RESEARCH ARTICLE

Evaluating approaches to identifying research supporting the United Nations Sustainable Development Goals

Yury Kashnitsky[1], Guillaume Roberge[2], Jingwen Mu[3], Kevin Kang[3], Weiwei Wang[3], Maurice Vanderfeesten[4], Maxime Rivest[2,5], Savvas Chamezopoulos[1], Robert Jaworek[6], Maéva Vignes[7], Bamini Jayabalasingham[1,8], Finne Boonen[1], Chris James[1], Marius Doornenbal[1], and Isabelle Labrosse[2]

[1]Elsevier BV, Amsterdam, Netherlands
[2]Elsevier BV, Montreal, Canada
[3]The University of Auckland Faculty of Science, Auckland, New Zealand
[4]Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[5]McGill University, Montreal, Canada
[6]Palacky University Olomouc, Olomouc, Czech Republic
[7]University of Southern Denmark, Odense, Denmark
[8]Elsevier BV, New York, NY, USA

Keywords: benchmarking, bibliometrics, machine learning, scientometrics, sustainability, Sustainable Development Goals

ABSTRACT

The United Nations (UN) Sustainable Development Goals (SDGs) challenge the global community to build a world where no one is left behind. Recognizing that research plays a fundamental part in supporting these goals, attempts have been made to classify research publications according to their relevance in supporting each of the UN's SDGs. In this paper, we outline the methodology that we followed when mapping research articles to SDGs and which is adopted by Times Higher Education in its Social Impact rankings. We compare our solution with other existing queries and models mapping research papers to SDGs. We also discuss various aspects in which the methodology can be improved and generalized to other types of content apart from research articles. The results presented in this paper are the outcome of the SDG Research Mapping Initiative, which was established as a partnership between the University of Southern Denmark, the Aurora European Universities Alliance (represented by Vrije Universiteit Amsterdam), the University of Auckland, and Elsevier to bring together broad expertise and share best practices on identifying research contributions to UN's Sustainable Development Goals.

1. INTRODUCTION

Numerous approaches to mapping research to the United Nations (UN) Sustainable Development Goals (SDGs)[1] have been documented (Armitage, Lorenz, & Mikki, 2020b; Bordignon, 2021; Confraria, Ciarli, & Noyons, 2022; Jayabalasingham, Boverhof et al., 2019; LaFleur, 2019). These approaches vary with regard to the framework used to define inclusion and

[1] https://metadata.un.org/sdg/.

an open access journal

Check for updates

Citation: Kashnitsky, Y., Roberge, G., Mu, J., Kang, K., Wang, W., Vanderfeesten, M., Rivest, M., Chamezopoulos, S., Jaworek, R., Vignes, M., Jayabalasingham, B., Boonen, F., James, C., Doornenbal, M., & Labrosse, I. (2024). Evaluating approaches to identifying research supporting the United Nations Sustainable Development Goals. *Quantitative Science Studies*, *5*(2), 408–425. https://doi.org/10.1162/qss_a_00304

DOI: https://doi.org/10.1162/qss_a_00304

Peer Review: https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss_a_00304

Supporting Information: https://doi.org/10.1162/qss_a_00304

Received: 1 May 2023
Accepted: 28 January 2024

Corresponding Author:
Yury Kashnitsky
y.kashnitskiy@elsevier.com

Handling Editor:
Vincent Larivière

The MIT Press

exclusion criteria, the methodology employed to retrieve publications, and the publication database used. For example, the approach to defining inclusion and exclusion criteria may be set conservatively to limit publications to those documenting actions made to achieve the SDG targets, or conversely, may be set using a more liberal approach, thereby including any papers that increase knowledge on the overall topic. With regard to the methodology employed to retrieve publications, publication sets for a specific SDG can use a Boolean approach only or be complemented by machine learning algorithms.

The source of publications that the methodology is applied to can also introduce variability, given the availability of many data sources, ranging from open access, subscription-based, or a mixture of both.

To date, there is no broadly agreed-upon methodology for mapping research to the SDGs and existing methods produce quite different results (Armitage et al., 2020b). A common approach to identifying research related to a topic is to use Boolean search expressions. The Boolean method involves the use of keywords, either alone or in combination, using conditional functions and applied to specified text sections (title, abstracts, keywords, etc.) of scientific publications, and results in the exclusive retrieval of articles within which the defined search expressions were found. Armitage et al. (2020b) applied the Boolean method, taking an approach to limit their SDG publication sets to publications with a direct contribution to targets and/or indicators, with efforts made to reduce the impact of issues raised on the Boolean technique, resulting in a more restrictive publication set. Bordignon's strategy (Bordignon, 2021) aimed at reducing the polysemy of terms by limiting keywords from Elsevier 2020 queries (Jayabalasingham et al., 2019) to relevant subject areas using the All-Science Journal Classification (ASJC). A text-mining tool (CorTexT) was then used to enrich those selected publications. The Aurora European Universities Alliance (Schmidt & Vanderfeesten, 2021) developed and released their 169 target-level SDG queries (Vanderfeesten, Otten, & Spielberg, 2020) also using keyword combinations and Boolean- and proximity operators. The University of Auckland (Jingwen & Weiwei, 2022) developed queries informed by the researchers within their network, resulting in a localized version that takes into account more papers that are specific to Australian and New Zealand research topics. Confraria et al. (2021) employed a two-step approach, involving building SDG-specific terms obtained from many sources (policy reports, publications, forums, etc.), applying a selection process to the terms, and then using the terms to identify citation-based communities of publications. However, as described by Armitage et al. (2020b), such a keyword-based approach involves challenges related to the interpretation of the themes and concepts of the SDGs, decisions around which publications to designate as a "contribution" to the chosen interpretation of the SDG, and the translation of concepts into a search query that will accurately identify publications.

An alternative or complementary approach to query-based methods involves using machine learning to map research articles to SDGs: either in a supervised manner (i.e., performing classification) or an unsupervised manner (i.e., performing clustering). Supervised methods typically resort to the same SDG queries to obtain a labeled data set to train the model (South African SDG Hub, 2023; Zhang, Vignes et al., 2020). Clustering is typically done with paper text representations or citation graphs where the resulting clusters are later mapped to SDGs either directly or via intermediate clusters (e.g., "topics"; Nakamura, Pendlebury et al., 2019; Wastl, Porter et al., 2020). Refer to Pukelis, Puig et al. (2020) for an overview of some more methods of classifying documents into SDGs. However, they all face the same challenges noted above, and machine learning further introduces the problem of interpretability of the model predictions or the clusters attained.

Since 2018, Elsevier has endeavored to map research to the SDGs, releasing publicly available queries to facilitate transparency and reproducibility (Jayabalasingham et al., 2019). Herein, we describe the approach taken to improve former attempts to map research to the SDGs, taking feedback into account, resulting in the creation of a more comprehensive query set with subqueries addressing targets and indicators and the application of a machine learning model to increase recall. This methodology ("Elsevier 2021 SDG mapping": Rivest, Kashnitsky et al. (2021)) captures on average twice as many articles as the 2020 version, while keeping precision above 80%. Times Higher Education (THE) is using Elsevier SDG mapping as part of its Social Impact rankings (Ross, 2022). "Elsevier 2023 SDG mapping" (Bedard-Vallee, James, & Roberge, 2023) is the most up-to-date simplified version of the queries & ML model, differing from the 2021 version in COVID-related enhancement to SDG 3 queries and queries designed for SDG 17: "Partnerships for the goals."

To evaluate the approach, the output generated using the developed methodology was compared to the results generated by Aurora European Universities Alliance (Vanderfeesten & Jaworek, 2022), the University of Auckland (Jingwen & Weiwei, 2022), the University of Bergen (Armitage et al., 2020b), SIRIS Academic (Duran-Silva, Fuster et al., 2019), Bordignon queries (Bordignon, 2021), and the ML classifier by the South-African SDG Hub (South African SDG Hub, 2023).

We have not seen much research aimed at doing similar benchmarking of different SDG mapping approaches with hand-labeled data sets. Wulff, Meier, and Mata (2023) is the closest investigation to ours: Apart from benchmarking, the authors explore the extent to which SDG queries produce false positives by marking non-SDG-related content with SDG labels. They also investigate the bias in SDG labeling systems defined as the normalized difference in the number of predicted and observed (i.e., put by human experts) SDG labels.

The novel contributions of this paper can be summarized as follows:

- We solve the problem of recall assessment for keyword queries mapping research articles to Sustainable Development Goals, while other approaches typically focus on precision.
- We are among the first to quantitively evaluate existing sets of such keyword queries against several validation data sets.

2. METHODOLOGY

2.1. Developing SDG Queries

The SDGs are goals to achieve rather than research topics, each SDG encompassing many targets. Using Boolean search expressions to build SDG-specific publication sets presents many challenges. Elsevier implemented a bottom-up approach to the construction of each SDG-relevant publication set, whereby several subqueries were first constructed for each SDG target, and then aggregated at the SDG level.

2.1.1. Building a query for each target within an SDG

The criteria for delineating the publication sets relevant to each SDG were designed by a team consisting of a minimum of four analysts and were based on an extensive literature review done by the team to gain an understanding of the SDG. As a first step, the SDG was further subdivided into themes to facilitate the creation of specific criteria linked to specific SDG targets. The criteria defined for each theme aimed to specify topics of focus as well as any requirements for "action terms" in association with the topics (Armitage et al., 2020b). For

example, for the topic of "poverty" the action term "alleviate," or other action terms holding similar meaning might be deemed a requirement. To ensure homogeneity in the approach, the criteria developed by the team of analysts were submitted to a review committee consisting of both those on the SDG team and those external to the team. The review committee was responsible for reviewing the criteria, recommending changes, and final approval of the criteria. Table 1 presents the criteria for SDG1 overall (SDG1-Main) and subcategories related to SDG1. These criteria were defined for each SDG and theme related to the SDG.

Following the establishment of criteria defining the research areas of focus relevant for each SDG (overall and per SDG-Theme), these criteria were used to guide the development of queries to retrieve publication sets. Where possible, the analyst responsible for query development was selected due to subject matter expertise in the field. Otherwise, the process was informed by a literature review. An iterative approach was taken to assess the precision with which individual keywords and sets of keywords identified publications that met the criteria. Keywords from the Elsevier 2020 (Jayabalasingham et al., 2019) and Aurora European Universities Alliance (Schmidt & Vanderfeesten, 2021) queries were assessed first. Additional keywords were identified using term-frequency and inverse-document frequency (TF-IDF) analyses of text from titles, abstracts, and author keywords from publications meeting the criteria. Additional efforts were taken to identify publications that may have been excluded based on the developed query. Specifically, the query results were analyzed to identify specialized journals that would be expected to include a high percentage of publications that fit the

Table 1. An example of SDG 1 subtopics and associated SDG targets

Subset code	Criteria	Associated target
SDG1-Main	Research focused on poverty and research as defined for any SDG1-subset below. "Action term" specified: The action term, "alleviate" was applied to make the topic term "poverty" more specific.	Target 1.1: Eradicate extreme poverty Target 1.2: Reduce poverty by half All Targets associated with SDG1-Subsets
SDG1-Theme1	Research focused on social programs, including all articles discussing social security systems related to health, finance, and work. No "action terms" were required for the inclusion of the topics above.	Target 1.3: Implement nationally appropriate social protection systems
SDG1-Theme2	Research focused on microfinance, access to the property, inheritance, natural resources, and new technologies as they relate to facilitating access, equality, and human rights. "Action term" specified: the action term "access to" was applied.	Target 1.4: Equal rights to economic resources and basic services
SDG1-Theme3	Research focused on resilience, exposure, and vulnerability to disasters (financial, climate-related, social ...), particularly on understanding poor and vulnerable people and communities. No "action terms" were required for the inclusion of the topics above.	Target 1.5: Build the resilience of the poor
SDG1-Theme4	Research focused on financial aid, policies, government support (such as food banks, and support distribution strategies), and strategies to eradicate poverty. No "action terms" were required for the inclusion of the topics above.	Target 1A: Ensure significant mobilization of resources from a variety of sources Target 1B: Create sound policy frameworks

criteria, and the citation network of the publications retrieved using the query was assessed to identify publications within the citation network of the results (i.e., publications citing or cited by the publications retrieved by the query) that were not retrieved by the query. Publications from these specialized journals or the citation network that were not being retrieved by the query were assessed to identify additional keywords to include in the query to increase recall. Relevant exclusions were built into the queries to increase precision and could result in the exclusion of specific terms using Boolean operators or the exclusion of fields of science deemed to be outside the scope of the criteria.

To facilitate the continuous evaluation of the query, publications were manually reviewed to assess their fit against the criteria shown above (see Table 1 for SDG1). An evaluation of a minimum of 100 random publications by two independent analysts was done to support the calculation of precision metrics for each query and a minimum precision threshold of 90% was required for a query to be considered acceptable. The recall was assessed against independent publication sets developed by an analyst consisting of publications from specialized journals identified to fit the criteria. As most specialized journals do not exclusively focus their content on a single SDG, a minimum recall of 60% was required for a query to be considered acceptable. In cases where no single journal was specific enough for all publications within that journal to fit the criteria set for an SDG or SDG-Theme, a publication set was constructed by manually selecting publications from a journal with high relevance to the SDG (or SDG-Theme), and recall was assessed against this set.

2.1.2. Precision assessment

As described earlier, queries were composed gradually, starting from the seed queries developed at first by analysts. These queries were developed by concatenating queries together with Boolean OR' expressions after evaluating the keywords suggested by the TF-IDF analysis on the seed data set. Before adding a new search expression to the global SDG data set, analysts were encouraged to sample at least 10 documents to ensure that high precision was maintained throughout most queries and not simply for the global SDG data set. This is quite important, as otherwise some keywords bringing a small number of new publications, but covering mostly content not relevant to the SDG, could be included in the data set, and while their impact on global precision would be relatively small, it would still mean that analysts would be forcing bad content with such terms. The sampling was performed directly in the exploration window which could be used to quickly draw random samples of publications containing the selected keywords. This enabled analysts to vet new keywords quickly, which was necessary given the complexity of the queries needed to delineate the SDGs properly. As a target, a 90% precision level was required to commit the tentative search expression; otherwise, a lower level would lead to diminished precision for the global data set at the end of the iterative process. This was especially critical for keywords adding a lot of new documents to the global data set, as lower precision for these would more greatly influence the global precision.

Although precision was assessed throughout the whole process, a more formal precision estimate was performed at the end of the whole process to provide a final assessment, which would guide analysts as to whether they could stop their work or if an additional effort was needed to remove content that was deemed too broad and resulted in lowered precision. A large sample of 100 publications was pulled from the global data set and analysts performed a manual inspection of these, the tool enabling us to tag publications as good, bad, and in-between for cases where the analyst was unsure whether documents should be included or not. This feature presented the advantage that final precision assessments were stored in the

tool and could be consulted at any time in the future. This was especially helpful when additional validation steps were performed by the QA analyst, which was able to validate the precision assessment by assessing the same sample. If final precision was in the 90–95% range or above, precision was deemed sufficient.

As a final step, a final QA was performed by an expert bibliometrician with more than a decade of experience in the field and in building data sets. Each query was analyzed by this expert, and tested again for its precision, reusing the samples pulled from each analyst, but often pulling new samples as well to further solidify confidence. This additional layer of validation helped cement the process, ensuring a unified view over all SDGs, in a similar way to what was accomplished when defining definitions as groups at the beginning of the process. The QA round led to multiple modifications, removals, and additions across most SDGs, often resulting in relatively minor changes in publication counts, but further increasing the robustness of the alignment between the definitions and the final content retrieved by the queries.

2.1.3. Recall assessment

To determine the recall of the queries developed by analysts, a selection of specialized journals was identified for each SDG to serve as a stand-in for a gold standard, representing the subjects at hand. This pragmatic "proxy" for recall measurement was developed in the absence of a true gold standard for testing the recall of the queries. The absence of a gold standard is unsurprising; should such a gold standard exist, it would imply that perfectly delineated document sets for SDGs would already exist, thus rendering the current exercise irrelevant. For each SDG, sets of highly relevant journals were identified using a combination of keyword searches in journal names and percentages of journal content covered by the keyword queries. This dual approach ensured that no relevant journals would be missed simply because their name was not declarative enough to be captured. After these journals were identified, analysts aimed to maximize recall across each of these journals, while maintaining high precision. Recall levels of 60–70% were set as the original minimal level for the current exercise based on two decades of expertise building such data sets. Increasing recall for some categories without comprising precision is sometimes easy in subjects relying on highly declarative vocabulary, while it can become quite tricky in others, especially those mixing multiple dimensions as their core concepts. In the case of the targets of the SDGs, this notion is especially relevant, as SDGs often mix basic research with economic and social concepts.

During the process, recall against the selected gold standard of journals was tested frequently to determine if more investigation was needed to add new keywords to the queries. Analysts performed recurring analyses of the content of these journals not captured by the queries to detect any research subject not covered. TF-IDF analyses on these documents not retrieved were performed to obtain lists of suggested terms for inclusion to further increase recall. At the end of the process, if recall remained low, corrected recalls were computed by sampling amongst the publications not retrieved with the keyword queries, estimating which part was truly relevant to the subject at hand. Indeed, specialized journals, while usually having targeted scopes, are not always fully relevant to the topic at stake. By sampling about 50 publications, analysts were able to compute corrected recall scores by estimating the fraction of the content not covered that was indeed relevant to subjects.

As a final step, a final QA was again performed by the expert bibliometrician. Each query was analyzed by this expert and tested for recall, investigating whether areas of each target

might have been missed or left out by the analyst. The QA round led to multiple modifications, removals, and additions across most SDGs, often resulting in relatively minor changes in publication counts, but further increasing the robustness of the alignment between the definitions and the final content retrieved by the queries.

Below we refer to the mentioned recall evaluation data set as to the Elsevier recall data set.

2.2. Machine Learning Applied to SDG Classification

On top of the mapping produced by the queries described above, additional articles are mapped to the SDGs by a machine learning model.

In a nutshell, the model is a logistic regression trained with TF-IDF representations of titles, keywords, abstracts, and two more optional text fields—main terms extracted from the full text and subject areas of the journal that published the paper. Thus, the model learns similar key phrases for each SDG and helps to improve the recall of the queries. To keep precision high, we keep only those papers that are classified by the model with 95% or higher predicted probability for some SDG.

In the Elsevier 2021 SDG mapping release (Rivest et al., 2021), the Elsevier team specifies the input data for the model, the targets with which it is trained, the technical details of the model itself, and model performance. Also, to ease the interpretation of the model classification outcomes, we share the SDG-specific key phrases learned by the model, as well as sample articles classified by the model. Refer to the mentioned documentation for more details on the machine learning component of our approach.

2.3. Combining the Queries and the Model

The end-to-end approach to mapping scholarly records to SDGs is two-staged:

- First, the keyword SDG queries are run (orange in Figure 1).
- Then, the ML model adds about 3.5% of papers (blue in Figure 1) on top of what is classified by the keyword queries. We only keep the most confident model predictions by thresholding predicted scores at 0.95.

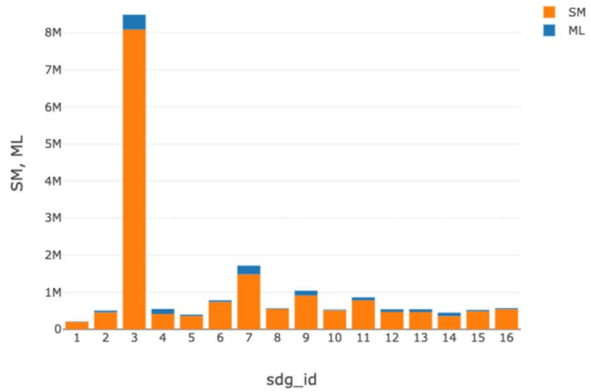

Figure 1. Distribution of the number of papers mapped by the queries (SM, orange) and by the model (ML, blue), by SDG (ignoring SDG 17).

Note that the approach is limited to the Scopus database, as the queries are written in Scopus search syntax.

3. RESULTS

3.1. Comparison Between the SDG Queries

Below we describe the SDG queries and validation data sets that we used for the comparison in terms of precision, recall, and F1 scores.

3.1.1. Query models

Table 2 describes the different classification methods that we compared. These can be either keyword queries ("Elsevier queries 2020," "Aurora queries v5," "Auckland queries v2," and "Bergen SDG queries") or machine learning models ("Aurora ML v0.2") or both ("Elsevier queries + ML 2021," "Elsevier queries + ML 2022").

3.1.2. Validations sets: Collection method, sizes, and quality

Table 3 provides details on the validation data sets used in the comparison. It also mentions the associated limitations and biases. It is important to mention that there is no single best validation data set to evaluate the output of SDG classification.

3.1.3. Performance; query models measured against validation sets

Table 4 provides the evaluation results for the SDG classification methods outlined in Table 2 and evaluation data sets described in Table 3. Each cell shows two values: microaverage F1-score and macroaverage F1-score (the microaverage F1-score aggregates performance metrics across all classes by treating each instance equally, while the macroaverage F1-score computes the F1-score for each class independently and then takes the average, giving equal weight to all classes regardless of their sizes), in percent (%). Both precision and recall were calculated with respect to the validation sets (i.e., all predictions beyond the validation sets were ignored):

- Precision is calculated as the number of correctly predicted SDG IDs divided by the number of Scopus IDs tagged with the same SDG ID in the given validation set.
- Recall is calculated as the proportion of correctly predicted SDG IDs within the given validation set.

To compare with Bergen queries, Table 5 provides similar metrics only considering a subset of 10 SDGs, namely, SDG 1 (No poverty), SDG 2 (Zero hunger), SDG 3 (Good health and well-being), SDG 4 (Quality education), SDG 7 (Affordable and clean energy), SDG 11 (Sustainable cities and communities), SDG 12 (Responsible consumption and production), SDG 13 (Climate action), and SDG 14 (Life below water), and SDG 15 (Life on land).

The same comparisons for precision and recall are found in the Supplementary material (see Tables S1–S4).

Note that microaveraging favors well-represented, frequent classes (like SDG 3 in our case) while high macroaveraged scores mean that the method works fairly well across all SDGs because bad results for a single SDG affect macroaveraged metrics much more than microaveraged ones. By attending to both micro- and macroaveraged F1-scores we try to assess both aspects: how good the method is at classifying papers into frequent or rare classes.

Table 2. SDG classification methods (both keyword queries and ML models) used in the evaluation (Armitage, Lorenz, & Mikki, 2020a)

Classification method	Description	Web location
Auckland queries v2 (Auckland_v2)	To gain a better understanding of our research contribution, the University of Auckland SDG Keywords Dictionary Project seeks to build on the processes developed by the United Nations and THE in order to create an expanded list of keywords that can be used to identify SDG-relevant research.	https://www.sdgmapping.auckland.ac.nz/
Aurora ML v0.2 (Aurora_ml)	"AI for mapping multi-lingual academic papers to the United Nations' Sustainable Development Goals (SDGs)" (Vanderfeesten & Jaworek, 2022).	https://doi.org/10.5281/zenodo.5603019
Aurora queries v5 (Aurora_v5)	"Mapping Research Output to the Sustainable Development Goals (SDGs)" (Vanderfeesten et al., 2020).	https://zenodo.org/records/4883250
Bergen SDG queries (Bergen_2023_baa and Bergen_2023_bta)	The Bergen approach created queries for Web of Science to retrieve SDG-related publications for a limited number of SDGs. The queries have been translated for Scopus and a sample of the results has been taken as positive examples. These have been supplemented by other publications which did not appear in the queries as negative examples. Two data sets were created, one based on the Action Approach queries and one based on the Topic Approach queries—referred to as Bergen BAA and Bergen BTA respectively (Armitage et al., 2020b).	https://zenodo.org/records/7711561
Elsevier queries 2020 (Els_2020)	"Identifying research supporting the United Nations Sustainable Development Goals" (Jayabalasingham et al., 2019).	https://elsevier.digitalcommonsdata.com/datasets/87txkw7khs/1
Elsevier queries + ML 2021 (Els_2021)	"Improving the Scopus and Aurora queries to identify research that supports the United Nations Sustainable Development Goals (SDGs) 2021" (Rivest et al., 2021).	https://elsevier.digitalcommonsdata.com/datasets/9sxdykm8s4/4
Elsevier queries + ML 2022 (Els_2022)	A simplified version of "Elsevier queries + ML 2021" with Covid-related addendum to SDG 3 (Roberge, Kashnitsky, & James, 2022).	https://elsevier.digitalcommonsdata.com/datasets/6bjy52jkm9/1
Elsevier queries + ML 2023 (Els_2023)	For 2023, the SDGs use the exact same search query and ML algorithm as the Elsevier 2022 SDG mappings, with only minor modifications to five SDGs, namely SDG 1, 4, 5, 7, and 14. In these cases, the queries were shortened by removing exclusion lists based on journal identifiers. These exclusion lists often contained thousands of items to filter out content in journals that were not core to the SDGs (Bedard-Vallee et al., 2023).	https://elsevier.digitalcommonsdata.com/datasets/y2zyy9vwzy/1
South African SDG hub (South_africa)	A machine learning model mapping text to SDGs.	https://sasdghub.up.ac.za/home/
SIRIS queries (SIRIS)	The SIRIS queries were developed by extracting key terms from the UN official list of goals, targets, and indicators as well as from relevant literature around SDGs. The query system has subsequently been expanded with a pre-trained word2vec model and an algorithm that selects related words from Wikipedia. There are multiple queries per SDG (Duran-Silva et al., 2019).	https://zenodo.org/records/4118028
Bordignon SDG queries (Bordignon)	These queries aimed at reducing the polysemy of terms by limiting keywords from Elsevier 2020 queries (Jayabalasingham et al., 2019) to relevant subject areas using the All-Science Journal Classification (ASJC) (Bordignon, 2021).	https://data.mendeley.com/datasets/xrx7ddbbb4/1

Table 3. SDG validation data sets used in the evaluation

Validation set	Description and method of collection	Web location	Size	Remarks on quality
Elsevier recall dataset (Elsevier_recall)	See Section 2.1.1	Shared via ICSR Lab; see Section 4.3	465k	The data set is noisy in the sense that not all papers from an SDG-specific journal are relevant to the same SDGs. Hence, we do not aim for 100% recall with respect to this data set
Aurora Survey (Aurora1)	"Survey data of 'Mapping Research output to the SDGs' by Aurora European Universities Alliance (AUR)": 244 senior researchers from different universities in Europe and the United States filled in a survey. They were only allowed to enter the survey if they were familiar with the SDG they had selected to evaluate. The first question was to provide a list of research papers they believe are relevant to that selected SDG. The second question was to handpick, from a given set of 100 randomly drawn papers in the Aurora query result set, the papers that they believe (based on reading the title, abstract, journal name, and authors) belong to the selected SDG. The suggested papers and the selected papers are included in the validation set.	https://zenodo.org/records/3813230#.YyG93uxBxYw	6,741	Bias: the researchers are located at western European universities.
Aurora Suggested Papers (Aurora2)	The papers suggested by researchers; see "Survey data of 'Mapping Research output to the SDGs'."	https://zenodo.org/records/3813230#.YyG93uxBxYw	3,964	The researchers involved in the survey identified themselves as having expertise in a specific SDG. They might also have the incentive to cite their own research
Elsevier multilabel SDG data set (Els_multilabel)	The data set consists of 6,000 papers annotated by three experts each. These papers come from five data sources to span as diverse as possible set of SDG-related papers. Thirty per cent of the papers are not mapped to any of SDGs.	Shared via ICSR Lab; see Section 4.3	6,000	Annotators are not as versed in SDGs as the analysts who developed Elsevier queries and the Elsevier recall data set
Chilean multilabel (Chile)	The data set is provided by Pontificia Universidad Católica (PUC) based in Chile and consists of about 1,200 papers self-assessed by PUC researchers and labeled with 0, 1, or 2 SDGs (Rodríguez, Delpiano et al., 2021).	https://repositorio.uc.cl/handle/11534/61951	1,200	Biases: self-assessment, only Chilean researchers
OSDG Community Dataset (OSDG)	A public data set of thousands of text excerpts, which were validated by approximately 1,000 OSDG Community Platform (OSDG-CP) citizen scientists from over 110 countries, with respect to the Sustainable Development Goals.	https://zenodo.org/records/6831287#.YyMF5OxBxYy	32,431	Crowd-sourced data set; the annotators are not versed in SDGs. In our benchmarks, we only kept the records with a difference between positive and negative votes greater than or equal to 2, thus leaving only 26,217 records.

Table 4. F1 scores with micro/macro averaging (percentages) for 10 classification methods and five validation data sets. Bold figures indicate the best result in the column; asterisks indicate multiple "winners" depending on micro- or macroaveraging

	Data set				
Method	**Aurora1**	**Aurora2**	**Els_multilabel**	**Chile**	**OSDG**
Auckland_v2	49/40	46/33	69/62	60/37	47/40
Aurora_ml	53/44*	39/32	64/57	55/38	**53/46**
Aurora_v5	55/42*	15/18	37/38	12/14	26/20
Els_2020	47/35	46/28	63/47	55/25	33/27
Els_2021	46/39	38/32	**73/67**	46/34	41/35
Els_2022	46/39	38/32	**73/67**	46/34	41/35
Els_2023	45/38	37/30	72/66	46/31	42/36
South_africa	51/40	45/35*	72/60	**65/41**	N/A
SIRIS	36/33	29/25	49/45	37/30	37/37
Bordignon	45/34	50/30*	60/48	61/32	N/A

The code reproducing the experiments presented in this subsection is found on GitHub[2]. Refer to the Data availability section for instructions on obtaining the data should you wish to reproduce the presented experiments.

We conclude that there is no single best approach performing well across all validation data sets: Some approaches are on average better at precision (e.g., Elsevier 2020 and South African SDG ML model; see Tables S1 and S2), others shine at recall (e.g., Auckland queries Q8 and Aurora ML model; see Tables S3 and S4). This finding supports the general criticism that SDG classification faces: Different mapping methods typically kick off with the same keywords but then result in poorly overlapping mappings (Armitage et al., 2020b; Purnell, 2022). Apart from these "qualitative" problems with SDG mappings we now establish the "quantitative" problem: When evaluated against several hand-labeled SDG data sets, different approaches fail to select a clear winner.

We notice a clear "overfitting" phenomenon: Elsevier queries + ML 2022 are best when validated against Elsevier's multilabeled data set while the Aurora queries v.5 and Aurora ML models achieve the highest F1-scores against the Aurora survey data set. A probable explanation is that the data sets were crafted for a specific definition/operationalization of SDGs and these definitions are undoubtedly different from one project to another.

It is important to conclude that there is no single "golden" SDG validation data set; each one considered in our experiments comes with its own shortcomings (see Table 3, remarks on quality), and each data set used in query development reflects some certain interpretation of SDGs by the query developers. Similarly to how Armitage et al. (2020b) concluded that there is a poor overlap in publications found by different sets of queries, we conclude that there is no clear winner among SDG classification methods when those are validated with available human-annotated SDG data sets.

[2] https://github.com/Yorko/sdg_mapping_queries_n_ml_benchmarks.

Table 5. F1 scores with micro/macro averaging (percentages) for 12 classification methods and five validation data sets. Here the validation is performed only against a subset of SDGs for which we have Bergen 2023 queries: 1, 2, 3, 4, 5, 11, 12, 13, 14, and 15

	Data set				
Method	**Aurora1**	**Aurora2**	**Els_multilabel**	**Chile**	**OSDG**
Auckland_v2	59/47	60/42*	78/70	69/46	59/57
Aurora_ml	61/48	47/37	72/63	66/48	**65/62**
Aurora_v5	**64/50**	17/23	40/46	13/18	29/27
Bergen_2023_baa	15/11	15/11	15/13	14/12	N/A
Bergen_2023_bta	17/16	17/15	22/25	16/14	N/A
Els_2020	54/39	61/37	71/52	63/33	39/37
Els_2021	55/45	46/38	**80/74**	53/42	47/44
Els_2022	55/45	46/38	**80/74**	53/42	48/45
Els_2023	54/43	46/37	80/73	52/38	48/46
South_africa	60/50	58/45	80/72	**74/56**	N/A
SIRIS	48/40	42/33	62/54	48/39	51/49
Bordignon	53/34	67/40*	68/52	72/40	N/A

3.2. Tracking the Progress of Elsevier Queries

The progress with SDG queries development at Elsevier was tracked both in terms of recall, as described in Section 2.1.3 and in terms of precision/recall/F1 when validated with the independently labelled Elsevier multilabel data set. Table 6 shows recall scores for different Elsevier queries as measured against the Elsevier recall data set described in detail in Section 2. Table 7 shows precision, recall, and F1 scores for different Elsevier queries as measured against the Elsevier multilabel SDG data set described in Table 3. Note that due to the specifics of the SDG query creation methodology, it makes sense to report only recall for the first data set. The reason is that it is labeled in a noisy way (the assumption that all papers from an SDG-specific journal contribute to the same Goal is far from perfect); thus, looking at precision (and hence F1) is not meaningful. However, reporting recall makes perfect sense—it shows how many SDG-related papers from this large data set the queries can detect.

Table 6. Elsevier queries validated against the Elsevier recall data set (see Section 2). Micro- and macroaveraged values for recall are reported

Method	**Elsevier recall data set, recall**
Els_2020	54/38
Els_2021	78/72
Els_2022	78/72
Els_2023	73/68

Table 7. Elsevier queries validated against the Elsevier multilabel SDG data set (see Table 3). P stands for precision, R for recall, and F1 for F1 score. Micro- and macroaveraged values are reported

	Elsevier multilabel data set		
Method	**P**	**R**	**F1**
Els_2020	72/62	57/42	63/45
Els_2021	69/63	78/75	73/63
Els_2022	69/63	78/75	73/63
Els_2023	68/62	76/73	72/62

From Tables 6 and 7, we see that all of the Elsevier 2021–2023 queries perform about the same in terms of metrics and provide a considerable improvement in recall (and hence F1) over the earlier 2020 version of the queries.

The metrics are close for the 2021–2023 versions of the queries because the 2022 and 2023 updates were not as considerable as the one in 2021. Namely, the 2022 version (Roberge et al., 2022) introduced only COVID-related changes to SDG 3. The 2023 version of the queries (Bedard-Vallee et al., 2023) introduced changes to SDGs 1, 4, 5, and 14, removing long lists of journal identifiers and replacing them with keywords.

4. DISCUSSION AND PERSPECTIVES

In previous sections, we described the methodology and evaluation results. Below, we outline possible improvements to the SDG mapping approach, including localization of the SDG queries, query generalization to non-English languages and extending the approach to nonarticle content.

4.1. Localization

Research activities do not stand alone; they are an integral part of the geographical place they were initiated and the communities they serve. An attempt to measure SDG-related research activities can be improved by infusing the local context within which the research activities take place. A localization approach can further foster understanding of, for example, the degree to which the prevailing SDG mapping approaches capture SDG research in the geographical region that may or may not have been described by keywords and key phrases with close semantic relatedness to the keyword-based queries (e.g., Elsevier 2020 queries).

The University of Auckland's approach (Wang, Kang, & Mu, 2023) is one such localization attempt based on Elsevier's earlier 2020 queries, a mixture of the UN official targets and indicators, and the suggested search terms by the Sustainable Development Solutions Network (SDSN). The *n*-gram model was applied to two samples of Scopus publication metadata: a global publication sample and a University of Auckland publication sample. The *n*-gram tokens were scored by a range of factors, including counts and measures of frequency, and were then ranked by those scores. Keywords with a high rank were then evaluated in more detail and manually reviewed and improved for SDG alignments. Table 8 shows the number of University of Auckland publications between 2009 and 2020 captured by the University's queries compared with those captured by Elsevier 2020 queries.

Table 8. Comparison of Auckland v2 queries and Elsevier 2021 queries

SDG Id	Auckland queries output (number of publications)	Elsevier 2020 queries output (number of publications)	Intersection (number of publications)
1	522	229	125
2	1,975	420	264
3	16,894	7,966	6,894
4	2,484	1,043	745
5	611	609	360
6	684	486	362
7	1,152	1,187	799
8	428	440	154
9	1,044	1,139	519
10	1,528	977	500
11	1,886	1,462	779
12	921	438	158
13	1,032	577	466
14	1,390	744	552
15	1,641	769	473
16	891	779	409

For 13 of the 16 SDGs documented in Table 8, the Auckland queries capture more SDG-related publications. In some cases, the number of publications captured by the Auckland queries doubled that captured by the Elsevier 2021 approach. A significant proportion of the additional publications are captured through localized keywords and search terms. For example, "Te Whāriki"—the New Zealand national curriculum document for early childhood education—was used as an SDG4 key phrase under the Auckland approach, as it pinpoints what makes a quality early childhood education curriculum with an indigenous Māori lens. It retrieved 19 SDG4 papers published by the University of Auckland, of which only six were counted by the Elsevier 2020 approach. A manual inspection of these 19 Te Whāriki papers unsurprisingly suggests the high relevance of all 19 papers to SDG4 Target 4.2 on ensuring quality early childhood development, care, and pre-primary education. In some other cases, the Auckland queries also gave rise to additional keywords potentially fitting for the global settings. For example, "marine biodiversity" as an Auckland key phrase retrieved 24 SDG14 papers published by the University of Auckland, of which 19 were counted by the Elsevier 2020 approach.

As shown in Figure 2, applying the Auckland approach to the Aurora (survey & suggested), Elsevier (multilabel), Chilean, and OSDG data sets generates F1 scores that are better for some SDGs (e.g., SDG3, 4, 7, and 14) than others. The F1 scores are notably low for SDGs such as SDG 1, 2, 10, and 12. This suggests that, while the localized approach adds useful keywords and themes in some contexts, further work is required to examine each keyword and key phrase independently to understand their impact on precision and recall and to refine the

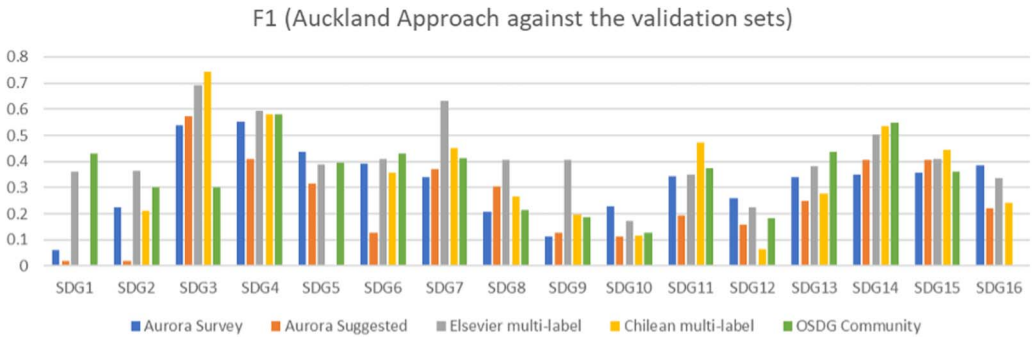

Figure 2. F1-scores for the Auckland approach applied to the Aurora, Elsevier, Chilean, and OSDG data sets.

search conditions upon which they should be applied. In future work, it would also be interesting to develop a contextualized SDG-label set that aligns with the contextualized SDG mapping approach (e.g., an Auckland SDG validation set) to better test out the performance of the contextualized approach against more generic, global approaches.

4.2. Multilingual Queries

In CRIS systems[3] and repositories there are many more publications that are not included in Scopus and are written in the local language of the country to serve a different audience. We found that we could not simply replace the keywords in the queries and have the search work the same in other languages, because of the syntax and morphology rules. That is why Aurora chose to train mBERT models to classify SDGs. Due to the lack of non-English SDG-labeled data, we used only English training data, specifically paper abstracts.

During the evaluation, the models for SDGs 1–5 and 11 were applied to classify 888 German paper titles. To have a qualitative benchmark, we performed a manual SDG classification only on titles as well. In doing the latter, we tended to take a strict approach and tried to stick very close to the respective SDG indicators (e.g., nonassignment of SDG4 to publications on teacher training in Germany, as the SDG indicators only refer to teacher training in the Global South).

The manual classification resulted in 43 SDG-related publications, while the ML models resulted in 58 SDG-related publications. The total overlap between these two methods was eight publications. This was mainly for SDG3—Good Health and Wellbeing (5/8) and can most likely be explained by great similarities in terminology between English and German for issues such as multiple sclerosis, psychotherapy, suicide, alcohol, and illegal drugs (in German: *Multiple Sklerose, Psychotherapie, Suizid, Alkohol, illegale Drogen*).

At the current phase of evaluation, the multilingual ability of the ML models for research output in German cannot be positively assessed. However, further analysis, including the abstracts of publications for the classification of the ML models, may offer improvements in classification quality.

4.3. Generalization to Other Types of Content

In addition to SDG-related research outputs, higher education institutions have a strong interest in understanding SDG-related educational activities, as done in the Aurora SDG Course Catalogue[4]. These SDG labels have been added manually by the course coordinators,

[3] https://en.wikipedia.org/wiki/Current_research_information_system.
[4] https://bit.ly/aurora-sdg-courses.

but such a process is labor-intensive and not sustainable, as this needs to be done year after year.

Similar to publication metadata (e.g., title, abstract, keywords), many course catalogues and curriculum management systems capture metadata in a similar way (e.g., title, course short description, course long description). Whether the SDG research mapping techniques can be translated and applied to SDG course mapping represents an interesting topic to many.

A study was conducted by the University of Auckland to apply the Auckland queries to classify courses taught by the university. The mapping results identified 792 SDG courses out of 2,441 courses in total offered to students in the academic year 2020. Compared with the frequency and distribution of keywords in research mapping, course mapping demonstrated a higher concentration of keywords used to convey the SDG topics. For example, the 24 University of Auckland courses related to SDG14 are fully captured by the top 10 keywords in the Auckland queries by frequency (i.e., marine; fisheries; coastal management; pollut*; aquaculture; marine environment; fisheries management; eutrophical*; aquatic ecosystem; alga*).

5. CONCLUSION

In this paper, we outlined the methodology behind research mapping to the United Nations (UN) Sustainable Development Goals (SDGs), how it compares to other existing methods, and how well it performs with existing SDG validation data sets. We conclude that there is no single best approach performing well across all validation data sets, although Elsevier queries are slightly more stable. We also conclude that there is no single "golden" SDG validation data set. Each one considered in our experiments comes with its own shortcomings, and each data set used in query development bears the intrinsic bias of the SDG interpretations by the query developers. We observed that Elsevier's queries have seen a measurable improvement from the original 2020 version to the 2021/2022/2023 versions. Finally, we discussed possible improvements to the existing approach: localization of the queries and generalization to other languages and data types.

ACKNOWLEDGMENTS

This work is partly an outcome of the SDG Research Mapping Initiative[5] that Elsevier initiated with the Aurora European Universities Alliance, the University of Auckland, and the University of Southern Denmark. We are also grateful to Scopus for providing data for the analysis.

AUTHOR CONTRIBUTIONS

Yury Kashnitsky: Conceptualization, Investigation, Methodology, Software, Supervision, Validation, Writing—original draft, Writing—review & editing. Guillaume Roberge: Conceptualization, Methodology, Writing—review & editing. Jingwen Mu: Investigation, Methodology, Resources. Kevin Kang: Software, Validation. Weiwei Wang: Investigation, Methodology. Maurice Vanderfeesten: Formal analysis, Methodology, Resources. Maxime Rivest: Conceptualization, Methodology, Validation, Writing—review & editing. Savvas Chamezopoulos: Software, Validation. Robert Jaworek: Formal analysis. Maéva Vignes: Investigation, Methodology, Resources. Bamini Jayabalasingham: Methodology, Validation. Finne Boonen: Methodology, Visualization. Chris James: Project administration. Marius Doornenbal: Resources, Supervision. Isabelle Labrosse: Methodology, Visualization.

[5] https://www.elsevier.com/about/sustainability/sdg-research-mapping-initiative.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

No funding has been received for this research.

DATA AVAILABILITY

The data underlying the results presented in the study (including the processed version of publicly available data sets listed in Table 3) are available, partially via GitHub[6] and partially (upon application) from Elsevier BV on the ICSR Lab[7]. ICSR Lab is intended for scholarly research only and is a cloud-based computational platform that enables researchers to analyze large structured data sets, including aggregated data from Scopus author profiles, PlumX Metrics, SciVal Topics, and Peer Review Workbench.

REFERENCES

Armitage, C., Lorenz, M., & Mikki, S. (2020a). *Replication data for: Mapping scholarly publications related to the Sustainable Development Goals: Do independent bibliometric approaches get the same results?* [Data set]. https://doi.org/10.18710/98CMDR

Armitage, C. S., Lorenz, M., & Mikki, S. (2020b). Mapping scholarly publications related to the Sustainable Development Goals: Do independent bibliometric approaches get the same results? *Quantitative Science Studies*, *1*(3), 1092–1108. https://doi.org/10.1162/qss_a_00071

Bedard-Vallee, A., James, C., & Roberge, G. (2023). *Elsevier 2023 Sustainable Development Goals (SDGs) mapping* [Data set]. Elsevier Data Repository. https://doi.org/10.17632/y2zyy9vwzy.1

Bordignon, F. (2021). Dataset of search queries to map scientific publications to the UN sustainable development goals. *Data in Brief*, *34*, 106731. https://doi.org/10.1016/j.dib.2021.106731, PubMed: 33537369

Confraria, H., Ciarli, T., & Noyons, E. (2022). *Countries' research priorities in relation to the Sustainable Development Goals*. MERIT Working Papers 2022-030, United Nations University—Maastricht Economic and Social Research Institute on Innovation and Technology (MERIT). [Web page]. Retrieved March 4, 2024, from https://ideas.repec.org/p/unm/unumer/2022030.html.

Duran-Silva, N., Fuster, E., Massucci, F. A., & Quinquillà, A. (2019). A controlled vocabulary defining the semantic perimeter of Sustainable Development Goals (1.2) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.3567769

Jayabalasingham, B., Boverhof, R., Agnew, K., & Klein, L. (2019). Identifying research supporting the United Nations Sustainable Development Goals (Version 1.0) [Data set]. *Mendeley*. https://doi.org/10.17632/87txkw7khs.1

Jingwen, M., & Weiwei, W. (2022). *The University of Auckland SDG keywords mapping* [Web page]. Retrieved March 4, 2024, from https://www.sdgmapping.auckland.ac.nz/.

LaFleur, M. (2019). *Art is long, life is short: An SDG Classification System for DESA Publications*. DESA Working Paper 159. Available at SSRN: https://dx.doi.org/10.2139/ssrn.3400135.

Nakamura, M., Pendlebury, J., Schnell, J., & Szomszor, M. (2019). *Navigating the structure of research on sustainable development goals* (3rd) [Web page]. Retrieved March 4, 2024, from https://clarivate.com/g/sustainable-development-goals/.

Pukelis, L., Puig, N. B., Skrynik, M., & Stanciauskas, V. (2020). OSDG—Open-source approach to classify text data by UN Sustainable Development Goals [Data set]. *arXiv*. https://doi.org/10.48550/arXiv.2005.14569

Purnell, P. J. (2022). A comparison of different methods of identifying publications related to the United Nations Sustainable Development Goals: Case study of SDG 13—Climate action. *Quantitative Science Studies*, *3*(4), 976–1002. https://doi.org/10.1162/qss_a_00215

Rivest, M., Kashnitsky, Y., Bédard-Vallée, A., Campbell, D., Khayat, P., ... James, C. (2021). Improving the Scopus and Aurora queries to identify research that supports the United Nations Sustainable Development Goals (SDGs) 2021 (Version 3.0) [Data set]. *Mendeley*. https://doi.org/10.17632/9sxdykm8s4.1

Roberge, G., Kashnitsky, Y., & James, C. (2022). Elsevier 2022 Sustainable Development Goals (SDG) Mapping (Version 1.0) [Data set]. *Digital Commons Data*. https://doi.org/10.17632/6bjy52jkm9.1

Rodríguez, P. C., Delpiano, R. R., Meneses, P. S., & Vargas, R. V. (2021). Conjunto de datos: Categorization of articles 2017 with authorship of Pontificia Universidad Católica de Chile, through the SDGS [Data set]. https://bibliotecadigital.oducal.com/Record/ir-11534-69668/Details

Ross, D. (2022). *Impact rankings 2022: Methodology* [Web page]. Retrieved March 4, 2024, from https://www.timeshighereducation.com/world-university-rankings/impact-rankings-2022-methodology.

Schmidt, F., & Vanderfeesten, M. (2021). Evaluation on accuracy of mapping science to the United Nations' Sustainable Development Goals (SDGs) of the Aurora SDG queries. *Zenodo*. https://doi.org/10.5281/zenodo.4917171

South African SDG Hub. (2023). *South African SDG Hub* [Web page]. Retrieved March 4, 2024, from https://sasdghub.up.ac.za/home/.

Vanderfeesten, M., & Jaworek, R. (2022). AI for mapping multilingual academic papers to the United Nations' Sustainable Development Goals (SDGs). *Zenodo*. https://doi.org/10.5281/zenodo.6487606

[6] https://github.com/Yorko/sdg_mapping_queries_n_ml_benchmarks.

[7] https://www.elsevier.com/insights/icsr/lab.

Vanderfeesten, M., Otten, R., & Spielberg, E. (2020). Search queries for "Mapping Research Output to the Sustainable Development Goals (SDGs)" v5.0.2 (Version 5.0.2) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.4883250

Wang, W., Kang, W., & Mu, J. (2023). Mapping research to the Sustainable Development Goals (SDGs), PREPRINT (Version 2). *Research Square*. https://doi.org/10.21203/rs.3.rs-2544385/v2

Wastl, J., Porter, S., Draux, H., Fane, B., & Hook, D. (2020). *Contextualizing sustainable development research*. https://doi.org/10.6084/m9.figshare.12200081.v2

Wulff, D. U., Meier, D. S., & Mata, R. (2023). Using novel data and ensemble models to improve automated labeling of Sustainable Development Goals. *arXiv*. https://doi.org/10.48550/arXiv.2301.11353

Zhang, R., Vignes, M., Steiner, U., & Zimek, A. (2020). Matching research publications to the United Nations' Sustainable Development Goals by multi-label-learning with hierarchical categories. In *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 516–525). Sydney, NSW, Australia. https://doi.org/10.1109/DSAA49011.2020.00066