Mapping STI ecosystems via Open Data: overcoming the limitations of conflicting taxonomies. A case study for Climate Change Research in Denmark

Nicandro Bovenzi1, Nicolau Duran-Silva1, Francesco Alessandro Massucci1*, Francesco Multari1, César Parra-Rojas1 and Josep Pujol-Llatse1

1SIRIS Lab, Research Division of SIRIS Academic, 08003, Barcelona, Spain

Abstract
Science, Technology and Innovation (STI) decision-makers often need to have a clear vision of what is researched and by whom to design effective policies. Such a vision is provided by effective and comprehensive mappings of the research activities carried out within their institutional boundaries. A major challenge to be faced in this context is the difficulty in accessing the relevant data and in combining information coming from different sources: indeed, traditionally, STI data has been confined within closed data sources and, when available, it is categorised with different taxonomies. Here, we present a proof-of-concept study of the use of Open Resources to map the research landscape on the Sustainable Development Goal (SDG) 13 – Climate Action, for an entire country, Denmark, and we map it on the 25 ERC panels.

Keywords
Science mapping, Text mining, Deep learning, Open data repositories, Sustainable Development Goals

1. Introduction
To inform their decisions, policy-makers in the Science, Technology and Innovation (STI) sector typically need "maps", either at a territorial or at an institutional level, to understand what is researched and by whom. Generally, those maps need to provide information about the research and innovation topics and about the relevant actors linked to them, so that effective policy-actions could be proposed, by covering the right scientific domains and by being catered for the adequate users. These maps need to be comprehensive to extensively cover i. the whole STI value chain (from basic research up to industrial innovation), ii. the different scientific domains and iii. all possible relevant actors. As such, these maps should rely on different data sources that could offer the broadest possible view of STI inputs and outputs. Some major challenges faced at a policy level arise because many of those data sources are not openly available (undermining therefore the participatory processes), they are not interoperable in terms of data classification schemes and institutional identification (therefore limiting transversal analyses) and they are hardly manageable by non-expert users [1].

In this paper, we present a proof of concept of merging different open datasets and of analysing
them with a common classification scheme specifically designed for the sake of our analyses. To do so, we gather data for the whole Danish\(^1\) STI ecosystem from 4 different data sources, namely: 

i. the CORDIS database\(^2\), through the UNICS [2] platform, for H2020–funded R&D projects, 

ii. the Kohesio linked–data portal\(^3\) for Regional R&D projects funded by the EC Cohesion initiative, 

iii. OpenAlex [3] and iv. OpenAIRE\(^4\) repositories for publications and other scientific outputs (these last being accessed through their respective APIs). After gathering these records, we use open knowledge–bases [5] and fine tune openly available Deep Learning models [6, 7] to:

i. **Identify STI documents linked with the Sustainable Goal 13, Climate Action [5]**, and

ii. **Categorise Danish research on Climate Action within the 25 panels of the European Research Council** (ERC).

In this way, we aim at showcasing how research in emerging fields (such as the SDGs [8]) can be gathered from open data sources and identified by means of modern, openly available AI models [9]. Finally, we demonstrate how gaps in taxonomic classifications across datasets may be filled by means of Deep Learning textual classifiers, by using the ERC panels as a paradigmatic example.

## 2. Main Results

We gathered, from a series of heterogeneous data sources, a dataset of scientific publications abstracts and R&D projects descriptions for the entire STI ecosystem of Denmark for the 2014–2019 time period. We then tagged each single record by means of a controlled vocabulary for SDG 13 – Climate Action [5] (that is, we identified the vocabulary terms in each single text by applying a series of textual matching rules). This enabled us to identify, within the initial dataset, all textual records linked with SDG 13.

The number of documents that we could identify in each data source as well as those we could link with SDG 13 are reported in Table 1. About 2% of Scientific publications in Denmark between 2014 and 2019, both from OpenAIRE and OpenAlex, are related to our SDG of interest. In contrast, European projects are more linked, in relative terms, to the issue of Climate Action: this comes as no surprise, given the orientation of EU policies towards the sustainability issues.

We finally applied Topic Modelling [10, 11] and we classified per ERC panels the SDG–related corpus, in order to obtain a series of topics characterising Climate Action–related research and to gain a “disciplinary” view of such research. In Fig. 1, we show a t-SNE visualisation of the automatically extracted topics [12] from textual data in publications (from OpenAlex and OpenAIRE) and projects (from CORDIS and Kohesio) concerning SDG 13.

In Fig. 1, each dot is a single document: as one can observe, we were able to extract a series of thematically different topics from the SDG 13 corpus, each dealing with a different aspect of climate action. At the centre of the figure, one finds topics related with the environment,

---

\(^1\)We choose Denmark as our case study because (i) it is a medium-size country and the size of its scientific production is such that one can practically retrieve all the documents from publications repositories, (ii) Danish R&D ecosystem is internationally visible and competitive and (iii) Denmark is internationally acknowledged as being one of the leading countries in terms of climate action policies and efforts (1st in the world according to the 2022 Environmental Performance Index (EPI)).

\(^2\)https://cordis.europa.eu/

\(^3\)https://kohesio.ec.europa.eu/
Table 1
Number of records with at least one author affiliation or beneficiary from Denmark (2014-2019) and relative volume mapped to SDG 13.

| Data source | Total Records in DK | Records related to SDG 13 |
|-------------|---------------------|--------------------------|
| OpenAlex    | 191,399             | 3,821 (2%)               |
| OpenAIRE    | 235,906             | 5,273 (2.2%)             |
| CORDIS      | 2,196               | 320 (14.6%)              |
| Kohesio     | 294                 | 14 (4.8%)                |

while going anti–clockwise from the top, it is possible to encounter topics related with energy, traditional and alternative fuels, emissions and pollution, impact on the biosphere and finally education and policy issues.

As a final pilot exercise, we proceeded to classify the documents linked to SDG 13 per ERC panels. To do so, we trained a Deep Learning textual classifier by fine-tuning [13, 6] the BERT-based SPECTER model [7] on a weakly supervised dataset, and we applied it to our Danish SDG 13 corpus. This effort allowed us to obtain a disciplinary classification of the records which is consistent across data sources and which may enable, in turn, a comparison of the Danish STI ecosystem with other geographical perimeters of interest. In Fig. 2, we present the distribution of documents by source and ERC panel. Perhaps surprisingly, the majority of the STI documents analysed were linked to Social Science issues, followed by Earth Sciences. Also, interestingly (and underscoring the importance of cross–platform analyses such as this one), one can see that the various data sources have a different coverage of the panels.

3. Conclusions

In this paper, we presented a proof–of–concept study of the use of Open Resources to map the research landscape on SDG 13 (Climate Action), for an entire country, Denmark. This type of mapping exercise is extremely useful for STI decision-makers, who, to design effective policies within their respective sphere of influence, need to have a clear vision of what is researched and by whom.

Here, we carried out a study of this sort by relying on Open Data for Research Projects (gathered from CORDIS and the Kohesio platform) and Scientific publications (collected from OpenAIRE and OpenAlex), by using an open vocabulary for mapping STI records on SDG 13 and by using openly available Deep Learning models to classify the corpus in accordance with the 25 ERC panels. The results we obtain are fairly encouraging: the coverage of the data analysed is extensive, both in absolute terms4 and in terms of scientific disciplines and actors. Interestingly, the data sources analysed offer a complementary view of the research domains, and allow one, when used in combination, to obtain a wide and precise overview of the local STI ecosystem.

4We made an overall comparison with results produced by Scopus: the number of records was much lower than what found both in OpenAIRE and OpenAlex, for the same country and time period
Figure 1: t-SNE visualisation of the 30 topics extracted from the SDG 13 corpus.

**Acknowledgments**

This work was partly funded by the European Commission H2020 Programme via the INODE project, under grant agreement No 863410.

**References**

[1] E. Fuster, F. Massucci, M. Matusiak, Identifying specialisation domains beyond taxonomies: mapping scientific and technological domains of specialisation via semantic analyses. (2020) 195–234. doi:10.4337/9781789905519.
Figure 2: Number of documents per source and by ERC Panel.

[2] X. Gimenez, A. Mosca, F. Roda, B. Rondelli, G. Rull, Unics: The open data platform for research and innovation, in: SEMANTiCS Posters&Demos, 2018.
[3] J. Priem, H. Piwowar, R. Orr, Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts (2022). URL: https://arxiv.org/abs/2205.01833. doi:10.48550/ARXIV.2205.01833.
[4] N. Rettberg, B. Schmidt, Openaire—building a collaborative open access infrastructure for european researchers., Liber Quarterly: The Journal of European Research Libraries 22 (2012).
[5] N. Duran-Silva, E. Fuster, F. A. Massucci, A. Quinquillà, A controlled vocabulary defining the semantic perimeter of sustainable development goals, Dataset, Zenodo (2019). URL: https://doi.org/10.5281/zenodo.4118028.
[6] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, in: CCL, 2019.
[7] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. S. Weld, Specter: Document-level representation learning using citation-informed transformers, 2020. arXiv:2004.07180.
[8] UN, Global indicator framework for the sustainable development goals and targets of the 2030 agenda for sustainable development (2017). URL: https://unstats.un.org/sdgs/indicators/indicators-list/.
[9] H. Chesbrough, Open innovation: a new paradigm for understanding industrial innovation, Open innovation: Researching a new paradigm 400 (2006) 0–19.
[10] T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National academy of Sciences 101 (2004) 5228–5235.
[11] V. R. Sangaraju, B. K. Bolla, D. K. Nayak, J. Kh, Topic modelling on consumer financial
protection bureau data: An approach using bert based embeddings, 2022.

[12] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (2008).

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).