Survey data of "Mapping Research output to the Sustainable Development Goals (SDGs)"

**This dataset contains information on what papers and concepts researchers find relevant to map domain specific research output to the 17 Sustainable Development Goals (SDGs).**

[Sustainable Development Goals](https://sustainabledevelopment.un.org/sdgs) are the 17 global challenges set by the United Nations. Within each of the goals specific targets and indicators are mentioned to monitor the progress of reaching those goals by 2030. In an effort to capture how research is contributing to move the needle on those challenges, we earlier have made an initial classification model than enables to quickly identify what research output is related to what SDG. (This [Aurora SDG dashboard](https://aurora-network.global/project/sdg-analysis-bibliometrics-relevance/) is the initial outcome as proof of practice.)

In order to validate our current classification model (on soundness/precision and completeness/recall), and receive input for improvement, a survey has been conducted to capture expert knowledge from senior researchers in their research domain related to the SDG. The survey was open to the world, but mainly distributed to researchers from the [Aurora Universities Network](https://aurora-network.global/). The survey was open from October 2019 till January 2020, and captured data from 244 respondents in Europe and North America.

17 surveys were created from a single template, where the content was made specific for each SDG. Content, like a random set of publications, of each survey was ingested by a data provisioning server. That collected research output metadata for each SDG in an earlier stage. It took on average 1 hour for a respondent to complete the survey. The outcome of the survey data can be used for validating current and optimizing future SDG classification models for mapping research output to the SDGs.

**The survey contains the following questions (see inside dataset for exact wording):**

* **Are you familiar with this SDG?**
  + Respondents could only proceed if they were familiar with the targets and indicators of this SDG. Goal of this question was to weed out un knowledgeable respondents and to increase the quality of the survey data.
* **Suggest research papers that are relevant for this SDG (upload list)**
  + This question, to provide a list, was put first to reduce influenced by the other questions. Goal of this question was to measure the completeness/recall of the papers in the result set of our current classification model. (To lower the bar, these lists could be provided by either uploading a file from a reference manager (preferred) in .ris of bibtex format, or by a list of titles. This heterogenous input was processed further on by hand into a uniform format.)
* **Select research papers that are relevant for this SDG (radio buttons: accept, reject)**
  + A randomly selected set of 100 papers was injected in the survey, out of the full list of thousands of papers in the result set of our current classification model. Goal of this question was to measure the soundness/precision of our current classification model.
* **Select and Suggest Keywords related to SDG (checkboxes: accept | text field: suggestions)**
  + The survey was injected with the top 100 most frequent keywords that appeared in the metadata of the papers in the result set of the current classification model. respondents could select relevant keywords we found, and add ones in a blank text field. Goal of this question was to get suggestions for keywords we can use to increase the recall of relevant papers in a new classification model.
* **Suggest SDG related glossaries with relevant keywords (text fields: url)**
  + Open text field to add URL to lists with hundreds of relevant keywords related to this SDG. Goal of this question was to get suggestions for keywords we can use to increase the recall of relevant papers in a new classification model.
* **Select and Suggest Journals fully related to SDG (checkboxes: accept | text field: suggestions)**
  + The survey was injected with the top 100 most frequent journals that appeared in the metadata of the papers in the result set of the current classification model. Respondents could select relevant journals we found, and add ones in a blank text field. Goal of this question was to get suggestions for complete journals we can use to increase the recall of relevant papers in a new classification model.
* **Suggest improvements for the current queries (text field: suggestions per target)**
  + We showed respondents the queries we used in our current classification model next to each of the targets within the goal. Open text fields were presented to change, add, re-order, delete something (keywords, boolean operators, etc. ) in the query to improve it in their opinion. Goal of this question was to get suggestions we can use to increase the recall and precision of relevant papers in a new classification model.

**In the dataset root you'll find the following folders and files:**

* **/00-survey-input/**
  + This contains the survey questions for all the individual SDGs. It also contains lists of EIDs categorised to the SDGs we used to make randomized selections from to present to the respondents.
* **/01-raw-data/**
  + This contains the raw survey output. (Excluding privacy sensitive information for public release.) This data needs to be combined with the data on the provisioning server to make sense.
* **/02-aggregated-data/**
  + This data is where individual responses are aggregated. Also the survey data is combined with the provisioning server, of all sdg surveys combined, responses are aggregated, and split per question type.
* **/03-scripts/**
  + This contains scripts to split data, and to add descriptive metadata for text analysis in a later stage.
* **/04-processed-data/**
  + This is the main final result that can be used for further analysis. Data is split by SDG into subdirectories, in there you'll find files per question type containing the aggregated data of the respondents.
* **/images/**
  + images of the results used in this README.md.
* **LICENSE.md**
  + terms and conditions for reusing this data.
* **README.md**
  + description of the dataset; each subfolders contains a README.md file to futher describe the content of each sub-folder.

**In the /04-processed-data/ you'll find in each SDG sub-folder the following files.:**

* **SDG-survey-questions.pdf**
  + This file contains the survey questions
* **SDG-survey-questions.doc**
  + This file contains the survey questions
* **SDG-survey-respondents-per-sdg.csv**
  + Basic information about the survey and responses
* **SDG-survey-city-heatmap.csv**
  + Origin of the respondents per SDG survey
* **SDG-survey-suggested-publications.txt**
  + Formatted list of research papers researchers have uploaded or listed they want to see back in the result-set for this SDG.
* **SDG-survey-suggested-publications-with-eid-match.csv**
  + same as above, only matched with an EID. EIDs are matched my Elsevier's internal fuzzy matching algorithm. Only papers with high confidence are show with a match of an EID, referring to a record in Scopus.
* **SDG-survey-selected-publications-accepted.csv**
  + Based on our previous result set of papers, researchers were presented random samples, they selected papers they believe represent this SDG. (TRUE=accepted)
* **SDG-survey-selected-publications-rejected.csv**
  + Based on our previous result set of papers, researchers were presented random samples, they selected papers they believe not to represent this SDG. (FALSE=rejected)
* **SDG-survey-selected-keywords.csv**
  + Based on our previous result set of papers, we presented researchers the keywords that are in the metadata of those papers, they selected keywords they believe represent this SDG.
* **SDG-survey-unselected-keywords.csv**
  + As "selected-keywords", this is the list of keywords that respondents have not selected to represent this SDG.
* **SDG-survey-suggested-keywords.csv**
  + List of keywords researchers suggest to use to find papers related to this SDG
* **SDG-survey-glossaries.csv**
  + List of glossaries, containing keywords, researchers suggest to use to find papers related to this SDG
* **SDG-survey-selected-journals.csv**
  + Based on our previous result set of papers, we presented researchers the journals that are in the metadata of those papers, they selected journals they believe represent this SDG.
* **SDG-survey-unselected-journals.csv**
  + As "selected-journals", this is the list of journals that respondents have not selected to represent this SDG.
* **SDG-survey-suggested-journals.csv**
  + List of journals researchers suggest to use to find papers related to this SDG
* **SDG-survey-suggested-query.csv**
  + List of query improvements researchers suggest to use to find papers related to this SDG

Table of Contents

[Metadata 5](#_Toc39740697)

[Dataset Content: 6](#_Toc39740698)

[Dataset Takeaways: 6](#_Toc39740699)

[License and Attribution, Acknowledgements and how to Cite 7](#_Toc39740700)

[Acknowledgements 7](#_Toc39740701)

[Please cite this data set as follows: 7](#_Toc39740702)

[License for reuse: 7](#_Toc39740703)

[License Attribution when reusing this data: 8](#_Toc39740704)

[Example Usage, some results: 8](#_Toc39740705)

[Contributors, full list 9](#_Toc39740706)





# Metadata

|  |  |
| --- | --- |
| meta | data |
| description | Survey data of "Mapping Research Output to the SDGs" presented here is the comprehensive raw, aggregated and processed output of a survey that ran from October 2019 till January 2020. The goal of the survey was to collect data from (senior) researchers, to add and reflect their conceptual knowledge of their specific research domain that contributes to the SDG (Sustainable Development Goals). The outcome of the survey data can be used for validating current and optimizing future SDG classification model for research output. |
| date | 2020-04-23 |
| authors | Maurice Vanderfeesten (VUA); Eike Spielberg (UDE); Yasin Gunes (VUA); |
| contributors | Alessandro Arienzo (UNA); Roberto delle Donne (UNA); Ignasi Salvadó Estivill (URV); José Luis González Ugarte (URV); Didier Vercueil (UGA); Nykohla Strong (UAB); Eike Spielberg (UDE); Felix Schmidt (UDE); Linda Hasse (UDE); Ane Sesma (UEA); Baldvin Zarioh (UIC); Friedrich Gaigg (UIN); René Otten (VUA); Nicolien van der Grijp (VUA); Yasin Gunes (VUA); Maurice Vanderfeesten (VUA) |
| related material | Aurora SDG queries version 4 |
| file formats | data files are all in comma separated files (csv) in UTF-8 encoding. |
| Software used | Data collector: [Aurora Data Collector](https://github.com/Aurora-Network-Global), Survey tool: [Survey Gizmo](https://app.surveygizmo.eu), Data aggregation: [Elastic Search](https://www.elastic.co/elasticsearch/) + [Kibana](https://www.elastic.co/kibana), Data wranging: [Python](https://www.python.org/), [Jupyter Notebook](https://jupyter.org/), [Elastic Painless](https://www.elastic.co/guide/en/elasticsearch/reference/current/modules-scripting-painless.html), [KNIME analytics platform](https://www.knime.com/knime-analytics-platform) |

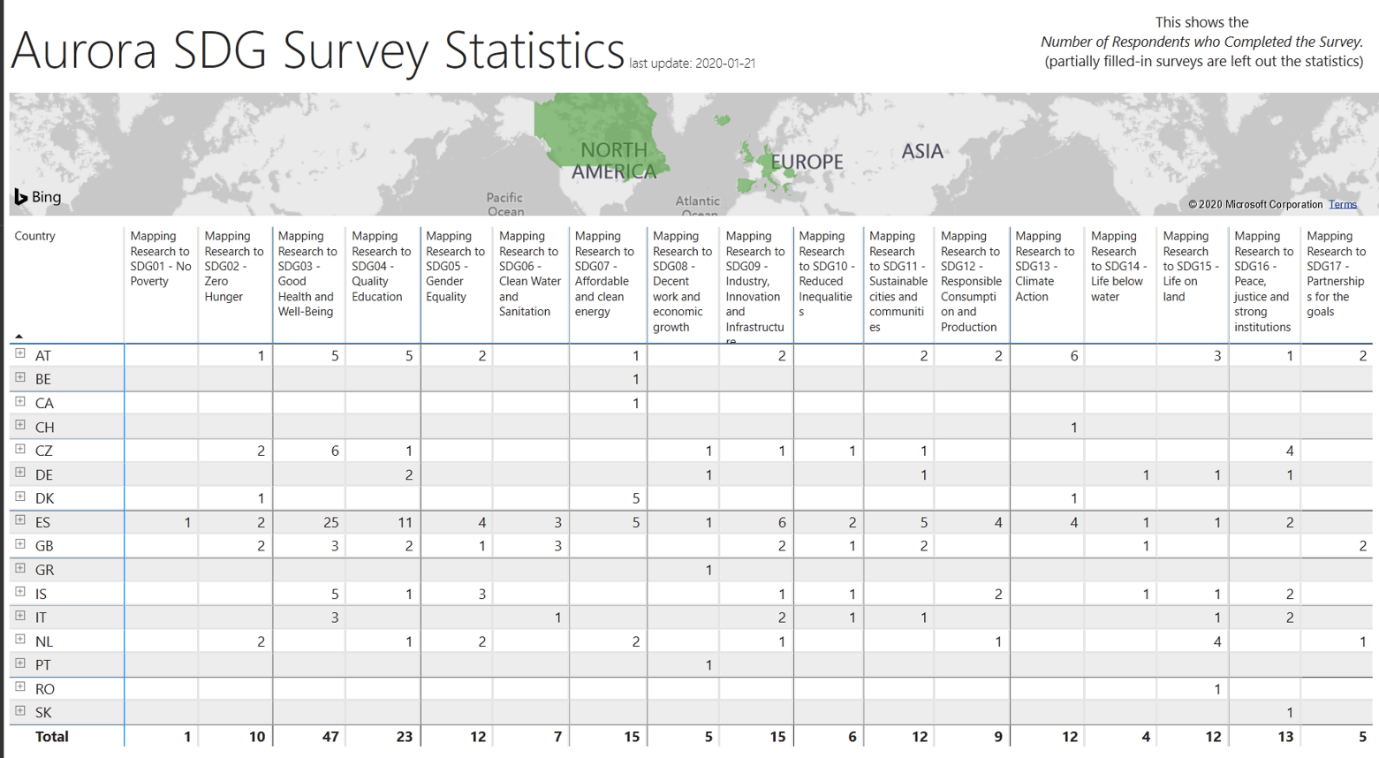
# Dataset Content:

Directories and files. Privacy sensitive data is removed in public release version of the data. All sub-directories contain [README.md](http://README.md) files explaining what data you see and how to interpret.

|  |  |
| --- | --- |
| /dir/ file.\* | description |
| /00-survey-input/ | This contains the survey questions for all the individual SDGs. It also contains lists of EIDs categorised to the SDGs we used to make randomized selections from to present to the respondents. |
| /01-raw-data/ | This contains the raw survey output. (Excluding privacy sensitive information for public release.) This data needs to be combined with the data on the provisioning server to make sense. |
| /02-aggregated-data/ | This data is where individual responses are aggregated. Also the survey data is combined with the provisioning server, of all sdg surveys combined, responses are aggregated, and split per question type. |
| /03-scripts/ | This contains scripts to split data, and to add descriptive metadata for text analysis in a later stage. |
| /04-processed-data/ | This is the main final result that can be used for further analysis. Data is split by SDG into subdirectories, in there you'll find files per question type containing the aggregated data of the respondents. |
| /images/ | images of the results used in this document. |
| LICENSE.md | terms and conditions for reusing this data. |

# Dataset Takeaways:

* data collection period: Oct 2019 till Jan 2020
* respondents: 224 from 16 counties in Europe and North America
* questions asked for each SDG:
  + Suggest research papers that are relevant for this SDG (upload list)
  + Select research papers that are relevant for this SDG (radio buttons: accept, reject)
  + Select and Suggest Keywords related to SDG (checkboxes: accept | text field: suggestions)
  + Suggest SDG related glossaries with relevant keywords (text fields: url)
  + Select and Suggest Journals fully related to SDG (checkboxes: accept | text field: suggestions)
  + Suggest improvements for the current queries (text field: suggestions per target)



# License and Attribution, Acknowledgements and how to Cite

## Acknowledgements

We would like to thank the presidents of the Aurora Universities for their support and executive representation in this project. Also we would like to thank all researchers involved to have shared their expertise. Many people from within the Aurora University Network were involved making this survey possible.

If you want to tribute this hard work, please reuse this data to create or improve your services and share your outcomes.

Do so by respecting the license and attribute the contributors.

## Please cite this data set as follows:

*Survey data of "Mapping Research output to the SDGs"* by Aurora Universities Network (AUR) [doi:10.5281/zenodo.3798385](http://doi.org/10.5281/zenodo.3798385)

## License for reuse:

[Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/legalcode)

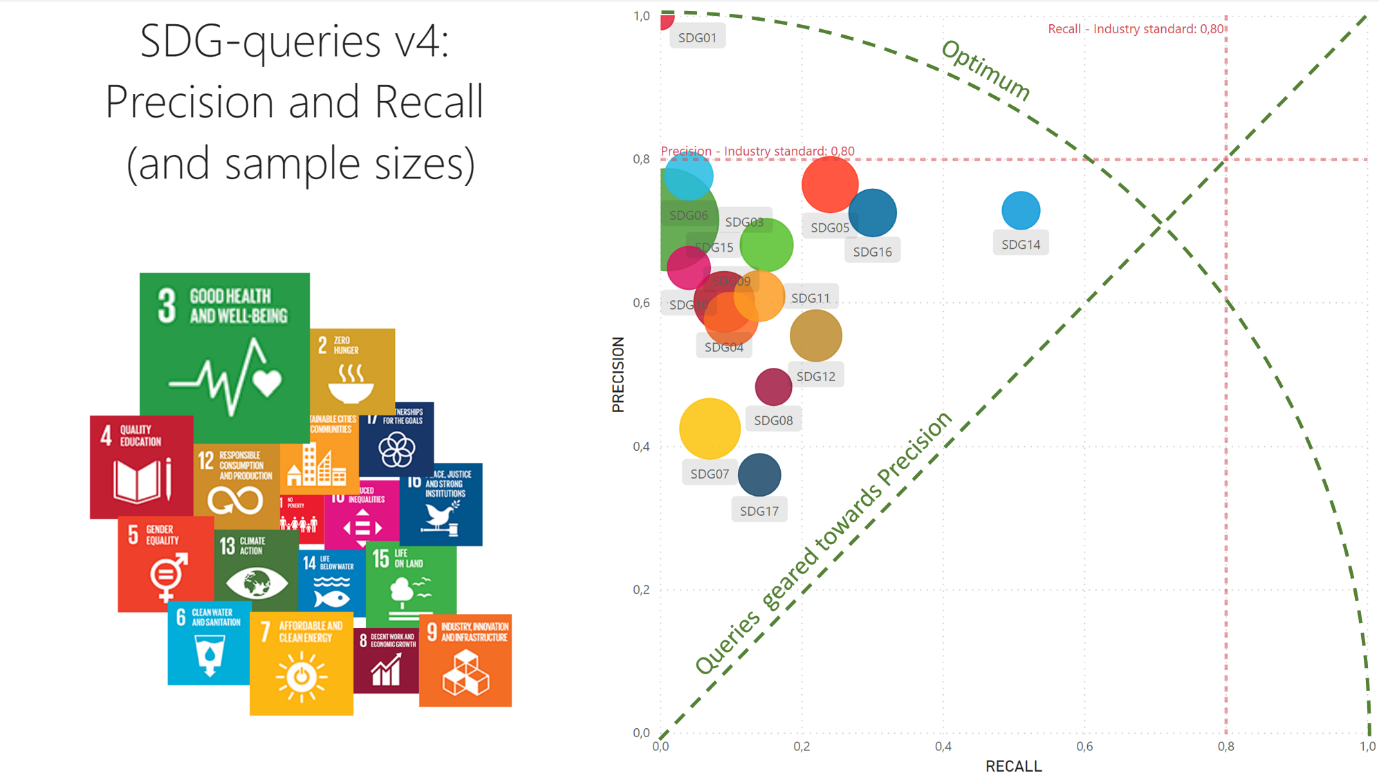
## License Attribution when reusing this data:

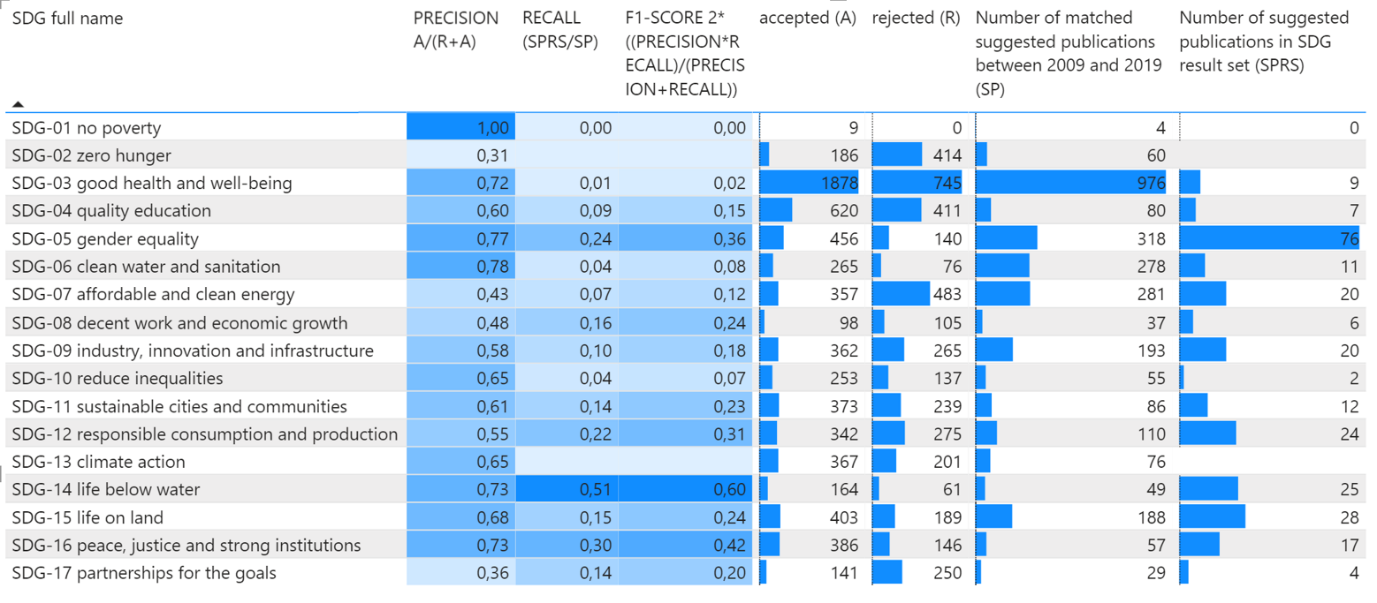
*Survey data of "Mapping Research output to the SDGs"* by Aurora Universities Network (AUR); Alessandro Arienzo (UNA); Roberto Delle Donne (UNA); Ignasi Salvadó Estivill (URV); José Luis González Ugarte (URV); Didier Vercueil (UGA); Nykohla Strong (UAB); Eike Spielberg (UDE); Felix Schmidt (UDE); Linda Hasse (UDE); Ane Sesma (UEA); Baldvin Zarioh (UIC); Friedrich Gaigg (UIN); René Otten (VUA); Nicolien van der Grijp (VUA); Yasin Gunes (VUA); Peter van den Besselaar (VUA); Joeri Both (VUA); Maurice Vanderfeesten (VUA); is licensed under a Creative Commons Attribution 4.0 International License. <https://aurora-network.global/project/sdg-analysis-bibliometrics-relevance/>

# Example Usage, some results:

We used the data for creating a baseline for the SDG classification model, and secondly to gather information on improving a next version.

* precision and recall: how good is our SDG classification model (SDG queries v4)
* suggestions for improving the queries to version 5.





# Contributors, full list

Below you'll see the full list of contributors to this project. With out them this was not possible.

|  |  |  |  |
| --- | --- | --- | --- |
| Family Name, First Name | University | ORCiD | role |
| Vanderfeesten, Maurice | (VUA) Vrije Universiteit Amsterdam | 0000-0001-6397-4759 | Project leader |
| Spielberg, Eike | (UDE) University of Duisburg-Essen | 0000-0002-3333-5814 | Work package leader |
| Gunes, Yassin | (VUA) Vrije Universiteit Amsterdam | ... | Project member |
| Arienzo, Alessandro | (UNA) University Federico II | 0000-0002-2867-5363 | Project member |
| Delle Donne, Roberto | (UNA) University Federico II | 0000-0001-8331-9436 | Project member |
| Salvadó Estivill, Ignasi | (URV) Universitat Rovira i Virgili | ... | Project member |
| González Ugarte, José Luis | (URV) Universitat Rovira i Virgili | ... | Project member |
| Vercueil, Didier | (UGA) Université Grenoble Alpes | ... | Project member |
| Strong, Nykohla | (UAB) University of Aberdeen | 0000-0002-6137-591X | Project member |
| Schmidt, Felix | (UDE) University of Duisburg-Essen | ... | Project member |
| Hasse, Linda | (UDE) University of Duisburg-Essen | ... | Project member |
| Sesma, Ane | (UEA) University of East Anglia | 0000-0003-3982-8932 | Project member |
| Zarioh, Baldvin | (UIC) University of Iceland | 0000-0001-9317-2597 | Project member |
| Gaigg, Friedrich | (UIN) Universität Innsbruck | ... | Project member |
| Otten, René | (VUA) Vrije Universiteit Amsterdam | 0000-0002-6485-8810 | Project member |
| Grijp, Nicolien van der | (VUA) Vrije Universiteit Amsterdam | 0000-0002-5119-3514 | Project member |
| Besselaar, Peter van den | (VUA) Vrije Universiteit Amsterdam | 0000-0002-8304-8565 | Supervisor |
| Both, Joeri | (VUA) Vrije Universiteit Amsterdam | ... | Supervisor |
| Kouwenaar, Kees | AURORA University Network | ... | Sponsor |
| Beukering, Pieter van | (VUA) Vrije Universiteit Amsterdam | 0000-0001-7146-4409 | Sponsor |