Pitfalls and promises of BIR in science studies: A case study of mapping scientific articles to the United Nations Sustainable Development Goals

Frédérique Bordignon

Pitfalls and promises of BIR in science studies: A case study of mapping scientific articles to the United Nations Sustainable Development Goals

Frédérique Bordignon [1,2]

[1] *École des Ponts, Marne-la-Vallée, France*
[2] *LISIS, INRAE, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France*

Abstract

In this keynote, I show the pitfalls and promises of Bibliometric-enhanced Information Retrieval taking the concrete example of the bibliometrician facing the difficulty of matching Sustainable Development Goals (SDGs) with articles from an institutional corpus.

Keywords

Bibliometrics, scientometrics, SDGs

My background is at the crossroads of four disciplines, or at least disciplinary fields, all four of which undeniably feed the research related to BIR. I was trained as a linguist, and then I worked in the field of knowledge management and more specifically in the field of scholarly communication, which relates to library sciences. A significant part of my activity is dedicated to developing and promoting Open Science among researchers; I also work as a bibliometrician and I produce scientometric studies, on various topics.

I am a typical end-user of the tools and techniques produced by the BIR community. Whenever knowledge is converted into tools or techniques, I am at the end of the chain. Practically speaking, this means that I am not a computer scientist. This implies limitations in the use of the numerous tools that are produced by computer scientists, sometimes in a very minimal way, for example scripts written in Python prove challenging to run when you're not a programmer.

It is in fact as if, on a regular basis, when collecting data, retrieving information, analyzing, formatting and sharing it, there is a forbidden zone that I cannot access, even though it would allow me to do all this much more quickly. Therefore, my activity consists of a mix of Excel macros, free scraping tools (e.g.: Octoparse), conversion tools (e.g.: PDF to TXT extractors) and where CSV reigns supreme and is the backbone of any analysis done with Excel, Google Sheets and above all Tableau Software. But in this maze, I end up navigating and making the most of what is available.

In this presentation, I would like to shed light on what it means to be a bibliometrician working for an institution, in other words, what this implies in terms of professional skills and practices. I will use a practical case study that focuses on the mapping of scholarly articles to the

BIR 2022: 12th International Workshop on Bibliometric-enhanced Information Retrieval at ECIR 2022, April 10, 2022, hybrid.
EMAIL: frederique.bordignon@enpc.fr
ORCID: 0000-0002-4918-9137

Sustainable Development Goals[1] defined by the United Nations. With this particular case study and the way it could be tackled, I will say a word about building up corpora and what I call "the hopeless quest for comprehensiveness" often experienced by the bibliometrician before even being able to start the analysis work. Then I will present the method I set up to match articles and SDGs.
In the end, through this practical example, the difficulties it presents as well as the hopes it may raise, I would like to draw attention to the "uncontrolled" use of tools and sources.

1. The United Nations Sustainable Development Goals

In 2015, the United Nations set 17 Sustainable Development Goals (SDGs) to be achieved in 2030: "SDGs are the blueprint to achieve a better and more sustainable future for all. They address the global challenges we face, including poverty, inequality, climate change, environmental degradation, peace and justice.". Research is of course a way to achieve those goals. Increasingly, public institutions and the private sector are seeking to know (and often to show) how they contribute or can contribute to achieving the UN objectives.
For an institution, it is a question of qualifying its scientific production in the light of the SDGs, and therefore of knowing itself better, positioning itself in relation to the challenges and possibly adjusting its policy. This is in fact the role that bibliometrics must play...

So the mapping of scholarly communication to SDGs has aroused great interest. The THE (Times Higher Education) international ranking agency has for example launched a ranking based on SDGs[2]. The bibliographic database Dimensions offers a feature (since 2020[3]) that allows the results of a query to be mapped to the SDGs, but my first trials were inconclusive with publications that were clearly climate change related not being flagged by Dimensions to be associated with the SDG "Climate Action". I think they have improved their functionality since then but I do not know exactly how. Very recently Clarivate mapped all publications from 1980 onwards to the SDGs in their tool InCites[4], which offers a visualization of the results of any query. And Elsevier added pre-generated search queries[5] in their advanced search form in Scopus.

But at the time when I needed to make this kind of projection, the few possibilities were not convincing and I had to innovate.

But before getting to the heart of the matter, I would like to dwell at some length on the stage prior to any bibliometric analysis, and from which this case study is no exception, which is gathering data, building the corpus. Before undertaking any kind of bibliometric study, regardless of the scope, one must begin by collecting bibliographic data. And for the bibliometrician, achieving comprehensiveness is a challenge.

2. The hopeless quest for comprehensiveness

In order to address the issue of the comprehensiveness of the data needed to build up the corpus on which the analyses are based, it is necessary to distinguish between research assessment, which is mainly covered by bibliometrics, and the research-on-research activity (meta-

[1] https://sdgs.un.org/goals
[2] https://www.timeshighereducation.com/rankings/impact/2021/overall
[3] https://www.dimensions.ai/blog/dimensions-includes-new-research-category-filters-for-sustainable-development-goals
[4] https://clarivate.com/blog/a-more-sustainable-future-for-all-introducing-the-un-sustainable-development-goals-in-incites
[5] https://blog.scopus.com/posts/sustainable-development-goals-sdgs-on-scopus

research) dedicated to the study of sciences, which is partially covered by scientometrics. The objectives and methods involved are in fact different.

For a bibliometric report designed for an institution's decision-makers, who will find intelligence to adjust their strategy, comprehensiveness is required. It is not possible to assess the output of an institution, and even less so of a researcher, if part of the production is not included in the corpus underlying the analysis. This would inevitably lead to bias. Comprehensiveness is therefore a key principle, a prerequisite for any analysis. Precision and recall are required. In contrast, in a scientometric study, more importance is attached to the representativeness of the data collected if the exhaustive collection of data potentially concerned by the object of study is not achievable. The process of delineating the corpus is very important here, and the description and justification of the choices made in assembling the data are essential in order to avoid introducing biases that would be detrimental to the future analyses.

In an institutional bibliometric report, the delineation of the corpus is straightforward and obvious, as it must cover all the output of the institution under consideration. Consequently, it is advisable to query as many sources as possible, in this case bibliographic databases (e.g., Scopus, Dimensions, Web of Science, The Lens, CrossRef) and to supplement this with the deposits of the authors themselves in institutional open archives (e.g., HAL in France) or even inhouse tools, whose metadata are often incomplete. This bibliographic data collation task is so long and tedious that it is difficult to reproduce it for all the institutions which one might want to compare to. This is probably an explanation for the success of international rankings which claim to provide an accurate benchmarking of institutions, but the truth is that they are far from carrying out this data collection properly, and in fact they generally rely on a single bibliographic source.

It is difficult to figure out what could quickly improve the situation, reduce this workload and increase the reliability of comparative studies. However, I have some high expectations for open repositories such as OpenCitations[6], The Initiative for Open Abstracts[7] and OpenAlex[8]. As their name suggests, they are non-commercial initiatives that rely on open data, but it remains to be seen how much coverage they can provide and whether they can be used as a single source to at least save time on data aggregation and harmonization.

Nevertheless, whatever the purpose of these studies, the files and their formats are eventually very similar. Another stage then begins, which consists of enriching this data to give it value and provide analyses that go beyond mere counting. The possibilities are undoubtedly infinite... but what about the consequences on the overall quality and the comprehensiveness of the initial corpus?

My point of view, or rather my experience, is that "the richer, the poorer": the more you try to enrich your data, the poorer you get. In a multidisciplinary environment, it is indeed impossible to retrieve the number of citations of *all* the publications in the corpus, nor their Open Access status, whether using Unpaywall[9], which is based on DOIs, or the DOAJ[10], which makes it possible to complete the data based on the journal status (Open Access or not). It is not possible to associate a disciplinary field to *each* article if a classification of journals by field is chosen.

[6] https://opencitations.net
[7] https://i4oa.org/#openabstracts
[8] https://openalex.org
[9] https://unpaywall.org
[10] https://doaj.org

It is not possible to retrieve the Altmetric score[11] (even if it is 0) for *all* publications, since, here too, a DOI (or an equivalent identifier) is required.

I did a back-of-the-envelope calculation for this keynote: I took a corpus of 5,478 articles, published within the 2016-2021 period by researchers of my institution (Ecole des Ponts) and at the end of various enrichment processes, if I want to provide an analysis on the publications for which the journal field(s), the Altmetric score, the Open Access status AND the number of citations can be retrieved, 40% of the initial corpus is lost.

The practitioner needs to navigate between the heterogeneity of databases and data, and the very availability of the data via these databases. And above all, the dependance on DOIs is very high. Overall, if we take the same corpus from my institution, the DOI is missing for less than 10% of the publications. But the situation is very different depending on the discipline: in Arts & Humanities and Social Sciences, there are still many journals there are still many journals not assigning DOIs to their articles. In some of the research units, up to 35% of DOIs are missing for articles.

So, at this point, if we go back to the mapping of articles to SDGs, we would be tempted to use the disciplinary fields of the journals and group them together to feed each SDG. But it is in fact impossible to establish such a correspondence. Each SDG is described very finely by the UN using specific targets, and even the journal classification schemes that include a relatively fine granularity level are not sufficient. It must be said that the 169 targets that the UN has presented to describe the SDGs are open to interpretation and that the scope of each SDG is very difficult to delineate since each SDG is intertwined with others. A recent paper has shown that these differences in interpretation and translation into queries can result in very different sets of publications [1].

It is therefore necessary to use textual analysis to assess SDG assignments at the paper level. The inventive bibliometrician has to move away from the metrics, scores and volume calculations associated with metadata and move into more subtlety, such as what text mining can provide.

3. The strive for nuance

The idea is obviously not new. Scientometrics has indeed benefited from the contribution of sociologists like Callon and Latour. And this movement led to the development of methods for mapping scientific topics as an alternative to citation analysis methods; Callon et al [3] stated: "A scientific text not only reveals the world-building strategy of its authors, but also the nature and force of the building blocks derived from the domain of science from which it draws and to which it contributes. The text thus provides access to the dynamics of science, to the shared worlds that constitute a means of mutual (and evolving) control.".

The search for occurrences and co-occurrences of words is often used in scientometrics; it allows a corpus to be explored and represented as clusters, for example, and to follow their evolution over time. Named Entity Recognition, including the possibility of retrieving locations from text, also facilitates the analysis of affiliation lines, which in France are the bibliometrician's nightmare both because the Higher Education system is complex and because

[11] https://www.altmetric.com

researchers are so creative in the way they sign articles. MeaningCloud[12], Netscity[13] and CorTexT[14] are very helpful to perform this mapping.
But what is most interesting is when a traditional bibliometric indicator and text analysis are combined to incorporate a new analysis into the existing framework, that is within the existing database or dynamic dashboard. It is important that the results enrich and inform existing traditional indicators and that they are not used or presented in a disconnected way.

For example, Scite[15] goes beyond calculating citations and infers their polarity in order to indicate whether the author supports or contradicts the works he/she cites, or merely mentions it to provide the background. Introducing more nuance in the study of citations would allow to identify controversies related to certain claims and to understand how science progresses. It can also improve the work dedicated to the detection of communities (with VOSviewer or CiteSpace for example) by nuancing the links between authors depending on whether the citation is critical or not. The analysis of the text, and more specifically the context of the citations, makes it possible to fine-tune the bibliometric indicator.

So it was with all these good examples in mind that I set out to develop a method to match articles with SDGs.

In the oral version of this keynote, I had the opportunity to present this work published in *Data in Brief,* and which therefore cannot be included *in extenso* in this document.
Nevertheless, here is a short presentation.

In a data paper [2], I present the method I developed to build a set of queries allowing the mapping of the United Nations SDGs with articles. This method, somehow improving the queries previously shared by Jayabalasingham [4], has the advantage of mitigating the polysemy of terms thanks to the combination of a bibliometric indicator (i.e.: disciplinary field(s) of journals in the ASJC classification scheme) and text retrieved from titles, abstracts and keywords describing articles.
It results in one Boolean query by SDG (except for the one related to international relations which can be considered as equivalent to a co-publication analysis highlighting the links between authors).

I tested this method with 81 researchers affiliated to Ecole des Ponts who were asked to associate an article with one or more SDGs. I then compared their answers with the SDGs retrieved from the queries. The results were good enough for me to apply this method in my institution. In the meantime, Elsevier has integrated "off-the-shelf" queries into Scopus advanced search, but unfortunately without using the ASJC classification, even though they are in the best position to take advantage of it.

Disclaimer: The articles mapped to SDGs through this process are not evidence of the commitment of authors and their institutions to actions towards the targets established by the UN. They should be carefully considered as describing research related to the various issues to be addressed according to the UN, but not as "deliberately" providing a way to do so.

[12] https://www.meaningcloud.com
[13] https://www.irit.fr/netscity
[14] https://www.cortext.net
[15] https://scite.ai

Improving these queries and in particular finding a way to extend them to other document types is a challenge for the BIR community.

4. Conclusion and discussion with the BIR community

In this keynote, I have shown the pitfalls and promises of Bibliometric-enhanced Information Retrieval taking the concrete example of the bibliometrician facing the difficulty of matching SDGs and articles from an institutional corpus.

Given the diversity of sources and formats, our dependence on DOIs, and the fluctuating coverage of the tools we use to enrich our data, we are insidiously and inescapably degrading our data. This is the pitfall that needs to be avoided by the bibliometrician. Furthermore, I provided examples where bibliometric indicators and text analysis combine well to provide innovative studies. And I pointed out a very large number of tools, both for analysis and for data retrieval. They are indeed a true reflection of field practices. However, if we take a step back, it is worth considering the consequences of this assemblage of disparate sources and tools.
This never-ending tinkering, although highly creative, might lead to the impossibility to compare the results obtained, whether on the scale of an individual, an institution, a topic, a country, a disciplinary field… Indeed, in the absence of a norm, it is likely that the practitioner is forced to pick and choose among the technical and methodological solutions that abound, mainly according to his or her skills, and the available of data.
I propose to the BIR community to engage in a discussion on this concern.

5. Acknowledgements

I would like to thank the BIR workshop organizers for their invitation, and Guillaume Cabanac for his insightful comments.

6. References

[1] C S Armitage, M Lorenz, and S Mikki. Mapping scholarly publications related to the Sustainable Development Goals: Do independent bibliometric approaches get the same results? *Quant. Sci. Stud.* 1, 3 (2020), 1092–1108 (2020). doi:10.1162/qss_a_00071
[2] F Bordignon. 2021. Dataset of search queries to map scientific publications to the UN sustainable development goals. *Data Brief* 34, 106731–106731 (2021). doi:10.1016/j.dib.2021.106731
[3] M Callon, J Law, and A Rip. How to study the Force of Science. In *Mapping the dynamics of science and technology - Sociology of Science in the Real World*. The Macmillan Pressltd, London, 1986.
[4] B Jayabalasingham, R Boverhof, K Agnew, and L Klein. 2019. Identifying research supporting the United Nations Sustainable Development Goals. *Mendeley Data* (2019). doi:10.17632/87TXKW7KHS.1