Mapping STI ecosystems via Open Data: Overcoming the limitations of conflicting taxonomies.

A case study for Climate Change Research in Denmark

N. Bovenzi, N. Duran-Silva, F. Massucci, F. Multari, C. Parra-Rojas & J. Pujol-Llatse.

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

Introduction

To inform their decisions, policy-makers in the Science, Technology and Innovation (STI) sector typically need "*maps*", to understand what are the relevant research domains and key actors within their territorial or institutional boundaries of interest. Generally, those maps need to enable effective policy-actions, so that they should generally be comprehensive to extensively cover i. the whole STI value chain (from basic research up to industrial innovation), ii. the different scientific domains of relevance and iii. all possible pertinent 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 possible 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 an hypothetical analytical work to support STI policy-making which only makes use of open data to overcome the above challenges. To do so, we merge different open datasets and we analyse them with a common classification scheme.

After gathering the records from their respective data sources, we use open knowledge-bases [2] and text mining to:

- Identify STI documents linked with the Sustainable Development Goal (SDG) 13 - Climate Action,
- Categorise documents within the 25 panels of the European Research Council (ERC)
- Automatically identify thematic clusters by topic modelling

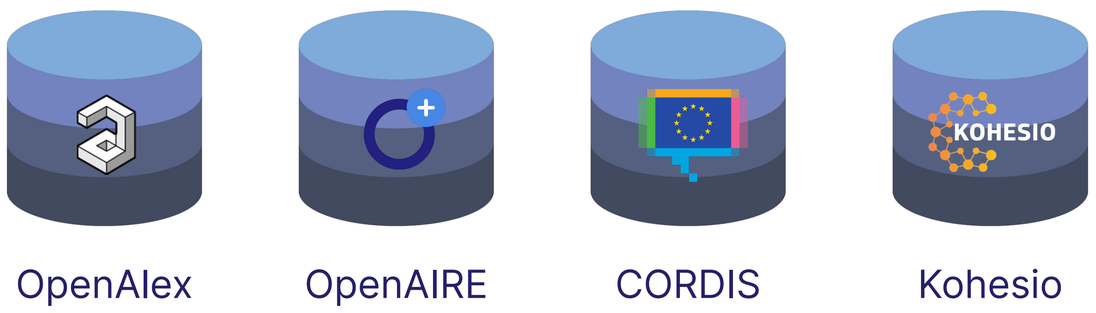
In this way, we aim at showcasing how research in emerging fields (such as the SDGs) can be gathered from open data sources and identified by means of modern, openly available AI models. Finally, we demonstrate how gaps in taxonomic classifications across datasets may be filled by means of Deep Learning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

Research Questions

In relation to Open Repositories, for the specific case of Science applied to Climate issues, the present work aims at tackling the following key questions:

1. Is it possible to use open repositories to inform policies in the field of STI?
2. How do the various available solutions cover the different disciplines?
3. What is the availability of textual information?
4. How can one obtain a reliable mapping of the available assets in a local STI ecosystem?
5. Is it possible to identify pertinent actors?

Data sources



OpenAlex OpenAIRE CORDIS Kohesio

Materials and methods

Task	Description
Mapping research concerning the Sustainable Development Goals (SDGs)	Text tagging based on a controlled vocabulary developed by SIRIS Academic, which combines expert knowledge with machine learning to scale the semantically rich "seed" concepts proposed by the official document [3].
Topic modelling	SPECTER (BERT-based model pre-trianed on scientific literature including citation information)[4] + KMeans clustering.
Text classification in the panels of the European Research Council (ERC)	Fine-tuning proposed by [5] of SPECTER model for text classification / training 25 single -label classifiers (one per ERC panel). The training set is built by using ERC grant proposals and publications, while tests are performed on proposals only, by obtaining an overall F-measure = 0.879.

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 [2] (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.

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%)

Table 1. Number of records with at least one author affiliation or beneficiary from Denmark (2014-2019) and relative volume mapped to SDG 13.

Besides thematic information, policy–making is best informed via the identification of relevant actors in the identified STI landscape. In this case, the identification of the most relevant actors across different data sources is affected by the different degree of affiliation disambiguation inside each individual source. So that the same actor can appear with several name variations in the different repositories.

We finally applied Topic Modelling in order to obtain both 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 from textual data in publications (from OpenAlex and OpenAIRE) and projects (from CORDIS and Kohesio) concerning SDG 13.

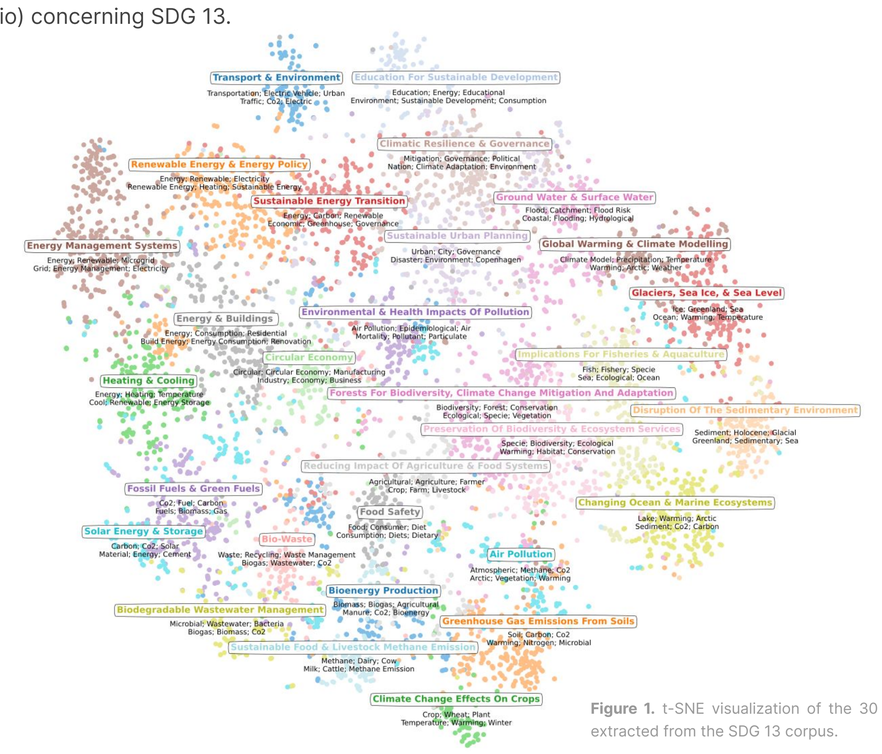


Figure 1. t-SNE visualization of the 30 topics extracted from the SDG 13 corpus.

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 the BERT-based SPECTER model [5, 6] 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 data source-independent 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 are differently distributed across panels.

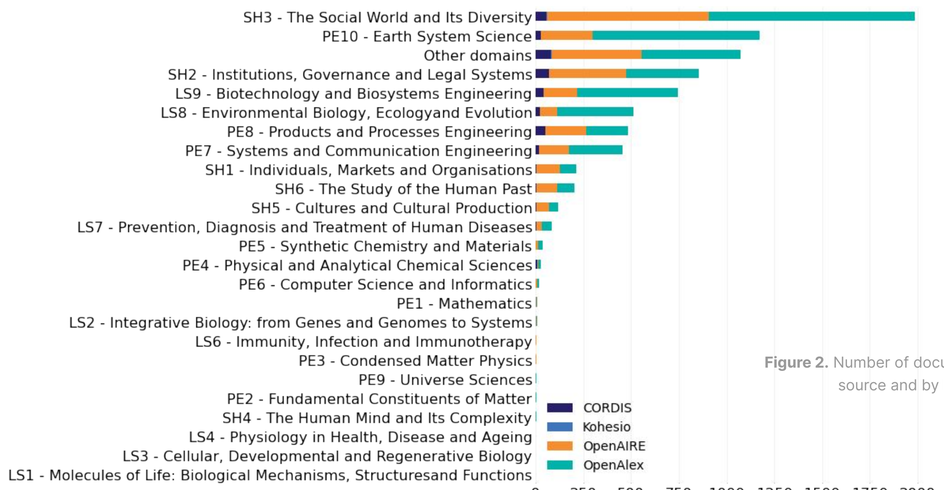


Figure 2. Number of documents per source and by ERC Panel.

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 terms 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.

References

[1] Fuster, E., Massucci, F. A., & Matusiak, M. (2020). Identifying specialisation domains beyond taxonomies: mapping scientific and technological domains of specialisation via semantic analyses. In *Quantitative Methods for Place-Based Innovation Policy* (pp. 195-234). Edward Elgar Publishing.

[2] Duran-Silva, N., Fuster, E., Massucci, F. A., & Quinquillà, A. (2019). A controlled vocabulary defining the semantic perimeter of Sustainable Development Goals. Dataset, *Zenodo.*

[3] United Nations. (2018). Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development.

[4] Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180.*

[5] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, October). How to fine-tune bert for text classification?. In *China national conference on Chinese computational linguistics* (pp. 194-206). Springer, Cham.

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

Acknowledgements

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

The full paper of this study is available here! →