files and run queries at a low cost. Only the cost of processing is important, given that the data themselves are public. Also, for researchers, Google's running policy is to provide research credits after submitting a valid case study [10]. The credits provided for case studies are enough for short and many medium term projects and a rough estimation of prices is about 5 euros each terabyte of queries.

To facilitate the uptake of this solution, this tutorial, whose processing costs on Google Big Query are covered by OpenAIRE, will aim to

- Provide a practical introduction to Google BigQuery, showcasing its capabilities for analysing large-scale open datasets;
- Guide participants in utilizing the OpenAIRE Graph, a rich dataset offering insights into research outputs beyond traditional publications, funding information, organisations, and more;
- Empower attendees to design and execute scientometric analyses using open tools and datasets.

By the end of the tutorial, participants will:

- Understand the strategic importance of open data in scientometrics research and responsible research assessment.
- Gain practical experience in using Google BigQuery to analyse the OpenAIRE Graph.
- Develop skills to design and implement their own scientometric studies using open datasets.
- Be equipped to advocate for and apply open data principles in their own work.

References

[1] Barcelona Declaration on Open Research Information, https://barcelona-declaration.org

[2] Agreement on Responsible Research Assessment (ARRA), https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf

[3] Coalition for Advancing Research Assessment (CoARA), https://coara.eu

[4] San Francisco Declaration on Research Assessment (DORA), https://sfdora.org/read

[5] Priem, Jason, Heather Piwowar, and Richard Orr. "OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts." arXiv, June 16, 2022. https://doi.org/10.48550/arXiv.2205.01833.

[6] Peroni, Silvio, and David Shotton. "OpenCitations, an Infrastructure Organization for Open Scholarship." *Quantitative Science Studies* 1, no. 1 (2020): 428–44. https://doi.org/10.1162/qss_a_00023.

[7] Massari, Arcangelo, Fabio Mariani, Ivan Heibi, Silvio Peroni, and David Shotton. "OpenCitations Meta." *Quantitative Science Studies* 5, no. 1 (March 1, 2024): 50–75. https://doi.org/10.1162/qss_a_00292.

[8] OpenAIRE Graph, https://graph.openaire.eu

[9] Manghi, P., Atzori, C., Bardi, A., Baglioni, M., Dimitropoulos, H., La Bruzzo, S., Foufoulas, I., Mannocci, A., Horst, M., Iatropoulou, K., Kokogiannaki, A., De Bonis, M., Artini, M., Lempesis, A., Ioannidis, A., Manola, N., Principe, P., Vergoulis, T., & Chatzopoulos, S. (2025). OpenAIRE Graph Dataset (9.0.0) [Data set]. OpenAIRE. https://doi.org/10.5281/zenodo.14582029

[10] Google Cloud research credits, https://curc.readthedocs.io/en/latest/cloud/gcp/Google-Cloud-research-credits.html

Target Audience and Prerequisites:

The tutorial targets researchers, PhD candidates, master students and data analysts working in the fields of scientometrics and bibliometrics who are willing to scale up the scope of their research and focus on the broader aspects of Open Science.

No background knowledge or skill is mandatorily requested to join the tutorial; however, having SQL rudiments can be beneficial. Similarly, experience with Python, Jupyter Notebooks, and Python libraries for data science, such as pandas, can support more advanced exercises.

During the tutorial, we will run a refresher of the relevant theoretical and practical concepts so that everyone can perform, or at least, follow the hands-on walkthrough.

Tutorial Length and Format:

We foresee a half-day tutorial, which will include explanations supported by slides in preparation for the hands-on walkthrough and solo exercises.

For the sake of interaction and participation, we prefer not to deliver the tutorial in hybrid mode (i.e., with a share of participants connected remotely).

Tutorial Outline and Content:

The following (tentative) outline is suggested:

- **Introduction to Google Cloud and Big Query platform:** Overview of the Google Cloud platform, its potential, costs and opportunities.
- **Introduction to the OpenAIRE Graph on BigQuery:** Overview of the dataset, the modelled entities and relations. We will describe how the data is structured, the information contained and highlight the unique opportunities in relation to other datasets available in the state of the art (e.g., WoS, Scopus, OpenAlex)
- **Gentle introduction to SQL:** We will deliver a refresher on SQL syntax and clauses for selecting, joining and aggregating data.
- **Simple queries, walkthroughs and exercises:** The audience will familiarise themselves with the OpenAIRE Graph data by starting with simple tasks and then moving to increasingly more structured queries and data explorations that get close to research questions.
- **Advanced queries walkthrough and exercises:** The use of join functions to connect information present in more tables or different datasets.
- **Data take-out and data analysis in Python notebooks:** Participants will query the data, bearing in mind that data will be exported to Python for further off-the-cloud analysis.

Required Materials and Software:

The tutorial will be structured in such a way that participants can follow theory and hands-on exercises without the need to replicate the queries mandatorily. However, if a participant would like to do so (which we suggest), they should **bring a laptop capable of Internet connection**.

In order to actively participate in hands-on exercises, participants are requested to have **an active Google account, to which they currently have access**.

Participant Requirements:

Relevant datasets and the Big Query platform will be set up beforehand, so participants are not required to perform any action in preparation for the event.

Key Takeaways:

Participants will familiarise themselves with the Google Big Cloud platform and how this can be used to scale up scientometrics research. Participants will also familiarise themselves with the OpenAIRE Graph and the data it contains, and they will gain proficiency in translating high-level research questions into practical queries against the offered datasets, as well as interpreting the results and performing troubleshooting.

We would like to highlight that the methodology based on Google Big Query is, in fact, data-agnostic, and while the tutorial is tailored around the OpenAIRE Graph, the general concepts and methods here introduced can be applied to other datasets, be they open or proprietary.

Instructor Backgrounds:

Dr. Andrea Mannocci is a Research Fellow at the InfraScience Laboratory within the Institute of Information Science and Technologies (ISTI), part of the Italian National Research Council (CNR) in Pisa, Italy. He holds a Ph.D. in Information Engineering from the University of Pisa. Before his current appointment, he was a Research Associate at the Knowledge Media Institute (KMI) of the Open University in Milton Keynes, UK, where he contributed to the SKM3 research team (Scholarly Knowledge Modelling, Mining, and Sense Making) and specialised in applying data science techniques to scholarly big data and research analytics.
His work focuses on the development, analysis and uptake of the OpenAIRE Graph for Scientometrics, with research interests on Open Science practices and Responsible Research Assessment.
Keen on Open Research Information and Open Scientometrics advocate, he is chairing the RDA WG on Scientific Knowledge Graphs - Interoperability Framework (SKG-IF), which focuses on enabling the exchange of information across different initiatives modelling research information.

Dr. Alysson Fernandes Mazoni holds a background as a control and automation engineer with a PhD on applications of Machine Learning at the University of Campinas, Brazil. Currently, he conducts research on quantitative methods for Geosciences and Scientometrics as applied to innovation economics and scientific information systems for the Global South. Affiliated as a post-doctoral research fellow also at University of Campinas.