Mapping scientific communities at scale

Hafsa Aallat[1]

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

February 2025

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

1. Motivation

1.1 Presentation of IPCC and IPBES: Working Groups and dates

The IPCC (Intergovernmental Panel on Climate Change) assesses scientific information on climate change, providing reports to guide policymakers. It has three working groups sees as three main themes :

- Working Group I (WGI) focuses on the **physical science** of climate change.
- Working Group II (WGII) examines climate change impacts, **adaptation**, and vulnerabilities.
- Working Group III (WGIII) addresses climate change **mitigation** strategies.

The Sixth Assessment Report (AR6) was released in stages between 2021 and 2022.

The IPBES (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services), established in 2012, assesses biodiversity and ecosystem services. It produces thematic and regional assessments, with the **Global Assessment Report (2019)** highlighting biodiversity loss and the need for urgent action.

Both platforms provide crucial scientific assessments that inform global climate and biodiversity policies.

1.2 Limits of the French Court of Audit study

In 2023, the French Court of Audit conducted a study on France's scientific output related to environmental transition. After hearings with the Directorate General for Research and Innovation (DGRI) and research operators, the Court analyzed the bibliography cited in the sixth IPCC report. The study found that French publications are the most cited in the physical sciences of climate change, highlighting the global impact of French research in this field.

However, this evaluation has important limitations. The IPCC bibliography is based on high-impact publications often from top journals, making it quite selective. This selection prioritizes more visible and well-known works, leaving out other important research that may not be as visible but still in the same themes as IPCC report. While this reflects France's scientific excellence, it does not fully represent the diversity of French scientific contributions to ecological transition.

1.3 How can we explore and recognize french publications related to the same topics as IPCC report from a global point of view ?

To fill this gap, we propose using a larger dataset, such as scanR. **ScanR has a significantly higher coverage** of publications with at least one French affiliation compared to other sources, contributing 92% to the overall aggregated corpus. This is much higher than databases like Scopus (67%), WoS (58%), or PubMed (29%), making ScanR a more comprehensive tool for capturing French scientific publications (Chaignon and Egret 2022). Unlike the IPCC's restricted approach, ScanR includes publications with at least one French affiliation, showing a larger view of research. This could allow us to capture a more diverse range of topics related to climate change physical science, adaptation and mitigation.

Initially, we will replicate the Court of Audit analysis of the IPCC bibliography to identify the main themes and their proportion of French contributions. Then, we will expand our study to know the top institutions, labs, regions, and researchers that provide solutions to the challenges of environemental transition in France, based on IPCC bibliography. In a second time, we will create a model that can recognize a publication about IPCC similar topics, and apply the model to scanR publications. At the same time, we will conduct a similar analysis for the IPBES bibliography, following the same approach to identify the French contributions, and exploring less visible but valuable research related to biodiversity and ecosystem services.

2. IPCC and IPBES Bibliography Analysis and Model

We propose a method to analyze the bibliographies of IPCC and IPBES reports.

2.1 Data Collection and Cleaning For each report, we collect the references:

- For IPCC report, we collect citations in .bib format for each chapter of each working group (n.d.a).
- For IPBES report, we gather all citations via Zotero (n.d.b).

Once the data is collected, we clean the DOI (Digital Object Identifier) of each publication. The DOI should follow a specific format starting with '10.'. Any publication without a valid DOI is not considered.

2.2 Data Enrichment After cleaning, the data contains features such as DOI, title, and main author. However, we still lack information such as institutions, researchers, countries, and topics associated with each publication. To fill in the gap, we enrich the data by importing additional features from OpenAlex for each publication with a valid DOI. These features include: countries, year, topics, title, author names, institutions, RORs (Research Organization Registry) and journals.

OpenAlex is an international open-access database that provides metadata on research papers, authors, journals, and institutions. It aims to make academic information more accessible and supports data analysis and knowledge discovery in various fields. OpenAlex is a valuable tool for researchers and educators. We use the Api to import the features.

Next, we use the Biblioglutton Python library to fill in missing DOIs based on the title and main author. We also verify that the year retrieved from OpenAlex matches the year in the original dataset.

2.3 Data storage and visualization

Once the data is enriched with openAlex features, we edit the data and push them on a cluster elastic-search. As the exemple, for one publication:

```
{
  "doi": "10.1126/science.aaw6974",
  "year": "2018",
  "title": "Impacts of 1.5 °C global warming on natural and human systems",
  "rors": [
    ["https://ror.org/00rqy9422", "AU"],
    ["https://ror.org/03ztgj037", "DE"],
    ["https://ror.org/03fkc8c64", "JM"],
    ["https://ror.org/03ztgj037", "DE"],
    ["https://ror.org/04jr1s763", "IT"],
    ["https://ror.org/01ryk1543", "GB"],
    ["https://ror.org/05wwcw481", "GB"],
    ["https://ror.org/05k07f122", "AR"],
    ["https://ror.org/03cqe8w59", "AR"],
    ["https://ror.org/0081fs513", "AR"],
    ["https://ror.org/05sbt2524", "FR"],
    ["https://ror.org/02feahw73", "FR"],
    ["https://ror.org/02rx3b187", "FR"],
    ["https://ror.org/01wwcfa26", "FR"],
    ["https://ror.org/05q3vnk25", "FR"],
    ["https://ror.org/03vmsb260", "JP"],
    ["https://ror.org/02j4mf075", "ID"],
    ["https://ror.org/00cvxb145", "US"],
    ["https://ror.org/03rp50x72", "ZA"],
    ["https://ror.org/04ex24z53", "FR"],
    ["https://ror.org/035xkbk20", "FR"],
    ["https://ror.org/01pa4h393", "FR"],
    ["https://ror.org/05q3vnk25", "FR"],
    ["https://ror.org/02feahw73", "FR"],
    ["https://ror.org/02hw5fp67", "JP"],
    ["https://ror.org/00ae7jd04", "US"],
    ["https://ror.org/013meh722", "GB"],
    ["https://ror.org/0524sp257", "GB"],
    ["https://ror.org/032e6b942", "DE"],
    ["https://ror.org/05a28rw58", "CH"],
    ["https://ror.org/02yr08r26", "DE"],
    ["https://ror.org/01c8qhb70", "BS"],
    ["https://ror.org/040tfy969", "GB"],
    ["https://ror.org/026k5mg93", "GB"],
    ["https://ror.org/034b53w38", "CN"]
  ],
  "ipcc": [
    { "name": "wg1_chap_01", "wg": "1", "chap": 1 },
    { "name": "wg2_chap_01", "wg": "2", "chap": 1 },
    { "name": "wg2_chap_02", "wg": "2", "chap": 2 },
    { "name": "wg2_chap_02", "wg": "2", "chap": 2 },
    { "name": "wg2_chap_04", "wg": "2", "chap": 4 },
    { "name": "wg2_chap_07", "wg": "2", "chap": 7 },
    { "name": "wg2_chap_08", "wg": "2", "chap": 8 },
    { "name": "wg2_chap_12", "wg": "2", "chap": 12 },
    { "name": "wg2_chap_13", "wg": "2", "chap": 13 },
    { "name": "wg2_chap_14", "wg": "2", "chap": 14 },
```

```
    { "name": "wg2_chap_15", "wg": "2", "chap": 15 },
    { "name": "wg2_chap_15", "wg": "2", "chap": 15 },
    { "name": "wg2_chap_16", "wg": "2", "chap": 16 },
    { "name": "wg2_cross_chap_1", "wg": "2_cross", "chap": 1 },
    { "name": "wg2_cross_chap_4", "wg": "2_cross", "chap": 4 },
    { "name": "wg2_cross_chap_4", "wg": "2_cross", "chap": 4 },
    { "name": "wg3_chap_01", "wg": "3", "chap": 1 },
    { "name": "wg3_chap_04", "wg": "3", "chap": 4 }
  ],
  "authors_name": [
    ["Ove Hoegh-Guldberg", ["AU"]],
    ["Daniela Jacob", ["DE"]],
    ["Michael A. Taylor", ["JM"]],
    ["Tania Guillén Bolaños", ["DE"]],
    ["Marco Bindi", ["IT"]],
    ["Sally Brown", ["GB"]],
    ["Inés Angela Camilloni", ["AR"]],
    ["Arona Diedhiou", ["FR"]],
    ["Riyanti Djalante", ["ID", "JP"]],
    ["Kristie L. Ebi", ["US"]],
    ["François Engelbrecht", ["ZA"]],
    ["Joël Guiot", ["FR"]],
    ["Yasuaki Hijioka", ["JP"]],
    ["Shagun Mehrotra", ["US"]],
    ["Chris Hope", ["GB"]],
    ["Antony J. Payne", ["GB"]],
    ["Hans-Otto Pörtner", ["DE"]],
    ["Sonia I. Seneviratne", ["CH"]],
    ["Adelle Thomas", ["BS", "DE"]],
    ["Rachel Warren", ["GB"]],
    ["Guangsheng Zhou", ["CN"]]
  ],
  "institutions_names": [
    ["University of Queensland", "AU"],
    ["German Climate Computing Centre", "DE"],
    ["University of the West Indies", "JM"],
    ["German Climate Computing Centre", "DE"],
    ["University of Florence", "IT"],
    ["University of Southampton", "GB"],
    ["Bournemouth University", "GB"],
    ["Instituto Franco-Argentino sobre Estudios de Clima y sus Impactos", "AR"],
    ["Consejo Nacional de Investigaciones Científicas y Técnicas", "AR"],
    ["University of Buenos Aires", "AR"],
    ["Grenoble Institute of Technology", "FR"],
    ["French National Centre for Scientific Research", "FR"],
    ["Université Grenoble Alpes", "FR"],
    ["Institute of Environmental Geosciences", "FR"],
    ["Institut de Recherche pour le Développement", "FR"],
    [
      "United Nations University Institute for the Advanced Study of Sustainability",
      "JP"
    ],
```

```
    ["Haluoleo University", "ID"],
    ["University of Washington", "US"],
    ["University of the Witwatersrand", "ZA"],
    ["Collège de France", "FR"],
    ["Aix-Marseille University", "FR"],
    ["Centre for Research and Teaching in Environmental Geoscience", "FR"],
    ["Institut de Recherche pour le Développement", "FR"],
    ["French National Centre for Scientific Research", "FR"],
    ["National Institute for Environmental Studies", "JP"],
    ["World Bank", "US"],
    ["University of Cambridge", "GB"],
    ["University of Bristol", "GB"],
    [
      "Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung",
      "DE"
    ],
    ["ETH Zurich", "CH"],
    ["Climate Analytics", "DE"],
    ["College of The Bahamas", "BS"],
    ["Tyndall Centre", "GB"],
    ["University of East Anglia", "GB"],
    ["Chinese Academy of Meteorological Sciences", "CN"]
  ],
  "countries": [
    "CHN",
    "GBR",
    "FRA",
    "ITA",
    "AUS",
    "JAM",
    "DEU",
    "JPN",
    "ZAF",
    "USA",
    "BHS",
    "CHE",
    "IDN",
    "ARG"
  ],
  "ipbes": [{ "chapter": "4" }],
  "topics": [
    "Impact of Climate Change on Human Migration",
    "Geoengineering and Climate Ethics",
    "Economic Implications of Climate Change Policies"
  ]
}
```


After that we used Highcharts, a graphic tool to visualize the graph.

2.4 Create a database

2.5 Train the model

3. Results

3.1

3.2 Custom perimeter

scanR offers this mapping tool for the entire indexed corpus, but it is also possible to adapt the tool to a restricted perimeter, at the user's discretion. For example, an institution or laboratory can define its own corpus (based on a list of publications) and a mapping tool dedicated to this perimeter is automatically created. Technically, elasticsearch queries are the same, with just an additional filter to query only the publications within the perimeter. The tool can be embedded in any website using an iframe. It's the same principle as the local barometer. This approach eliminates the need for automatic alignment of affiliations, which remains a highly complex task. Automation is possible to a certain extent (**???**), but human curation remains necessary in the majority of cases (Jeangirard, Bracco, and L'Hôte 2024). In this way, users retain control over the definition of their perimeter, and can, if they wish, have several distinct perimeters.

4. Code availibility

The code developed for the scanR web application is open source and available online on GitHub https://github.com/dataesr/scanr-ui

References

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.

Jeangirard, Eric, Laetitia Bracco, and Anne L'Hôte. 2024. "Works-magnet : aucune de perdue, 10 000 de retrouvées." Abes; Journées Abes 2024. https://doi.org/10.5281/zenodo.11471247.

n.d.a. https://www.ipcc.ch/report/ar6.

n.d.b. https://www.zotero.org/groups/2333077/ipbes_global_assessment/library.