Mapping scientific communities at scale

Victor Barbier[1] and Eric Jeangirard[1]

[1]French Ministry of Higher Education and Research, Paris, France

January 2025

Keywords: open access, open science, open data, open source

1. Motivation

Analysing and mapping scientific communities provides an insight into the structure and evolution of academic disciplines. This involves providing an analytical and visual representation of the relationships between entities (e.g. researchers, research laboratories, research themes), with the aim, in particular, of understanding the networks and dynamics of scientific collaboration, and identifying collaborative groups and their influences. From the point of view of decision-makers, this type of tool is useful for strategic decision-making with a view to public policy and funding.

These maps are generally deduced from data in bibliographic databases (open or proprietary), based on co-publication or citation information. In the case of co-publications, two entities (authors, for example) will be linked if they have collaborated (co-published) on a piece of research. These links are then symmetrical. In the case of citation links, two authors will be linked if one cites the research work of another, in the list of references. This is a directed link, as one author may cite another without this being reciprocal. A lot of recent work uses this second approach, for example by trying to calculate composite indicators of novelty (or innovation) based on citation links.

The quality and completeness of the bibliographic metadata used are, of course, essential if we are to produce a relevant map. Today, the quality of open citation data still needs to be improved, cf (Alperin et al. 2024). On the other hand, it is possible to obtain quality metadata on publications (and therefore links to co-publications). For example, the French Open Science Monitor (BSO) has compiled a corpus of French publications with good coverage cf (Chaignon and Egret 2022). This corpus is exposed in the French research portal scanR (Jeangirard 2024). This is a corpus containing about 4 millions publications in all disciplines. These publications have been enriched with disambuation persistent identifier (PID) on authors, affiliations and topics. For authors and affiliations, French-specific PID have been used (idref for authors and RNSR for laboratories) because they have the best coverage, even if not perfect. For topics, wikidata identifiers has been used cf (Foppiano and Romary 2020). Other enrichments, like software detection are also present, and thus usable as entities to map.

1.1 Current limits of the scanR application

1.3 Network analysis limits

2. Network analysis at scale

Focusing on strongest interactions

Elasticsearch impl

VOSviewer implem

LLM trick

3. Making insightful maps

Citation / hot topics

User interaction

References

Alperin, Juan Pablo, Jason Portenoy, Kyle Demes, Vincent Larivière, and Stefanie Haustein. 2024. "An Analysis of the Suitability of Openalex for Bibliometric Analyses." https://arxiv.org/abs/2404.17663.

Chaignon, Lauranne, and Daniel Egret. 2022. "Identifying Scientific Publications Countrywide and Measuring Their Open Access: The Case of the French Open Science Barometer (Bso)." *Quantitative Science Studies* 3 (1): 18–36. https://doi.org/10.1162/qss_a_00179.

Foppiano, Luca, and Laurent Romary. 2020. "Entity-Fishing: A Dariah Entity Recognition and Disambiguation Service." *Journal of the Japanese Association for Digital Humanities* 5 (1): 22–60.

Jeangirard, Eric. 2024. "scanR - Explore public data on French research and innovation." In *euroCRIS SMM 2024*. Paris, France: euroCRIS. https://hal.science/hal-04813230.