CS TRACK
Investigating Citizen Science

Horizon 2020 / Science with and for Society Programme
Grant agreement number: 872522

Web Analytics Toolset and Workbench

D3.2

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 872522

The European Commission's support for this project does not constitute an endorsement of the contents, which reflect the views only of the partners, and the Commission cannot be held responsible for any use which may be made of the information contained

Title of project	CS Track
Full title of project	Expanding our knowledge on Citizen Science through analytics and analysis
Title of this document	Web Analytics Toolset and Workbench
Number of this document	D3.2
Dissemination level	Public
Due date	November 30th, 2021
Actual delivery	December 15th, 2021
Versioning history	Previous versions elaborated over the last two months have been captured in a shared online document.
Authors	Cleo Schulten (RIAS), Sven Manske (RIAS), Fernando Martínez Martínez (URJC), David Roldán Álvarez (URJC), H. Ulrich Hoppe (RIAS)
Executive summary	The "Analytics Workbench" is a software platform developed within the CS Track project as a product of Task T3.2. The main deliverable is a software published on GitHub under an open-source software license. In addition to the software, this document provides an overview of the functional characteristics of the Analytics Workbench. In month 24, a workshop with potential users from the stakeholder groups has been conducted to evaluate the usability and utility of the tool. Preliminary results of the evaluation are included in this deliverable.

Table of Contents

1. Introduction

In D3.1 we presented several approaches and methods that can be utilized to extract information or analyse semantics. We combined and integrated a selection of these approaches as functional components into one web application called the "Analytics Workbench". This accomplishes milestone MS13 as a result of task T3.2 as described in the WP3 specification. The actual deliverable is indeed the Analytics Workbench as a piece of software. This document labelled D3.2 describes the software in terms of its technical implementation (chapter 2) and includes a report on a first evaluation (section 3) and a brief conclusion. Accordingly, the written version of D3.2 should be understood as a documentation and not as a research paper.

CS Track maintains a continuously growing and updated database of CS projects (see D2.1). The goal is to generate a comprehensive and up-to-date overview of CS activities. This includes a sometimes repetitive application of sophisticated algorithms to the database. Since, these algorithms and methods are mostly expert tools, but the stakeholders of the results are most likely not experts with respect to these methods and tools, there is an interest in providing a tool that makes the results of several analyses accessible to non-expert stakeholders. This includes browsing the results providing different views (e.g., network views, word clouds, taxonomical categories like SDGs & research areas, geo maps) on the data to highlight different aspects of CS activities, discovery of related projects, but also adding new projects (not included in the database until now). There are already existing applications serving similar goals like RapidMiner (Mierswa et al., 2006; Mierswa & Klinkenberg 2018), which provide a comprehensive toolsuite for datamining experts. However, we aim at supporting less experienced users for more selective predefined tasks. Business analytics suites like PowerBI[1], a tool for automatic analysis of business figures, or data visualisation tools like Metabase[2] that combine data from different datasources in a dashboard view are easy-to-use but not sufficiently geared towards the specific processing needs and applications of our target users. Specifically, Metabase has been used in the project in parallel to the development of the WP2 database. It provides basic data visualisations but does not support functional or productive applications. In this sense, the Analytics Workbench supports routine data management beyond just visualisation and incorporates algorithms that have been developed and applied for these purposes, now in a more standardised form. An important feature of the workbench that many visual data mining tools still do not have is that it is web-based application which does not require local installations.

The tasks supported by the Analytics Workbench are primarily related to the meso-level of analysis in the distinction introduced in D3.1 (repetitive tasks that can be applied to collections of projects in a largely uniform way with smaller "manual" corrections and adaptations). The main data sources for the Analytics Workbench are the descriptors collected in the WP2 database. The descriptors are filled with data from the WP2 crawler (cf. deliverable D2.1), which extracts unstructured information, particularly textual descriptions, from citizen science project websites. Therefore, the selection of methods related to the ones comprised in D3.1 has a strong focus on text mining and semantic analyses. Micro-level analysis based on project-specific logfile data require expert skills and are not included here. In addition to text-based analyses,

[1] https://powerbi.microsoft.com/

[2] https://www.metabase.com/

the workbench provides different types of data visualisation, including network visualisations. However, the application of sophisticated exploratory network analysis techniques also requires expert skills. In the CS Track project, we rely on tools like Gephi[3] (Heyman & Jacomi, 2009) for these purposes. As a recent extension of the workbench, we have integrated a Lynguo-based module for the analysis of Twitter feeds (macro-level). Again, network and other visualisations are included as specific post-processing steps.

The Analytics Workbench is a web-based application that can be accessed via a graphical user interface running in the browser. In addition, the workbench provides several data interfaces (REST APIs) which can be integrated into other applications. For example, the implemented extraction methods can be directly accessed through the crawler developed in WP2 (cf. D2.1) to enrich the database, particularly the collection of descriptors for citizen science projects.

The utility (functional value) and usability of the Analytics Workbench have been evaluated in a user workshop (milestone MS13), which is documented in Section 3 of this deliverable. The workshop involved possible end users of the Analytics Workbench to exemplify and demonstrate how the results of web analytics can be interpreted. For future work in CS Track, this will help to conduct the Case Studies (T3.3) and to interface with other analyses (along with WP4) and export the results into the community platform (T3.4).

The source code of the Analytics Workbench (which is the core of D3.2) will be published at GitHub under an open-source software license (AGPL v3) with the following URL:

https://github.com/cstrack-code/analytics-workbench

[3] https://gephi.org/[4] Pallets, Welcome to Flask (2010), https://flask.palletsprojects.com/en/2.0.x/. Accessed: 2021-12-08

2. Technical Implementation

This section describes the technical implementation of the Analytics Workbench.

2.1 Architecture, components, and tools

When implementing the workbench, the various functionalities were implemented as separate components. This was done to have them run independent from each other. Additionally, this keeps the functionality separate from the presentation of results which enables us to use the functionalities in different contexts. Therefore, we divided the workbench in frontend, middleware and multiple backends (see Figure 1). Each analysis tool we implemented in a backend that serves the corresponding analysis results. In addition, one backend was implemented to serve as data management for the web application. The frontend manages the web application and holds the APIs for the workbench to make its functionalities available outside of the main application (mainly to the crawler implemented in WP2). The middleware is used to connect frontend to its backends.

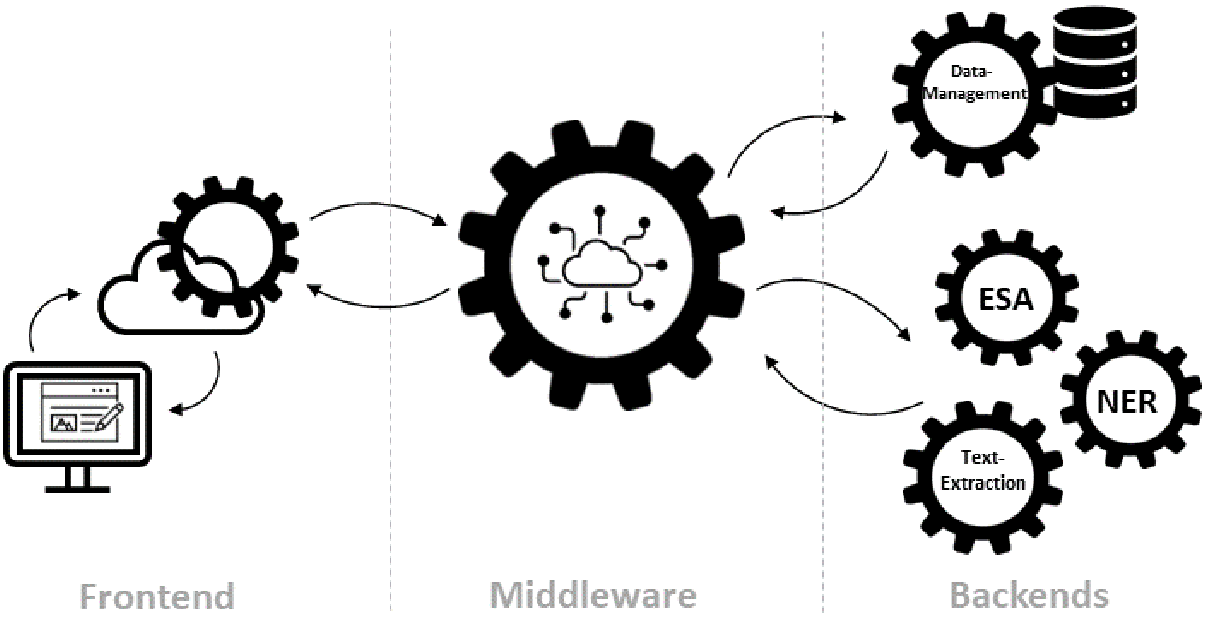


Figure 1: Workbench Architecture

For the realization of the Analytics Workbench, we based our implementation on pre-existing concepts and used already existing tools, mostly in form of python modules, which will be described in the next part before the tools we developed in this context will be described.

2.1.1 Components

Middleware and Backend components of the Analytics Workbench were implemented in python utilizing Flask[4] as a framework. The Text-Extraction Backend is

[4] Pallets, Welcome to Flask (2010), https://flask.palletsprojects.com/en/2.0.x/. Accessed: 2021-12-08

an exception as it was implemented using Node.js. The components communicate with each other via http requests and responses.

Frontend. The Frontend holds the web application as well as an API that can be used from the crawler implemented in WP2. It serves as the public point of contact with the functionalities provided in the respective backends. An overview over the functionalities as well as screenshots of their presentation in the web application can be found in section 2.2 Functional features. Additionally, the Frontend holds the code for the Twitter Analysis described in section 2.2.3.

Middleware. The Middleware serves as the connection between frontend and the backends. In this capacity it is also responsible to check if requested projects have already been analysed as to not re-analyse them needlessly. Since we went through some iterations of code in the backends it also checks against a current version number to ensure that present results are not outdated. In case no or outdated results are found the middleware will have them analysed to provide the results.

Text-Extraction Backend. The Text-Extraction Backend can be used to extract text from websites. It uses the Mercury Web Parser[5] and is implemented in Node.js. It is utilized to extract the reference texts for SDGs ("sustainable development goals") and research areas in the set-up of the Workbench. It can also be used to extract project descriptions from given URLs.

NER Backend. Using the NER ("named entity recognition", cf. D3.1) implemented in spaCy[6] the NER Backend of the Workbench identifies Named Entities from a given project description. For this we use a pretrained model by spacy.

ESA Backend. The ESA Backend holds an implementation of the ESA approach ("explicit semantic analysis", cf. D3.1). In this we use a predefined vector space model based on a Wikipedia dump. It compares given project descriptions to precalculated text vectors of research areas or SDGs. For this the research areas were set with the taxonomy provided within Web of Science[7] and text vectors were calculated from related Wikipedia articles. For the SDGs the set list of 17 SDGs was used and here likewise the related Wikipedia articles were used to calculate text vectors. All precalculated vectors are stored in a MySQL database. The similarities between a project description and all research areas/SDGs are calculated and ranked. At 75% of the highest reached similarity between a research area/SDG and this project a cutoff score is set to assign all research areas/SDGs exceeding this score. This percentage is based on previous trial and error. Additionally, we utilized tf-idf ("term frequency – inverse document frequency", cf. D3.1) to reduce the number of terms from a text that are used to calculate the text vector.

Data Management Backend. The Data Management Backend holds the Workbench's MongoDB database. In the database the collected project data is saved. The database can either be fed through using the Workbench's Web interface, over the API or with local program code that feeds an excel list of projects into the database. The most basic information for a project would be project name,

[5] Postlight, Mercury (2021). https://mercury.postlight.com/. Accessed: 2021-11-25

[6] spaCy, Entity Recognizer: https://spacy.io/api/entityrecognizer. Accessed: 2021-11-25

[7] Clarivate, Web of Science Core Collection Help – Research Areas: https://images.webofknowledge.com/images/help/WOS/hp_research_areas_easca.html. Accessed: 2021-11-25

project URL, and project description, whereas the structure of a completely analysed and user corrected project would be as follows:

```
{
”project_name”: < project name >,
”project_link”: < project link >,
”description”: < description >,
”ra_results”: {
   ”top_classification_areas_with_sim”:
        < list of assigned research areas with similarities >,
   ”classification_areas_with_sim”:
        < list of all research areas with similarities >,
   ”used_tokens”: < list of for esa used tokens >

   ”classification_scheme”: < research_areas or sdgs >

   ”version_control”: < control number for used setting version >
},
”sdg_results”: {
   ”top_classification_areas_with_sim”:
        < list of assigned sdgs with similarities >,
   ”classification_areas_with_sim”:
        < list of all sdgs with similarities >,
   ”used_tokens”: < list of for esa used tokens >

   ”classification_scheme”: < research_areas or sdgs >

   ”version_control”: < control number for used setting version >
},
”ner_results”: {
   ”ner_list”:
       < list of named entity results with label, descriptor, start
         and end character >,
   ”all_descriptors”:
       < list of all available labels with descriptors >
}
”user_ra_results”: < same data structure as ra_results >,
”user_sdg_results”: < same data structure as sdg_results >,
”user_ner_results”: < same data structure as ner_results >
}
```

The analysis of the collected data is also done in the Data Management Backend. For this the data is read from the database and divided in projects with analysis results and projects without results. From the projects with analysis results the Backend creates lists of occurring named entities, research areas and SDGs while also counting each number of occurrences.

The main element of the analysis is the creation of a network of all projects, research areas, SDGs and named entities which is done using NetworkX[8]. In NetworkX it is possible to give each node additional information, with that the nodes can later still be identified as project, research area, SDG or named entity.

For the network that will be shown in the Web Interface’s dashboard non-project nodes (i.e., research areas, SDGs and named entities) that are also leaves in the graph

[8] NetworkX, NetworkX: https://networkx.org/documentation/stable/index.html. Accessed: 2021-11-25

are removed to reduce the memory load in the browser. Additionally, a folded network is created that shows connections between projects with the edges representing common elements between projects. The created networks are translated in vis.js networks before they are handed toward the frontend as this is the format, they will be represented in.

The network also serves as the basis for recommendations. Recommendations are calculated using the PageRank implementation in NetworkX[9]. PageRank can be used both with and without personalization.

2.1.2 Data sources

In this section, we describe all the connected data sources used in the Analytics Workbench that are either functioning as a target for analysis or help to perform the analysis, for example as knowledge sources or ontologies. The data sources can be divided into unstructured, semi-structured and structured data sources (cf. D2.1).

Unstructured data. Project descriptions of Citizen Science (CS) projects are intentionally written by CS project leaders and their respective team to provide information and explanations about the project and its context to the readers of the description. Those descriptions are published on project websites and harvested by the WP2 crawler. As a means of explaining the project to human readers, the descriptions are free and unstructured texts. Therefore, such descriptions are valuable data sources for the Analytics Workbench to be transformed into structured information. The analysis of project descriptions may contain goals, research topics, organizations, stakeholders, or covered SDGs.

Semi-structured data. To gain a larger picture of the CS landscape, we harvest Twitter data ("macro level of analytics") using the Lynguo tool (cf. D3.1). Within the Lynguo tool, queries to monitor Twitter have been created. If a Tweet matches the query, it will be stored in the Lynguo database. Lynguo provides a CSV interface to download all Tweets from the internal storage. Although this dataset is tabular, the fields contain unstructured data such as the Tweet text itself, but also structured data such as account names, references, or hashtags.

Structured data. Algorithms included with the Analytics Workbench support the mapping of unstructured and semi-structured data to semantic interpretations and representations. For this purpose, the workbench has access to connected knowledge sources and external web services that serve structured data. For the method of Explicit Semantic Analysis (ESA, see D3.1, section 3.3.3), knowledge bases necessary to construct the concept space must be used. For the extraction of research areas (cf. section 2.2.1) a mapping between a taxonomy of research areas and corresponding textual descriptions was needed. This led to the creation of a mapping from categories to subordinate research areas and finally to links of corresponding Wikipedia pages. From this knowledge base, text vectors and subsequently concept vectors were created to span the concept space. For the extraction of SDGs (cf. section 2.2.1), an equal mapping from each SDG to a corresponding Wikipedia article was created from which likewise concept vectors are calculated to span the concept space representing the SDGs.

[9] NetworkX Guide, PageRank algorithm (2021). https://networkx.guide/algorithms/link-analysis/pagerank/. Accessed: 2021-11-25

2.2 Functional features

In this section the functional features that can be found in the web application of the workbench are described. These entail Natural Language Processing (NLP) based tools for the analysis of individual projects, project networks that investigate the collected data and a twitter analysis tool.

2.2.1 NLP-based analyses

As stated out in Deliverable D3.1, the processing of natural language ("NLP" = natural language processing) and text analytics are in the core set of methods and tools for CS Track. For most of the projects in the corpus of the CS Track Database (cf. WP2, D2.1), a valuable information source to characterize a project or an activity is the project description. The Analytics Workbench uses the descriptions stored in the database by the crawler and processes information to generate semantic characterisations. This process is fully automated so that all the methods are valid for the whole set of CS projects, if a project description is stored in the database.

The analysis tasks performed in the Analytics Workbench using methods of NLP are shown in *Table 1*.

Table 1: Information extraction tasks and corresponding NLP methods in the Analytics Workbench.

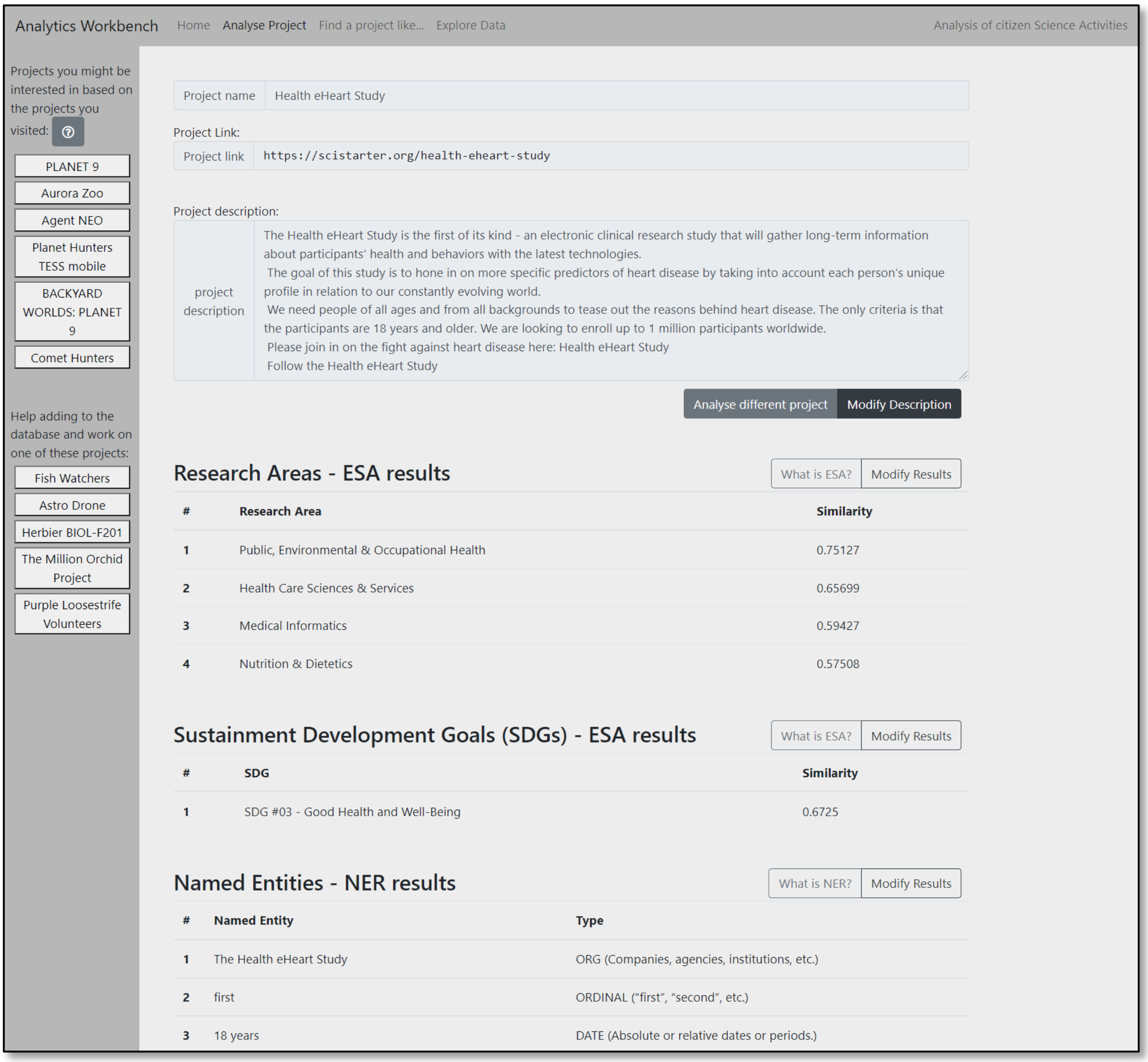
Information Extraction Task	**Method**	**Knowledge Sources**	**Representations / Visualizations**
Extraction of Research Areas	Explicit Semantic Analysis (ESA)	Wikipedia, Research Area Taxonomy ("WoS Core"), tf-idf model (pre-processed), DBpedia	List (per project), bar chart (aggregated)
Extraction of SDGs	Explicit Semantic Analysis (ESA)	Wikipedia, Mapping of SDGs, tf-idf model	List (per project), bar chart (aggregated)
Extraction of Named Entities	Named Entity Recognition (NER)	-	List (per project), network of projects (aggregated)

The representation of the results is dependent on the specific method. ESA delivers a list of explicit concepts such as the extracted research areas or SDGs from a project description. The resulting concepts needed to be defined in the external knowledge source which is used to span the concept space and to construct the word-concept-matrix for the ESA algorithm (cf. D3.1, section 3.3.3). For named entity recognition, the results are typed according to the entity type extracted. This can be cardinal numbers, dates, persons, geolocations, organizations and much more. The full list can be found in the documentation of spaCy[10].

The advantages and mechanisms of methods such as ESA or Named Entity Recognition have been discussed in D3.1, section 3. By using knowledge sources and

[10] spaCy, named entity recognition: https://spacy.io/api/data-formats#named-entities. Accessed: 2021-11-24.

a closed vocabulary for ESA, the results always match the right terminology (i.e., existing names of research areas or SDGs). However, they still might be incorrect due to a "wrong" context or inaccurate due to the quality of the textual description. Therefore, the Analytics Workbench contains functions for the correction of results per project. Users might uncheck or add research areas, SDGs or named entities to a project. In addition to correcting the results of the mechanisms, users also might add new projects or change descriptions stored in the database. However, the changes do not interfere with the WP2 database, as they are kept in the local database of the Analytics Workbench (cf. section 2.1.1, Data Management). Figure 2 shows the analysis of project descriptions in the user interface of the workbench.



Projects you might be interested in based on the projects you visited:

Help adding to the database and work on one of these projects:

Project name: Health eHeart Study

Project Link:

Project link: `https://scistarter.org/health-eheart-study`

Project description:

project description: The Health eHeart Study is the first of its kind - an electronic clinical research study that will gather long-term information about participants' health and behaviors with the latest technologies.
The goal of this study is to hone in on more specific predictors of heart disease by taking into account each person's unique profile in relation to our constantly evolving world.
We need people of all ages and from all backgrounds to tease out the reasons behind heart disease. The only criteria is that the participants are 18 years and older. We are looking to enroll up to 1 million participants worldwide.
Please join in on the fight against heart disease here: Health eHeart Study
Follow the Health eHeart Study

Research Areas - ESA results

#	Research Area	Similarity
1	Public, Environmental & Occupational Health	0.75127
2	Health Care Sciences & Services	0.65699
3	Medical Informatics	0.59427
4	Nutrition & Dietetics	0.57508

Sustainment Development Goals (SDGs) - ESA results

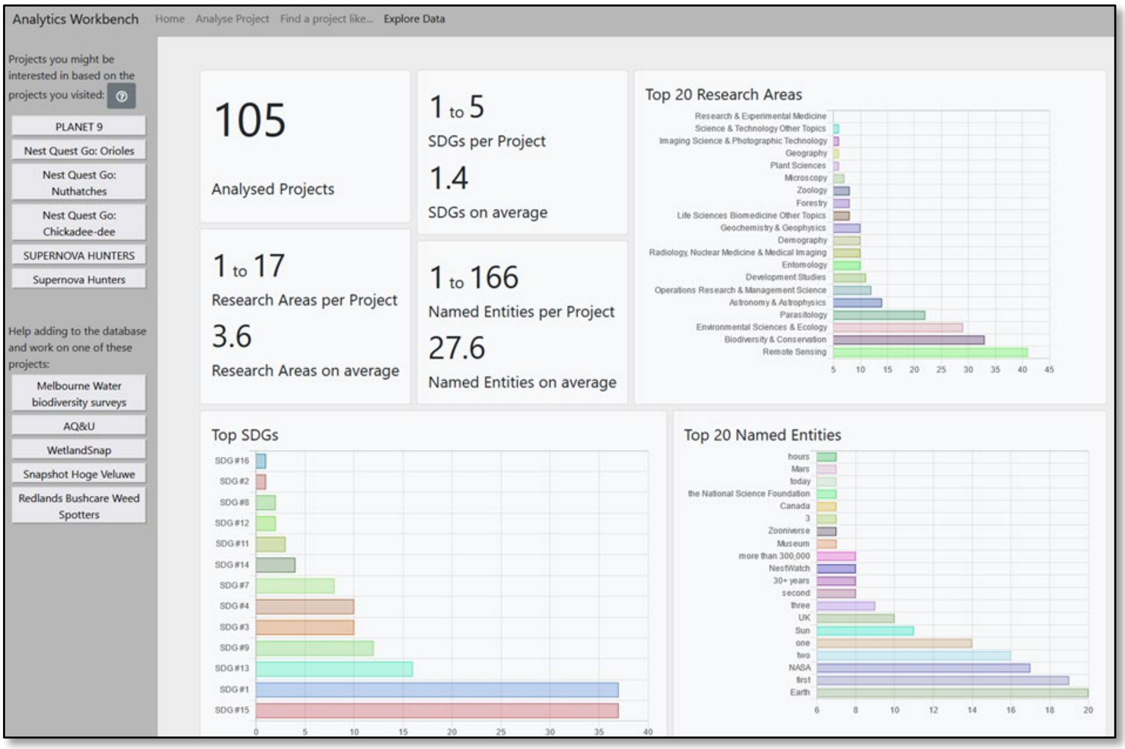
#	SDG	Similarity
1	SDG #03 - Good Health and Well-Being	0.6725

Named Entities - NER results

#	Named Entity	Type
1	The Health eHeart Study	ORG (Companies, agencies, institutions, etc.)
2	first	ORDINAL ("first", "second", etc.)
3	18 years	DATE (Absolute or relative dates or periods.)

Figure 2: Interface for the analysis of project descriptions. Users select a project, and if it already exists in the database, the project description is fetched from the database. The analysis triggers all tasks asynchronously and informs the user once it is finished.

In addition to the individual analysis of projects, all the results are also visualized in aggregated representations over the whole set of projects. The named entities also play an important role in the construction of project networks because they can be conceived as connectors between projects (see next section).



Projects you might be interested in based on the projects you visited:

Help adding to the database and work on one of these projects:

105

Analysed Projects

1 to 5

SDGs per Project

1.4

SDGs on average

1 to 17

Research Areas per Project

3.6

Research Areas on average

1 to 166

Named Entities per Project

27.6

Named Entities on average

Figure 3: Aggregated results from the extraction of research areas, SDGs and named entities.

2.2.2 Project networks

Although the projects do not have an inherent or explicit network structure, we might observe aspects that connect projects. Different activities might share similar research topics, similar sustainable development goals or even organizations. Therefore, the backend of the Analytics Workbench creates networks based on the (semantic) connectors that are derived from the NLP functionality (see previous section).

The following features of the Analytics Workbench facilitate network structures (for Twitter-based network features see section 2.2.3):

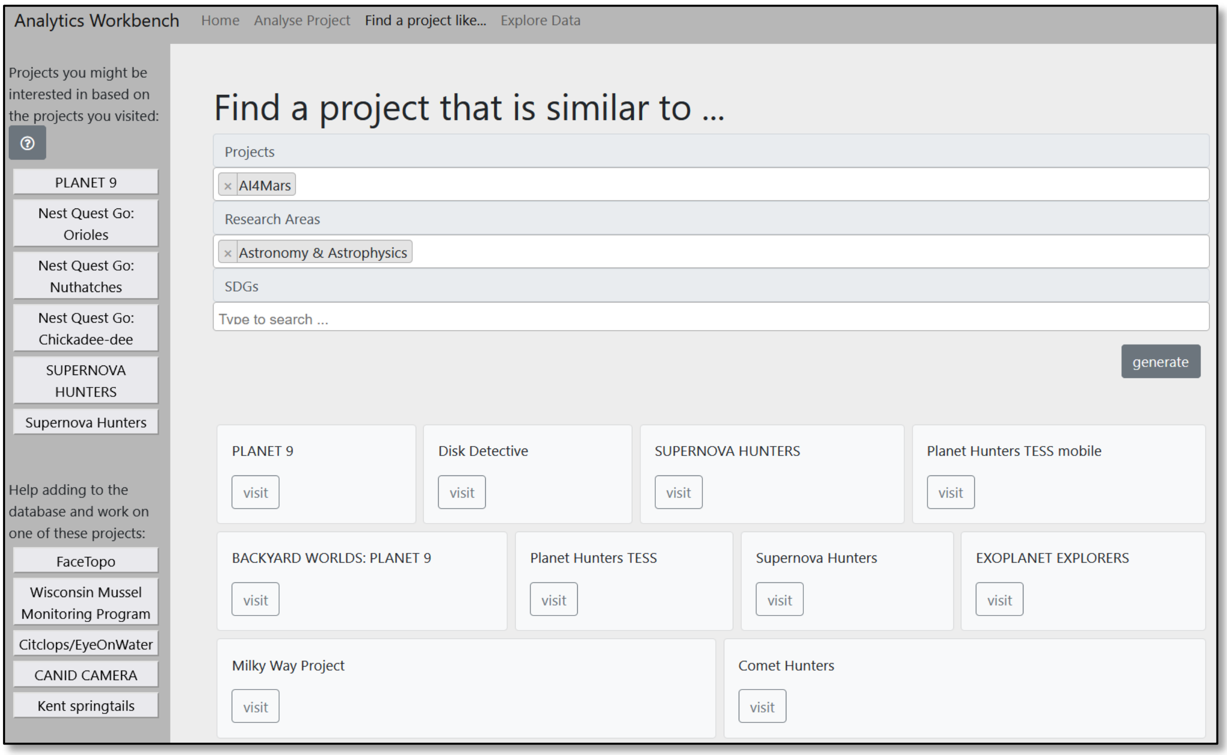
- Recommender system (for project discovery):
    - Find similar projects (globally).
    - Find similar projects based on individual trajectories (personalized PageRank).
- Explore and filter connections of projects
    - Multi-mode network (projects, research areas, SDGs, named entities),
    - Fold the network to a one-mode network (project-project relation),
    - Filter networks (node type, by project name),
    - k-Core filtering,
    - find and highlight nodes.

For the construction of the network, the following rules are applied:

- Nodes are created for projects and extracted entities from text analytics such as research areas, SDGs and named entities (organizations, persons, locations, …).
- An edge between two nodes is established based on co-occurrences, particularly, when an item I is extracted from a description of project P, then the edge (P, I) is added to the set of edges.

By folding the network, it is possible to create a network of projects and from this network, we can determine measure for the similarity of projects. The recommendation is based on a personalized and non-personalized version of the PageRank algorithm (Page et. al, 1999). In the "Find a project like…" tab, users can explore the recommender system. The main panel (Figure 4) shows the global recommender system. This is based on a PageRank algorithm, where the user can select a single or multiple projects, research area(s) or SDGs as seeds for the PageRank. The results are displayed as a list with links to the recommended projects. This enables to further inspect the adjacent project descriptions, research areas, SDGs or named entities.

In the left panel of the recommender system, the user has the access to recommendations from a personalized PageRank, which is based on the individual trajectories of projects visited. In the bottom part of this tab, another category of recommendations which are not based on network features is placed. Those recommendations are based on missing data in the database of the Analytics Workbench.



Projects you might be interested in based on the projects you visited:

Help adding to the database and work on one of these projects:

Find a project that is similar to ...

Projects: × AI4Mars

Research Areas: × Astronomy & Astrophysics

SDGs: Type to search ...

Figure 4: User interface for the recommender system. Main panel: actively triggered (and controlled) recommendations; Left panel: personalized recommendations.

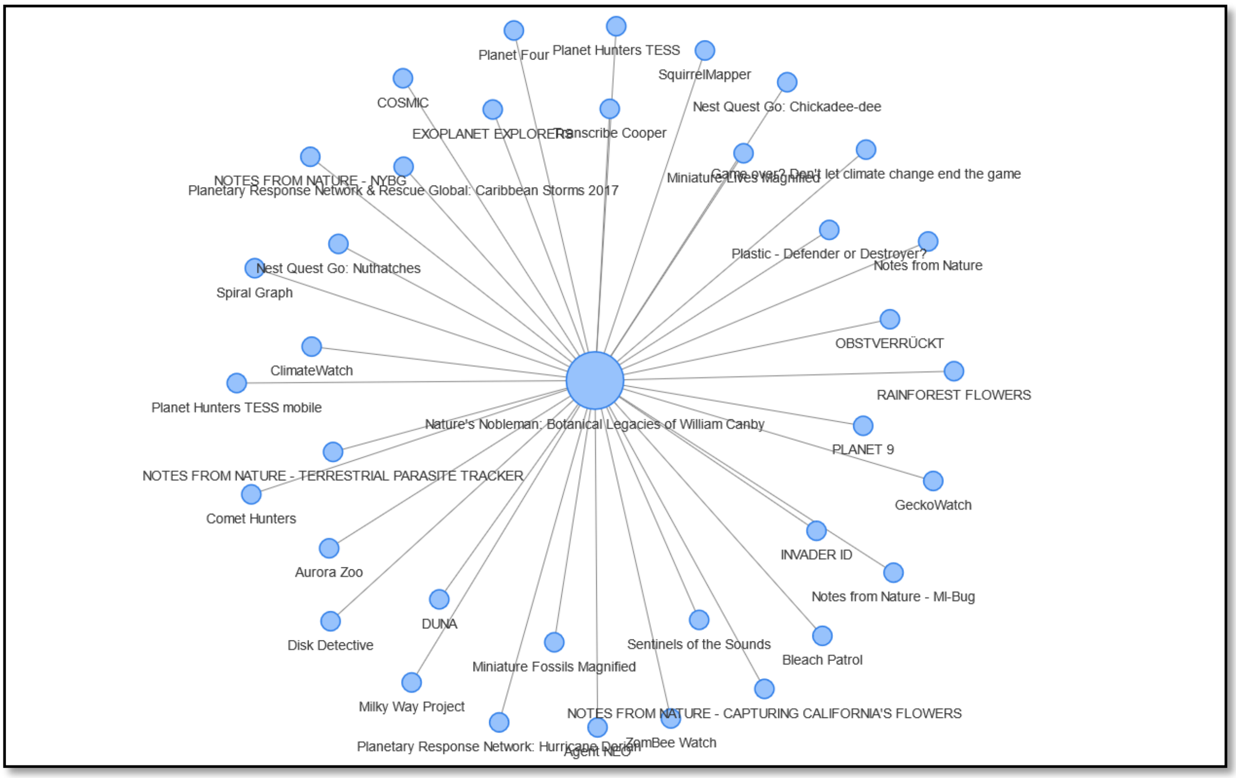

Planet Four
Planet Hunters TESS
SquirrelMapper
COSMIC
Nest Quest Go: Chickadee-dee
EXOPLANET EXPLORERS
Transcribe Cooper
NOTES FROM NATURE - NYBG
Planetary Response Network & Rescue Global: Caribbean Storms 2017
Miniature Lives Magnified
Game over? Don't let climate change end the game
Plastic - Defender or Destroyer?
Notes from Nature
Nest Quest Go: Nuthatches
Spiral Graph
OBSTVERRÜCKT
ClimateWatch
RAINFOREST FLOWERS
Planet Hunters TESS mobile
Nature's Nobleman: Botanical Legacies of William Canby
PLANET 9
NOTES FROM NATURE - TERRESTRIAL PARASITE TRACKER
GeckoWatch
Comet Hunters
INVADER ID
Aurora Zoo
Notes from Nature - MI-Bug
DUNA
Disk Detective
Sentinels of the Sounds
Bleach Patrol
Miniature Fossils Magnified
Milky Way Project
NOTES FROM NATURE - CAPTURING CALIFORNIA'S FLOWERS
Planetary Response Network: Hurricane Dorian
Agent NEO
ZomBee Watch

Figure 5: Ego-network for a single project. Connected nodes are similar to the central node regarding on of the specific connectors (research areas, SDGs, named entities).

The similarity of projects can be also perceived in the "Explore projects" tab, which shows by default an ego-network for the previously selected project. Figure 5 shows this network for the "Nature's Nobleman: Botanical Legacies of William Canby" project. All nodes connected to this project are other similar projects. The similarity is, as stated out previously, determined by the connection through a node of one of the types (1) *research area*, (2) *SDG*, or (3) *named entity* (including various sub-types). For this view, the network is folded to a one-mode network only showing the projects.

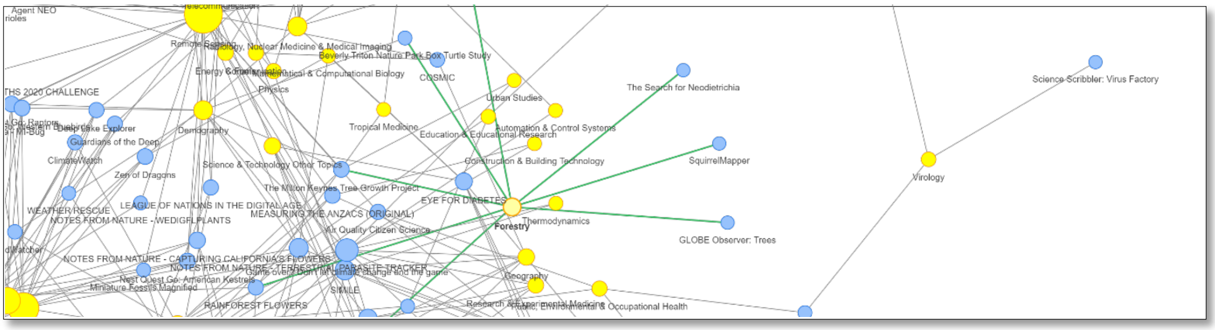
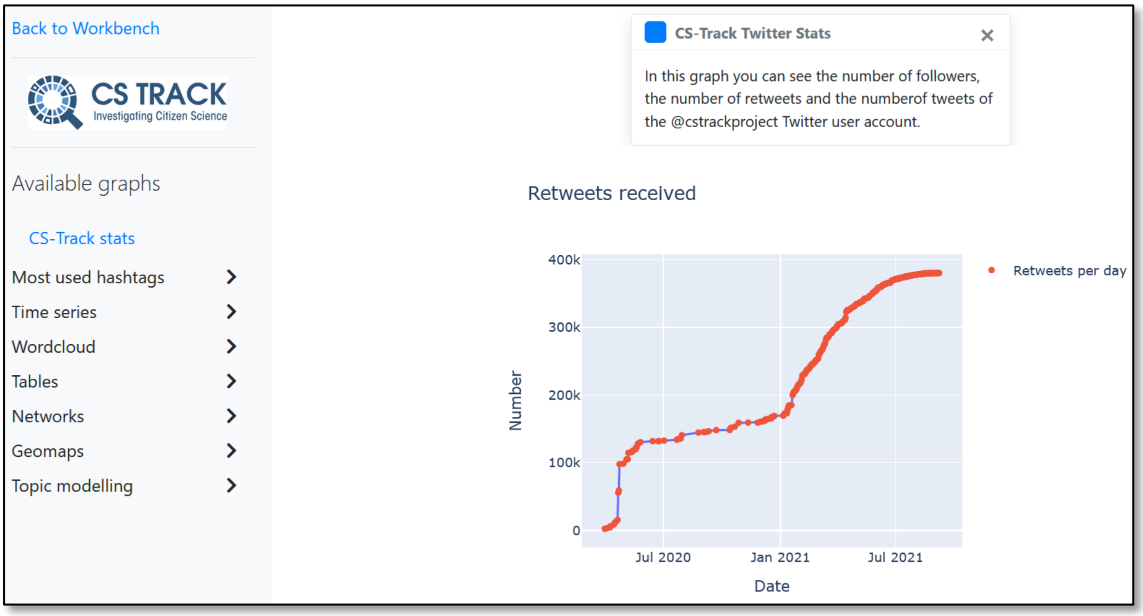

Agent NEO
Beverly-Triton Nature Park Box Turtle Study
COSMIC
Physics
Urban Studies
The Search for Neodietrichia
Science Scribbler: Virus Factory
Demography
Tropical Medicine
Education & Educational Research
Automation & Control Systems
Guardians of the Deep
ClimateWatch
Zen of Dragons
Science & Technology Other Topics
Construction & Building Technology
SquirrelMapper
Virology
The Million Keynes Tree Growth Project
WEATHER RESCUE
LEAGUE OF NATIONS IN THE DIGITAL AGE
NOTES FROM NATURE - WEDIGFLPLANTS
MEASURING THE ANZACS (ORIGINAL)
EYE FOR DIABETES
Forestry
Thermodynamics
Air Quality Citizen Science
GLOBE Observer: Trees
NOTES FROM NATURE - CAPTURING CALIFORNIA'S FLOWERS
NOTES FROM NATURE - TERRESTRIAL PARASITE TRACKER
Game over? Don't let climate change end the game
Nest Quest Go: American Kestrels
Miniature Fossils Magnified
Geography
SIMILE
RAINFOREST FLOWERS
Public, Environmental & Occupational Health

Figure 6: Portion of the extracted and displayed project network in the "explore results" tab of the Analytics Workbench. Blue nodes represent projects that are connected to a certain research topic (yellow nodes). The connections are established based on the information extraction for projects.

Figure 6 shows a network of projects (blue nodes) that are connected through research areas (yellow nodes). In this figure, the research area forestry and the adjacent projects are highlighted. This relation has been established, because in all the adjacent projects' descriptions, the research area "forestry" has been extracted using the ESA algorithm (see previous section).

2.2.3 Twitter analysis



In this graph you can see the number of followers, the number of retweets and the numberof tweets of the @cstrackproject Twitter user account.

Figure 7: Dashboard startpage

In the Twitter analysis Dashboard (Figure 7), the main features of the tweets are explored and visualized using several techniques. They provide different views on the CS projects from the Twitter data.

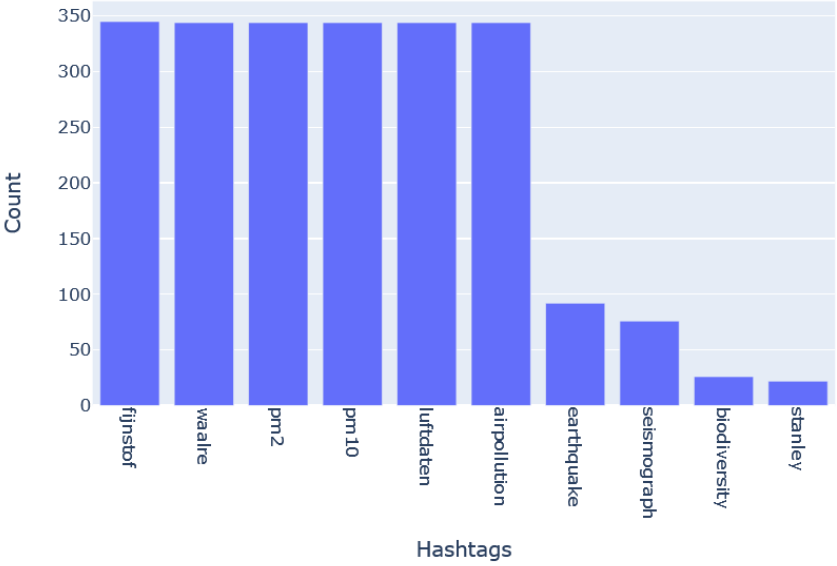
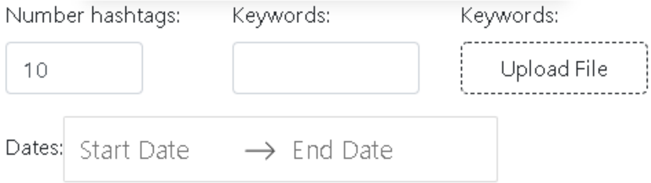


Figure 8: Most used hashtags

Most used hashtags. This section presents the bar charts of most used hashtags and most retweeted hashtags. In the most used hashtags, the presence of each hashtag in the Lynguo database is counted. This can be filtered applying keywords, uploading a list of predefined keywords, and selecting a starting and end date to refine the search of hashtags. Besides, it can be changed the number of hashtags displayed in the graph. The appearance of this graph is shown in Figure 8.

In the most retweeted hashtags analysis, hashtags that received the higher number of retweets are displayed. Like in the section before all the analysis can be filtered according to the same possibilities. The filtering process shown in Figure 9 is also available in the time series analysis.



Number hashtags: 10

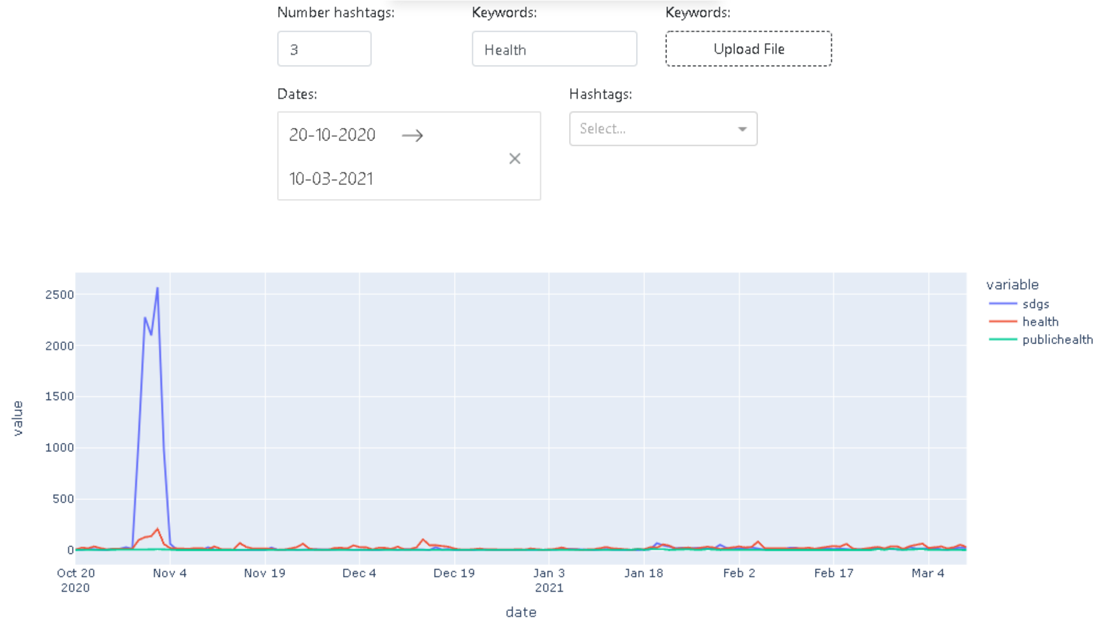
Keywords:

Keywords: Upload File

Dates: Start Date → End Date

Figure 9: Filters for hashtags analysis

Time series. In the time series analysis, every use of a hashtag or retweet is analysed and displayed through the time. In this example the starting point was the 29th of September 2020 so the usage or number of retweets can be displayed starting from that point. Once more these graphs can be filtered according to the previous filters described alongside a new functionality which allows the selection of one specific hashtag. An example of use of this section is shown in Figure 10.



Number hashtags: 3

Keywords: Health

Keywords: Upload File

Dates: 20-10-2020 → 10-03-2021

Hashtags: Select...

Figure 10: Use of the time series analysis

Word cloud. In the Word cloud section, the most used noun in the dataset are displayed.

Degree and sentiment analysis. The Degree analysis shows the users with the higher degrees in the database, which is calculated according to the number of retweets these users received.
The sentiment analysis the users with higher sentiment are presented. This sentiment is calculated through natural language processing methods, text analysis,

computational linguistics, and biometrics to show a subjective appreciation to each user tweets.

Retweets, Retweets Communities and Two-mode networks. The network of retweets graph shows the users that receive the highest number of retweets in a redder colour and bigger size as it can be observed in Figure 10.

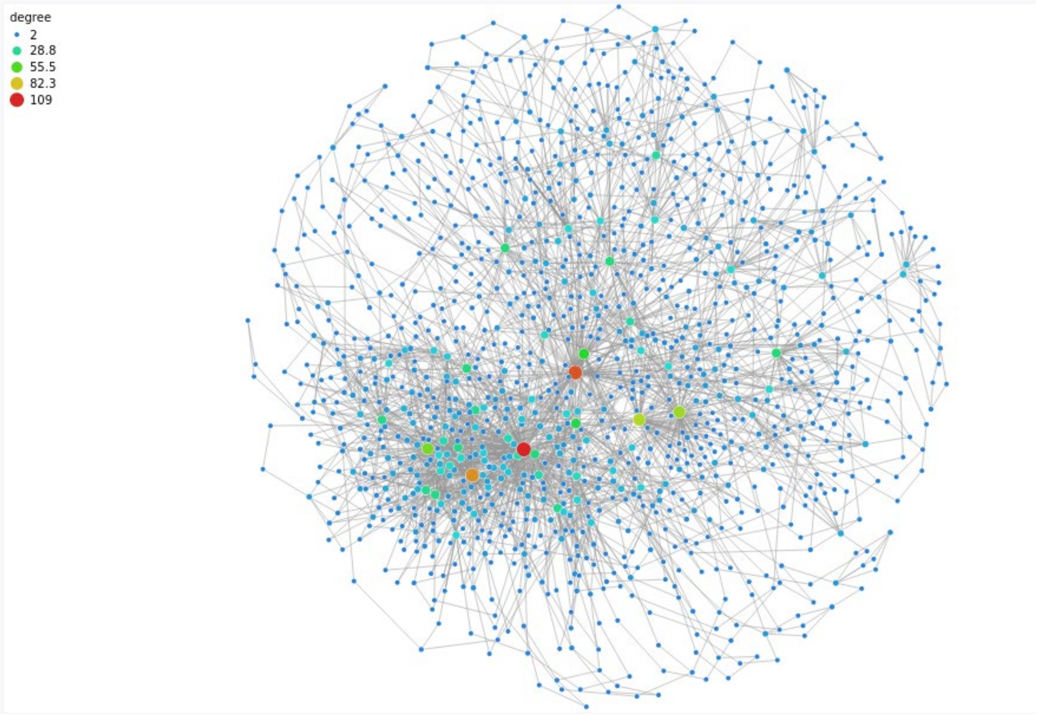


Figure 11: Network of retweets

The Retweets Communities allow to see the communities formed around certain users. At first glance a graph with nodes of different colours can be seen and once we select one node (community) the whole community graph is displayed allowing the analysis of individual nodes such as the central node around which the community evolves around.

The Two-mode section displays a bipartite directed graph. This graph is composed of two set of nodes, the first set is the user and the other the tweets. Thus, this allows the visualization of the most retweeted tweets in the database.

Geomaps. A Geomaps section presents the information about location and countries of the users in the database. The first subsection, Tweets and Followers per country, the number of tweets per country will be shown and the number of followers to users

Figure 11.

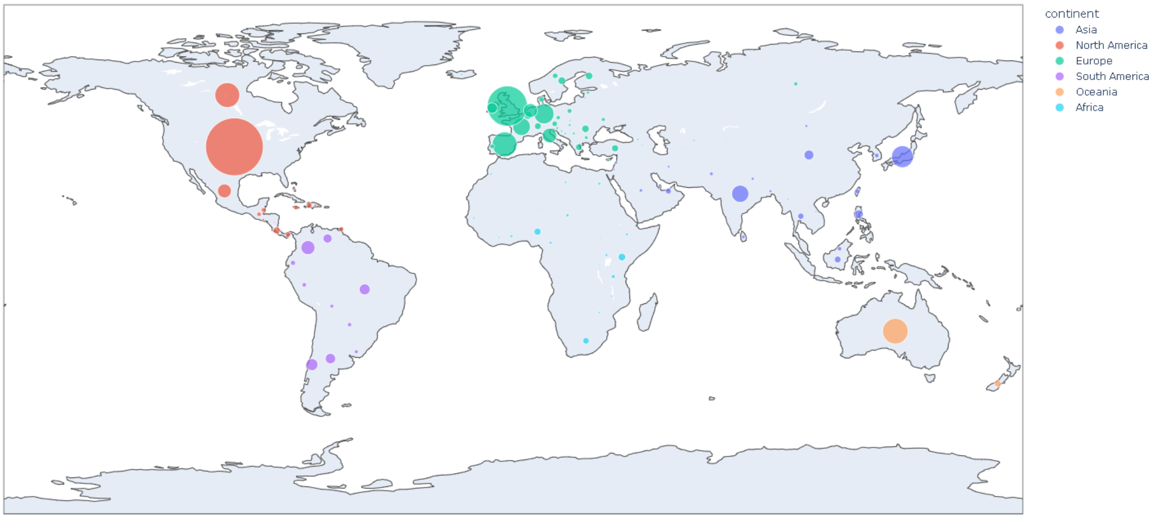


Figure 12: Tweets per country

Additionally, a world map displaying the location of the users in the database see Figure 12

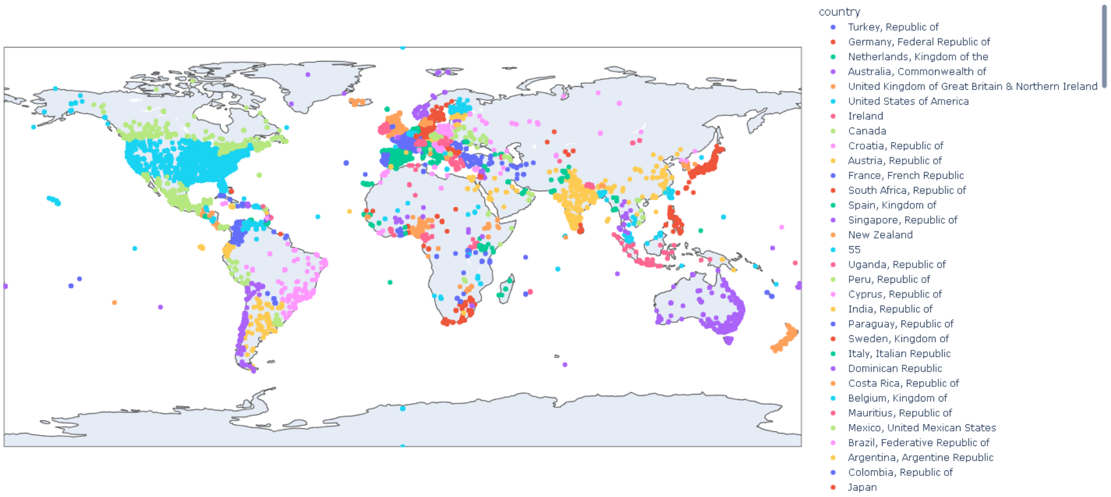


Figure 13: Locations of the users in the Lynguo database

3. Evaluation

We ran a workshop in mid-November to present the Analytic Workbench to participants and use the data collected in the Workshop to evaluate the Workbench under aspects of utility and usability. The usability orientation followed the User Experience Questionnaire (UEQ-S; Schrepp, Hinderks & Thomaschewski, 2017). The utility evaluation was based on predefined tasks leading the participants through a prototypical application case. The guidance was provided through a structured online questionnaire (SoSci Survey[11]). This section documents the workshop and its results.

3.1 Workshop documentation

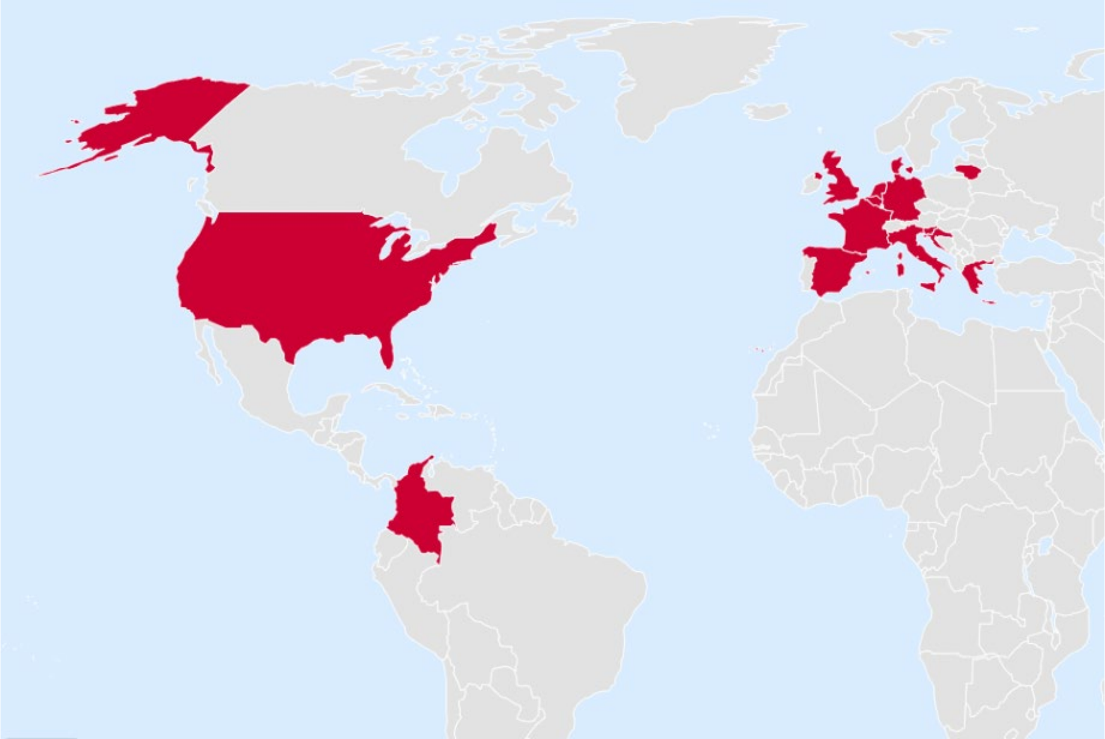


Figure 14: Global distribution of participants

The workshop was held online using zoom and was visited by 22 participants (plus 3 hosts from ATiT and 3 presenters from RIAS). As shown in Figure 16 the participants came predominantly from Europe, more concretely from nine EU member states as well as Switzerland and the United Kingdom. Additionally, we had one participant each from the United States and Colombia. After an introduction and a demo of the Workbench the participants were instructed to follow a questionnaire to analyse a project and answer questions in the process. The schedule for the workshop was the following:

[11] SoSci Survey GmbH, soSci, https://www.soscisurvey.de/

Time	Presenter	Topic
14:00	H. Ulrich Hoppe	Reception and CS Track introduction
14:10	Sven Manske	Analytics Workbench introduction
14:20	Cleo Schulten	Analytics Workbench demo
14:30	Cleo Schulten	Hands-on task description
14:35	Participants	Hands-on task / questionnaire
15:15	H. Ulrich Hoppe	Questions and remarks
15:25	H. Ulrich Hoppe and Sally Reynolds	Closing of the Workshop

In the beginning of the interaction with the questionnaire, each participant was randomly given an individual citizen science project that they could then use in the then following questions that relate to a specific project.

The first questions asked the participants to read the project description of their assigned project and give their own opinion on possible research areas and SDGs that would fit the project. Next, they were supposed to have the Workbench analyse their project and report on some of the given results as well as indicate whether they found the results fitting.

After that they were asked to use the dashboard view to investigate connections between their project and other projects in the database. Additionally, they were supposed to report on the predominant research area, SDG, and named entity as well as a project each is connected to respectively.

As a last interaction with the Workbench the participants were asked to generate recommendations based on their project and a list of given interests.

Subsequently followed the short User Experience Questionnaire (UEQ-S; Schrepp, Hinderks & Thomaschewski, 2017) as well as open feedback questions and questions regarding the helpfulness of the individual functionalities of the workbench.

Finally, the participants were asked what role they see themselves in within the CS context.

For the concrete questions in the questionnaire see Annex 1.

3.2 Evaluation of results

Participants.17 participants of the workshop started the questionnaire and 13 completed it. Of these 13 participants seven stated they were engaged in CS-related research, one identified him-/herself as a professional scientist, the other five selected other options and gave more information on their role in an open text field. Out of those five, one indicated the role as a researcher in related topics, one said (s)he was supporting researchers using CS methodology, one declared to be a community manager for a CS project, another one works in the design of CS tools and infrastructures and the last one just declared not being an expert in this context.

Individual projects. Of the 15 participants that answered the questions regarding the assigned research areas, SDGs and organisations for their project (cf. Annex 1, questions 3, 5 and 7), 12 correctly reported the assigned research areas, one gave only the similarities, one used the wrong project but gave the correct research area for that project and one participant's answer was uninterpretable. The same pattern was found in the replies for the SDGs. For the organisations all but one participant gave one or more identified organisation for the project they analysed, the remaining one gave a product instead.

Finding connections in the network view. In the questions asking the participants to find projects connected to their project with the network view (cf. Annex 1, questions 8-10) only two out of 14 participants were entirely unable to give correct project names. Two other participants seem to have misunderstood one of the questions and gave an organisation or an incomplete project name. The same pattern reoccurred at the question asking for the connector that links their project to the most other projects (i.e., the neighbour with the highest degree).

Global view on collected data. All 14 participants correctly identified the predominant research area in the workbench (cf. Annex 1, question 11) and all except one gave a correct project linked to this research area. Equally all participants correctly named the predominant SDG (cf. Annex 1, question 12), though for this two failed to name a linked project. When reporting one of the named entities listed in the top 20 (cf. Annex 1, question 13), all 14 participants correctly named one, but four failed to name a connected project, with one giving no answer at all.

Recommendations. When creating recommendations based on their project the participants were asked to report their input for the recommendation (cf. Annex 1, question 14). Eleven participants chose appropriate input values ranging from the project name to assigned research areas or SDGs and various combinations. The remaining two that answered these questions gave an unconnected research area and an entirely different project name respectively.

UEQ-S. The results of the performed short UEQ-S (cf. Annex 1, question 17) are shown in Figure 17. These are based on the answers of all 13 participants that answered this part of the questionnaire. In pragmatic quality the Workbench reaches a value of 0.904, for hedonic quality 1.173 and a value of 1.038 overall. These values are on a scale from -3 to 3 and values above 0.8 constitute a positive result regarding the usability of the Analytics Workbench.

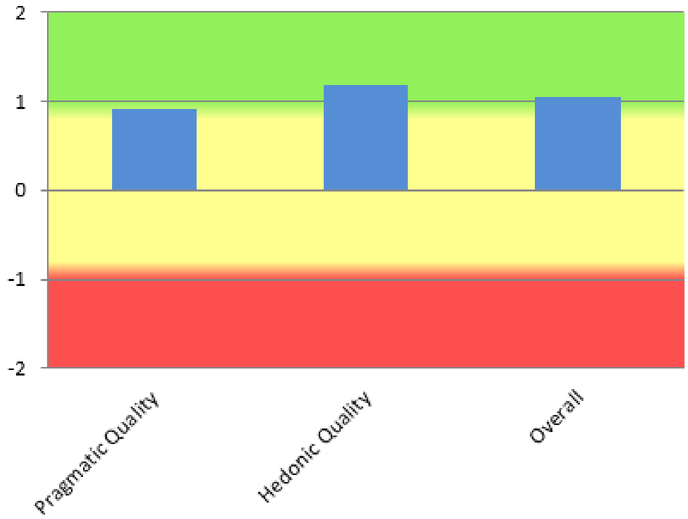


Figure 15: UEQ-S scores

Helpfulness. Additionally, we asked the participants to rate each functionality on their perceived helpfulness. The corresponding results are pictured in Figure 18. To calculate the mean answer per functionality we coded the answers as follows: *not helpful* – 0, *somewhat helpful* – 1, *helpful* - 2 and *very helpful* – 3.

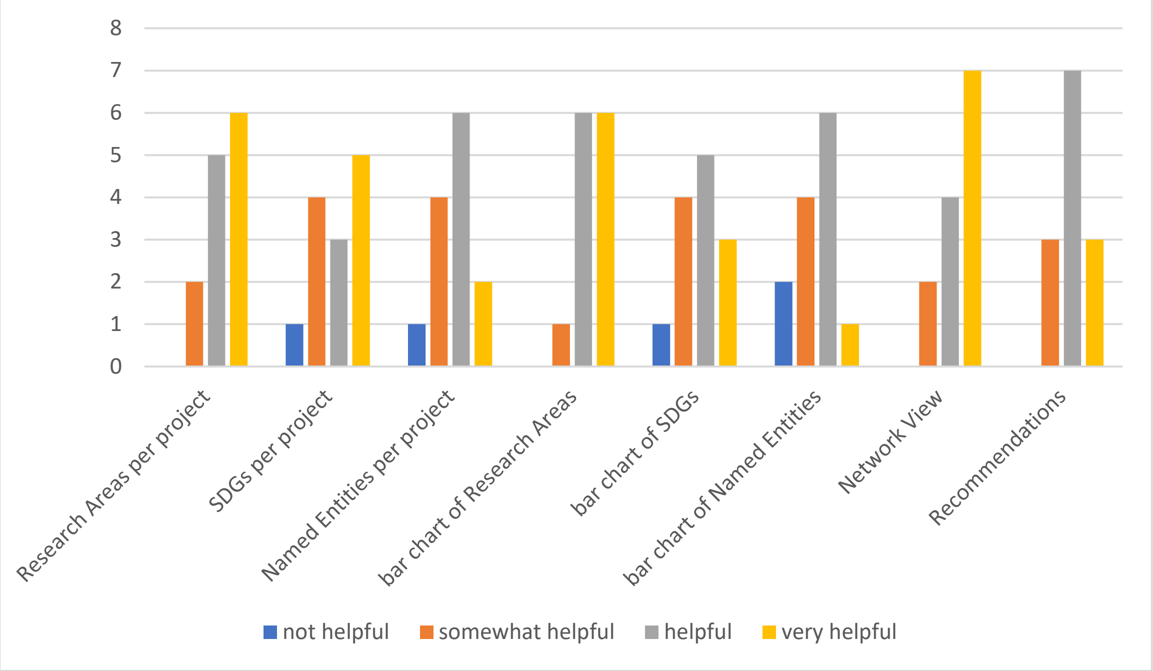


Figure 16: Helpfulness ratings per functionality

For the results per project, we asked separately for research areas, SDGs and named entities. The research areas were rated *somewhat helpful* by two participants, *helpful* by five and *very helpful* by six, with a mean of *M*=2.31 (*SD*=0.75). The assigned SDGs are rated as *not helpful* by one participant, *somewhat helpful* by four, *helpful* by three and *very helpful* by five (*M*=1.92, *SD*=1.04). Named entities has the lowest mean value out of the three with *M*=1.69 (*SD*=0.85). This is based on one participant rating them as *not helpful*, four as *somewhat helpful*, six as *helpful* and two as *not helpful*. Still the mean value of 1.69 can be translated as *somewhat helpful* with a considerable tendency to *helpful*.

For the bar charts included in the dashboard - that show the top research areas, SDGs and named entities - we equally asked separately for each chart. The research area bar chart was rated as *somewhat helpful* by one participant, *helpful* by six and *very helpful* by six, leading to a higher mean than the research area results per project with *M*=2.38 (*SD*=0.65). The SDG bar chart was rated as *not helpful* by one participant, *somewhat helpful* by four, *helpful* by five and *very helpful* by three (*M*=1.77, *SD*=0.93). The named entities bar chart was rated as *not helpful* by two participants, *somewhat helpful* by four, *helpful* by six and *very helpful* by one. With a mean of *M*=1.46 (*SD*=0.88) this feature is rated as the least helpful, which can be translated as *somewhat helpful* with a slight tendency to *helpful*.

The network view is deemed *somewhat helpful* by two participants, *helpful* by four and *very helpful* by seven, making it the best rated feature with a mean of *M*=2.38 (*SD*=0.77). The recommendation is rated as *helpful* by seven participants and as *somewhat helpful* and *very helpful* by two each, giving it a mean of *M*=2 (*SD*=0.71).

The overall positive rating of helpfulness for the Workbench's features indicates a positive result regarding the utility of the Workbench. Looking back to the evaluation of the guided questions those show us that most of the participants were able to utilize the features of the Workbench in order to answer the given questions correctly.

Open ended questions. In the evaluative part of the questionnaire, we inquired whether the participant's expectations of this workbench were met, if something surprised them, what else they might have expected to see or what else they wished to see (cf. Annex 1, questions 18 and 19). One of the findings is, that the participants are interested in meta data about the workbench processing from other participants- such as how often results of a project were analysed and if modifications diverge. Furthermore, the participants were interested in additional information about the projects, e.g., about the type of involvement and tasks of citizen scientists. They also had ideas for additional analytics approaches. One participant requested additional tooltips and some improvement of the interface of the Workbench.

Final open discussion. In the final open discussion, in addition to an overall very positive feedback on the workshop, several participants indicated their interest in further using the toolset and workbench. One specific intended application would require dealing with German project descriptions. Here, our suggestion was not to convert the text analytics (in this case, ESA) but to use a source translation. Another participant showed an interest in using NER (Named Entity Recognition) to support the standardization of labels for organisations (here specifically NASA) and their related projects and institutions. It was reported that several databases showed inconsistencies in this respect. Another outcome was the invitation to report on the workbench and further aspects of analytics work in CS Track in a research colloquium at KIT Karlsruhe (Germany).

4. Conclusion

This deliverable has documented the (open source) software release underlying the Analytics Workbench, which is the main outcome of Task T3.2 in the CS Track project. The Analytics Workbench provides a toolset of analytics tools and methods based on the previous deliverable D3.1, with a strong focus on computational content analysis, especially text analytics and network extraction both from project descriptions as well as Twitter-based analyses. The set of tools and methods can be bundled in Analytics Workbench facilitates the interactive usage of the mentioned analytics methods to generate insights about CS activities.

A first workshop with stakeholders and potential users of the workbench has been conducted in month 24 of the project. Within this workshop, the usability and utility of the tools and the system as a whole have been evaluated. It turns out that most of the tools have been perceived as being (very) helpful and that the stakeholders see a significant potential regarding the generation of insights about citizen science projects in their scope or beyond.

During the next phase of the project, the Analytics Workbench will be further employed to support the analyses in CS Track and thus it will also serve as one ingredient of the triangulation approach conducted within WP4.

References

Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. *3rd International AAAI Conference on Weblogs and Social Media*. San José (CA), May 2009.

Mierswa, I. & Klinkenberg, R. (2018). *RapidMiner Studio*. RapidMiner, Inc., 12 Dec. 12, 2018. Source: rapidminer.com

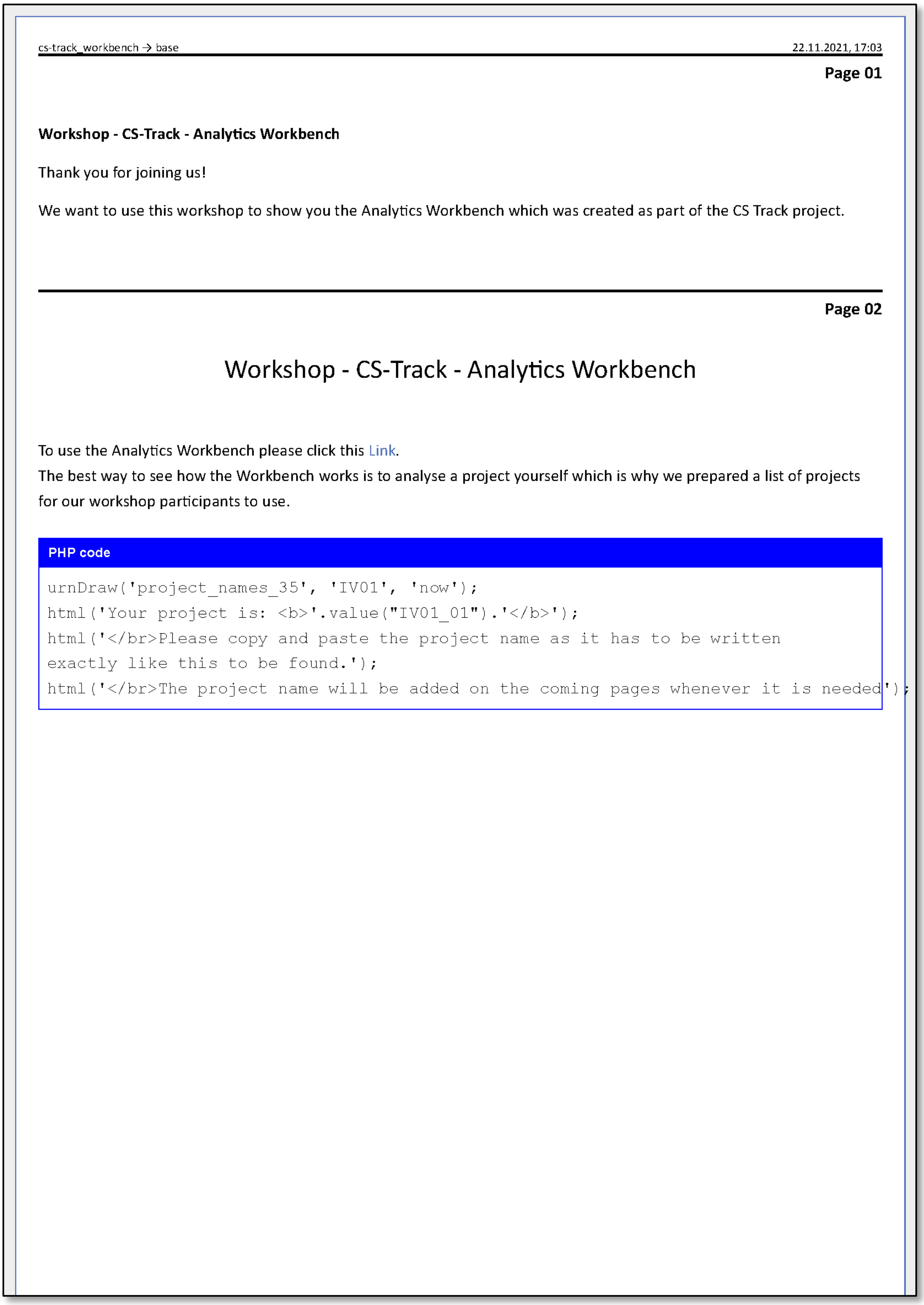
Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. & Euler, T. (2006). *YALE: Rapid Prototyping for Complex Data Mining Tasks.* Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006. 10.1145/1150402.1150531.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web*. (Technical Report Nr. 1999-66). Stanford InfoLab. doi: 10.1.1.31.1768

Schrepp, M., Hinderks, A. & Thomaschewski, J. (2017). Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). In: *IJIMAI 4 (6)*, 103–108.

Annexes

Annex 1: Questionnaire (Workshop)



cs-track_workbench → base 22.11.2021, 17:03

Page 01

Workshop - CS-Track - Analytics Workbench

Thank you for joining us!

We want to use this workshop to show you the Analytics Workbench which was created as part of the CS Track project.

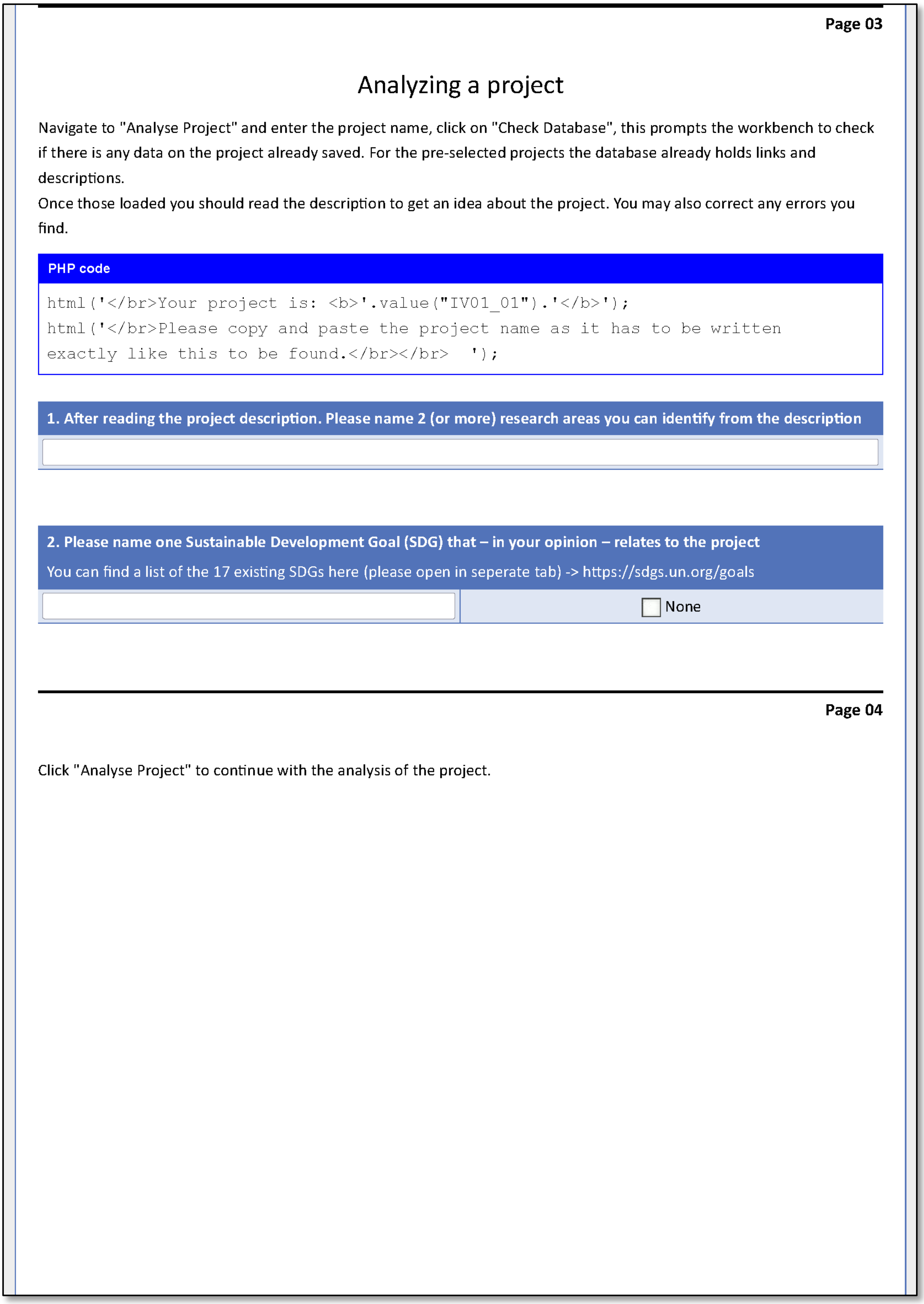
Page 02

Workshop - CS-Track - Analytics Workbench

To use the Analytics Workbench please click this Link.
The best way to see how the Workbench works is to analyse a project yourself which is why we prepared a list of projects for our workshop participants to use.

PHP code

```
urnDraw('project_names_35', 'IV01', 'now');
html('Your project is: <b>'.value("IV01_01").'</b>');
html('</br>Please copy and paste the project name as it has to be written
exactly like this to be found.');
html('</br>The project name will be added on the coming pages whenever it is needed');
```



Page 03

Analyzing a project

Navigate to "Analyse Project" and enter the project name, click on "Check Database", this prompts the workbench to check if there is any data on the project already saved. For the pre-selected projects the database already holds links and descriptions.

Once those loaded you should read the description to get an idea about the project. You may also correct any errors you find.

PHP code

```
html('</br>Your project is: <b>'.value("IV01_01").'</b>');
html('</br>Please copy and paste the project name as it has to be written
exactly like this to be found.</br></br>  ');
```

1. After reading the project description. Please name 2 (or more) research areas you can identify from the description

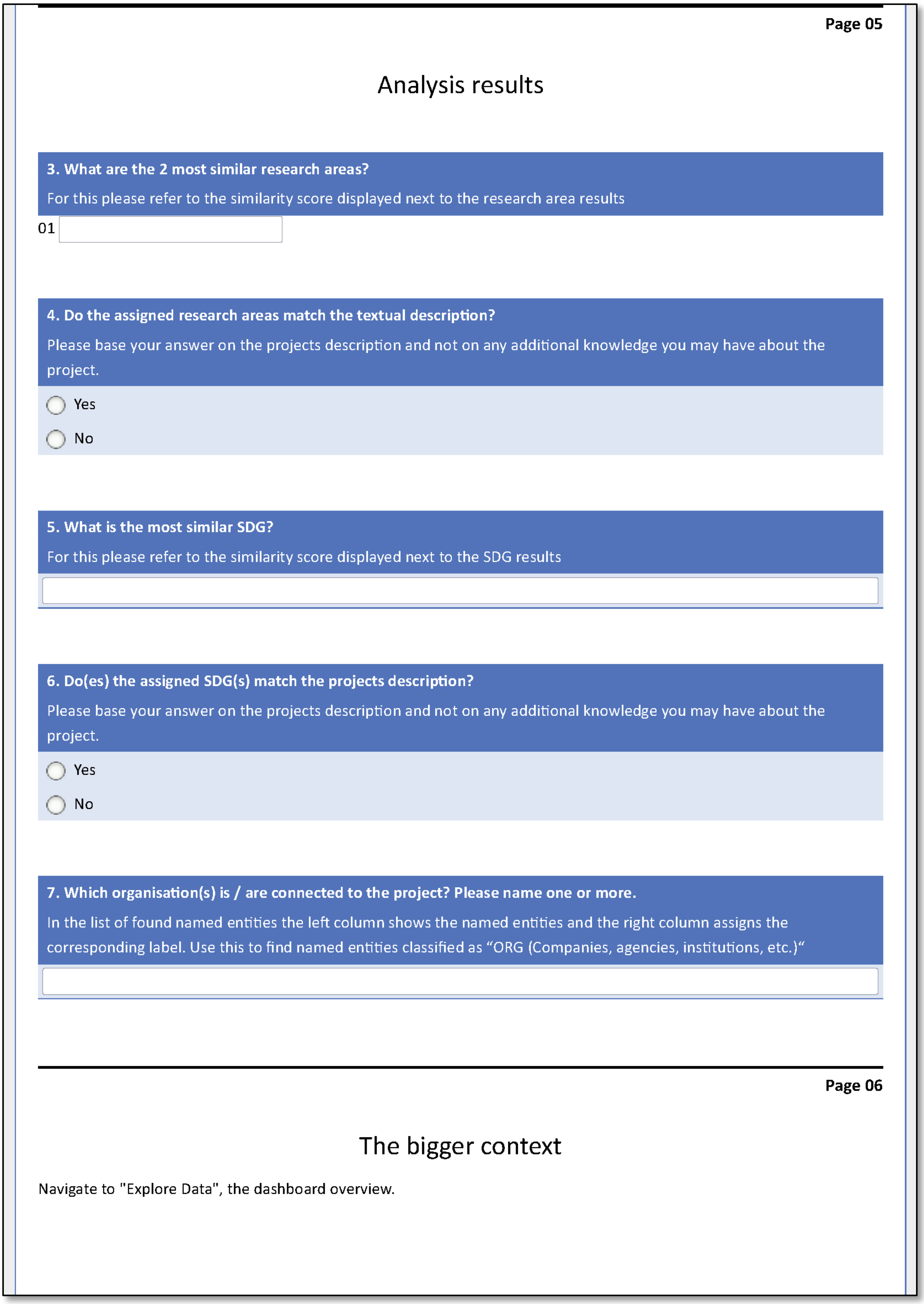
2. Please name one Sustainable Development Goal (SDG) that – in your opinion – relates to the project

You can find a list of the 17 existing SDGs here (please open in seperate tab) -> https://sdgs.un.org/goals

□ None

Page 04

Click "Analyse Project" to continue with the analysis of the project.



Page 05

Analysis results

3. What are the 2 most similar research areas?

For this please refer to the similarity score displayed next to the research area results

01

4. Do the assigned research areas match the textual description?

Please base your answer on the projects description and not on any additional knowledge you may have about the project.

○ Yes

○ No

5. What is the most similar SDG?

For this please refer to the similarity score displayed next to the SDG results

6. Do(es) the assigned SDG(s) match the projects description?

Please base your answer on the projects description and not on any additional knowledge you may have about the project.

○ Yes

○ No

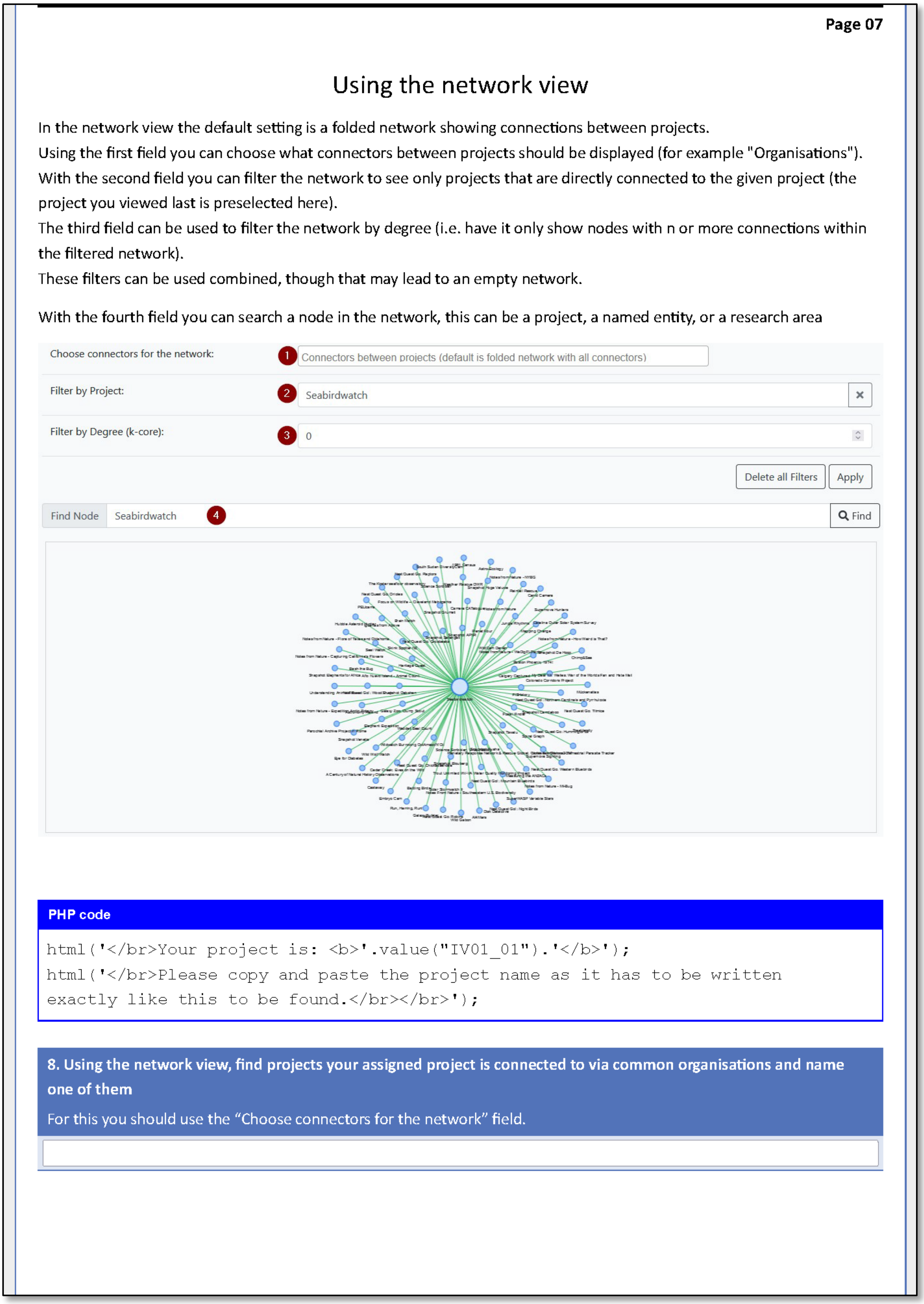
7. Which organisation(s) is / are connected to the project? Please name one or more.

In the list of found named entities the left column shows the named entities and the right column assigns the corresponding label. Use this to find named entities classified as “ORG (Companies, agencies, institutions, etc.)“

Page 06

The bigger context

Navigate to "Explore Data", the dashboard overview.



Page 07

Using the network view

In the network view the default setting is a folded network showing connections between projects.
Using the first field you can choose what connectors between projects should be displayed (for example "Organisations").
With the second field you can filter the network to see only projects that are directly connected to the given project (the project you viewed last is preselected here).
The third field can be used to filter the network by degree (i.e. have it only show nodes with n or more connections within the filtered network).
These filters can be used combined, though that may lead to an empty network.

With the fourth field you can search a node in the network, this can be a project, a named entity, or a research area

Choose connectors for the network: 1 Connectors between projects (default is folded network with all connectors)

Filter by Project: 2 Seabirdwatch

Filter by Degree (k-core): 3 0

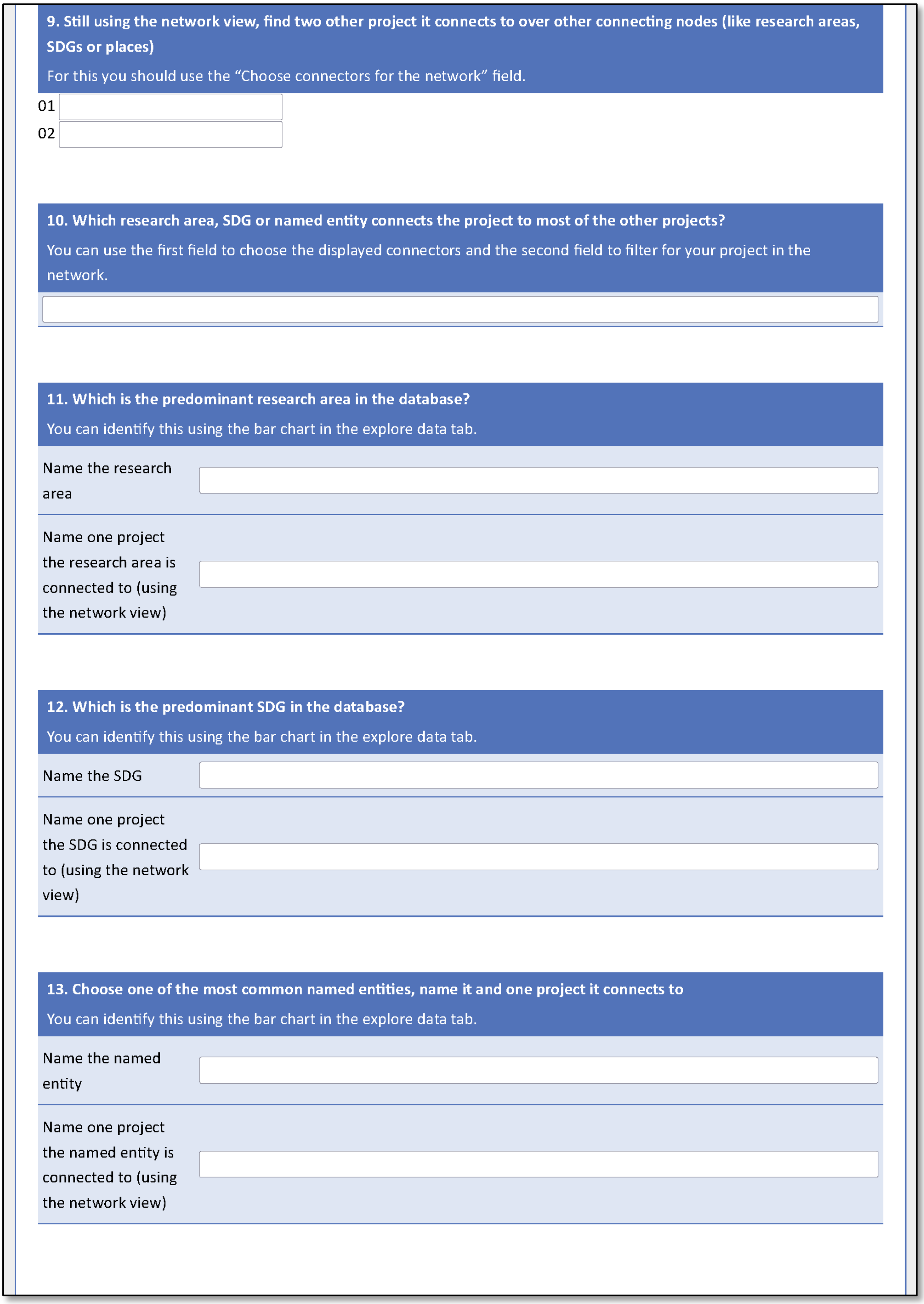
Find Node Seabirdwatch 4

PHP code

```
html('</br>Your project is: <b>'.value("IV01_01").'</b>');
html('</br>Please copy and paste the project name as it has to be written
exactly like this to be found.</br></br>');
```

8. Using the network view, find projects your assigned project is connected to via common organisations and name one of them

For this you should use the “Choose connectors for the network” field.



9. Still using the network view, find two other project it connects to over other connecting nodes (like research areas, SDGs or places)

For this you should use the "Choose connectors for the network" field.

01

02

10. Which research area, SDG or named entity connects the project to most of the other projects?

You can use the first field to choose the displayed connectors and the second field to filter for your project in the network.

11. Which is the predominant research area in the database?

You can identify this using the bar chart in the explore data tab.

Name the research area

Name one project the research area is connected to (using the network view)

12. Which is the predominant SDG in the database?

You can identify this using the bar chart in the explore data tab.

Name the SDG

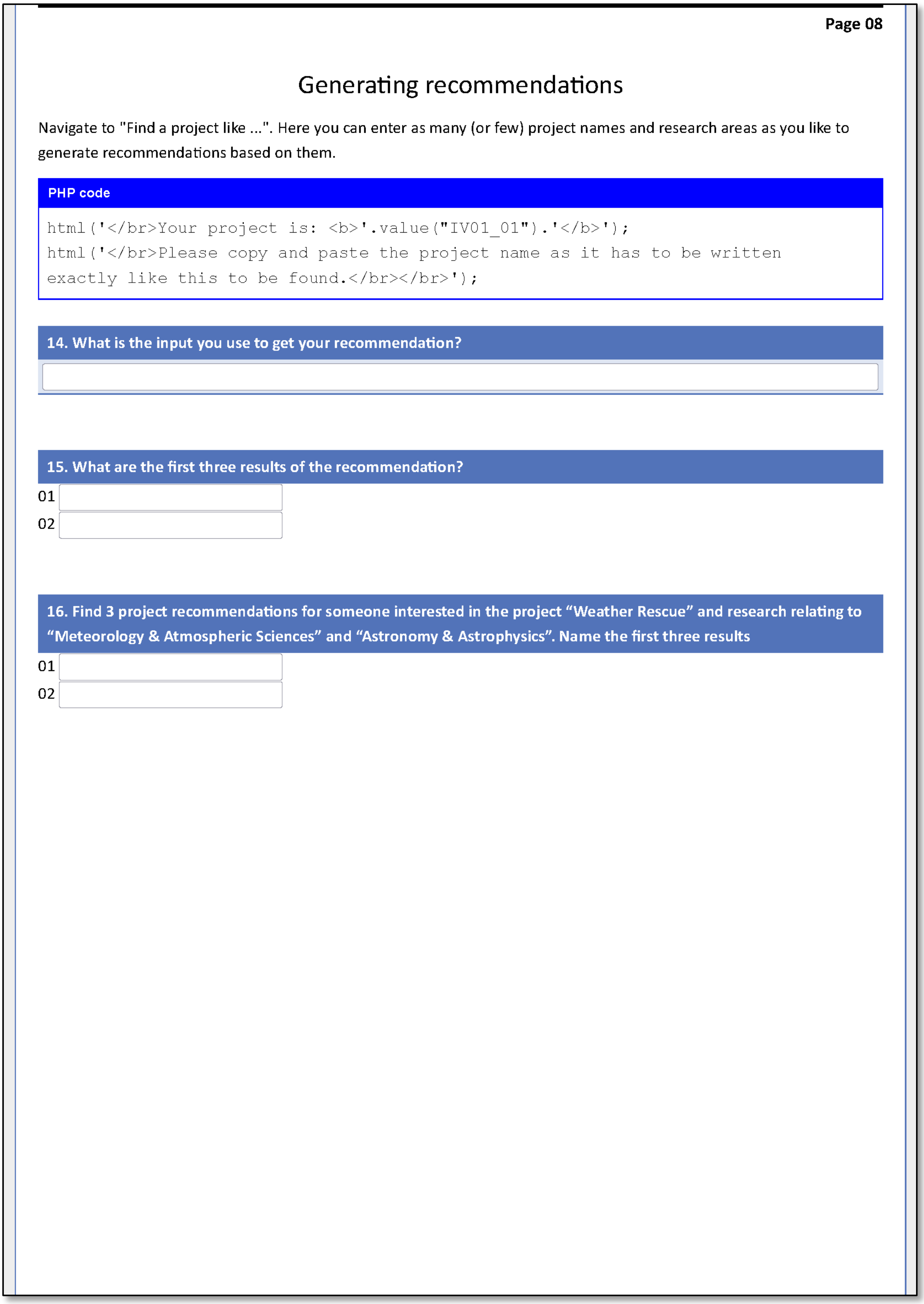
Name one project the SDG is connected to (using the network view)

13. Choose one of the most common named entities, name it and one project it connects to

You can identify this using the bar chart in the explore data tab.

Name the named entity

Name one project the named entity is connected to (using the network view)



Page 08

Generating recommendations

Navigate to "Find a project like ...". Here you can enter as many (or few) project names and research areas as you like to generate recommendations based on them.

PHP code

```
html('</br>Your project is: <b>'.value("IV01_01").'</b>');
html('</br>Please copy and paste the project name as it has to be written
exactly like this to be found.</br></br>');
```

14. What is the input you use to get your recommendation?

15. What are the first three results of the recommendation?

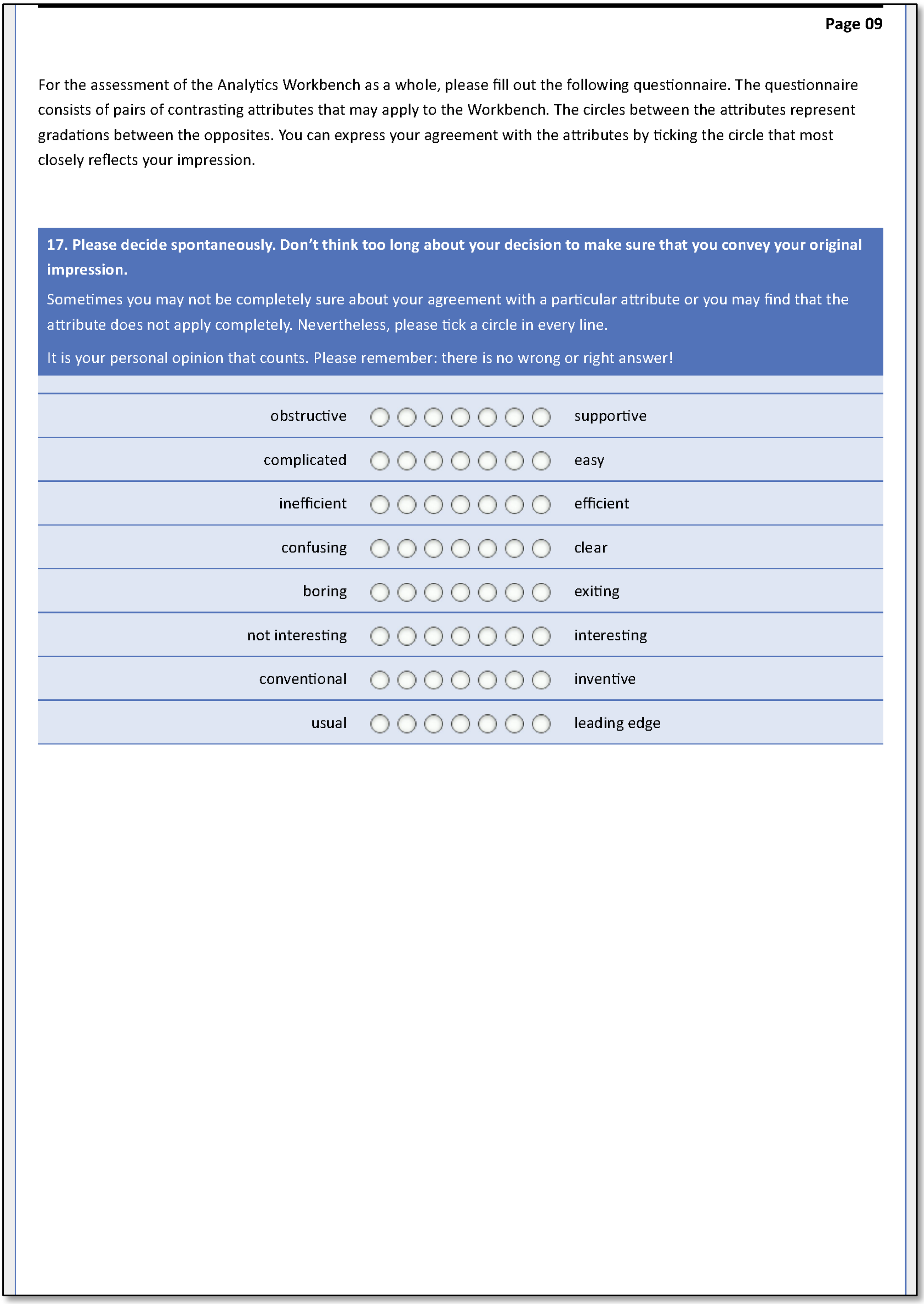
01

02

16. Find 3 project recommendations for someone interested in the project "Weather Rescue" and research relating to "Meteorology & Atmospheric Sciences" and "Astronomy & Astrophysics". Name the first three results

01

02



Page 09

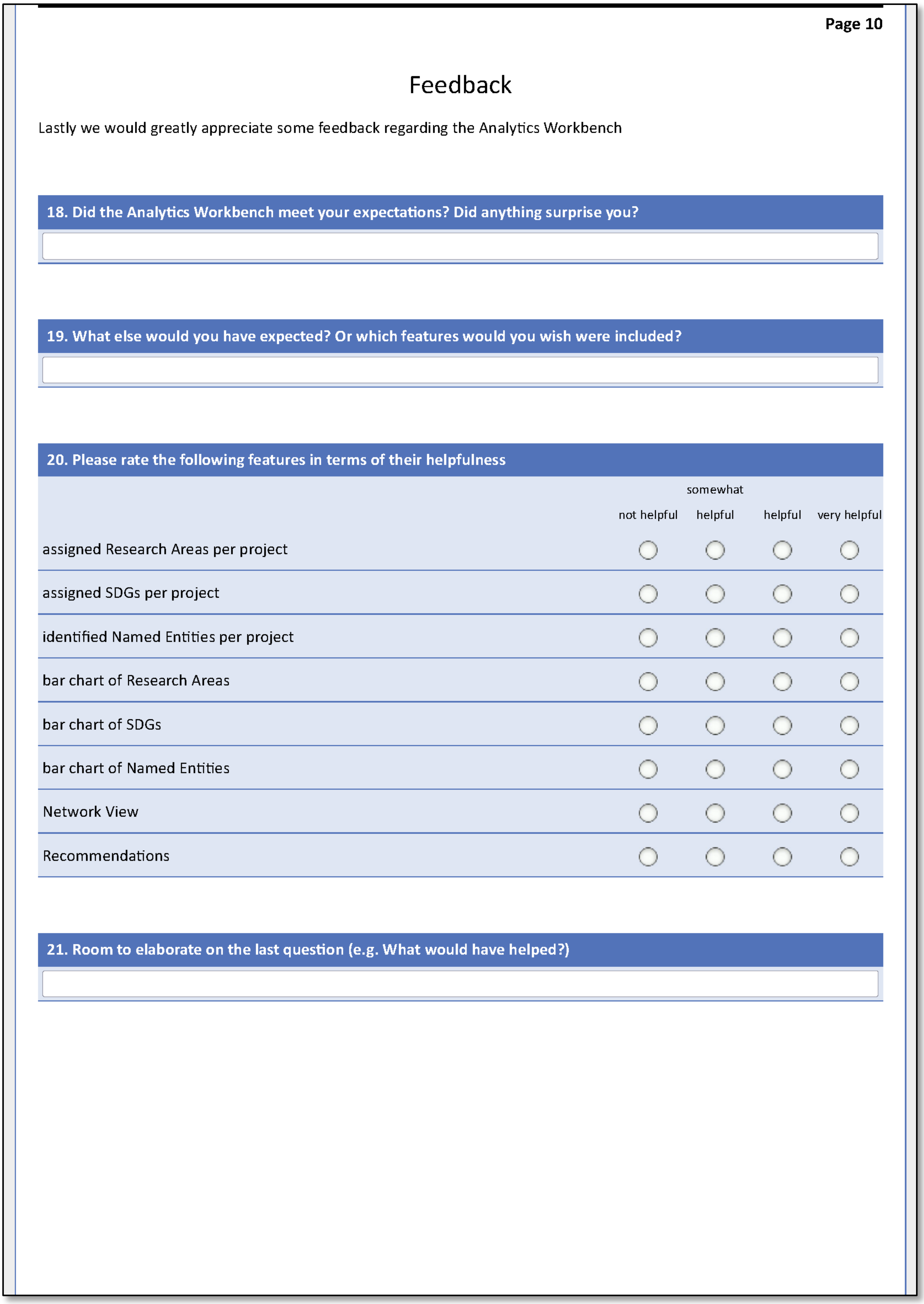
For the assessment of the Analytics Workbench as a whole, please fill out the following questionnaire. The questionnaire consists of pairs of contrasting attributes that may apply to the Workbench. The circles between the attributes represent gradations between the opposites. You can express your agreement with the attributes by ticking the circle that most closely reflects your impression.

17. Please decide spontaneously. Don’t think too long about your decision to make sure that you convey your original impression.

Sometimes you may not be completely sure about your agreement with a particular attribute or you may find that the attribute does not apply completely. Nevertheless, please tick a circle in every line.

It is your personal opinion that counts. Please remember: there is no wrong or right answer!

obstructive	○	○	○	○	○	○	○	supportive
complicated	○	○	○	○	○	○	○	easy
inefficient	○	○	○	○	○	○	○	efficient
confusing	○	○	○	○	○	○	○	clear
boring	○	○	○	○	○	○	○	exiting
not interesting	○	○	○	○	○	○	○	interesting
conventional	○	○	○	○	○	○	○	inventive
usual	○	○	○	○	○	○	○	leading edge



Page 10

Feedback

Lastly we would greatly appreciate some feedback regarding the Analytics Workbench

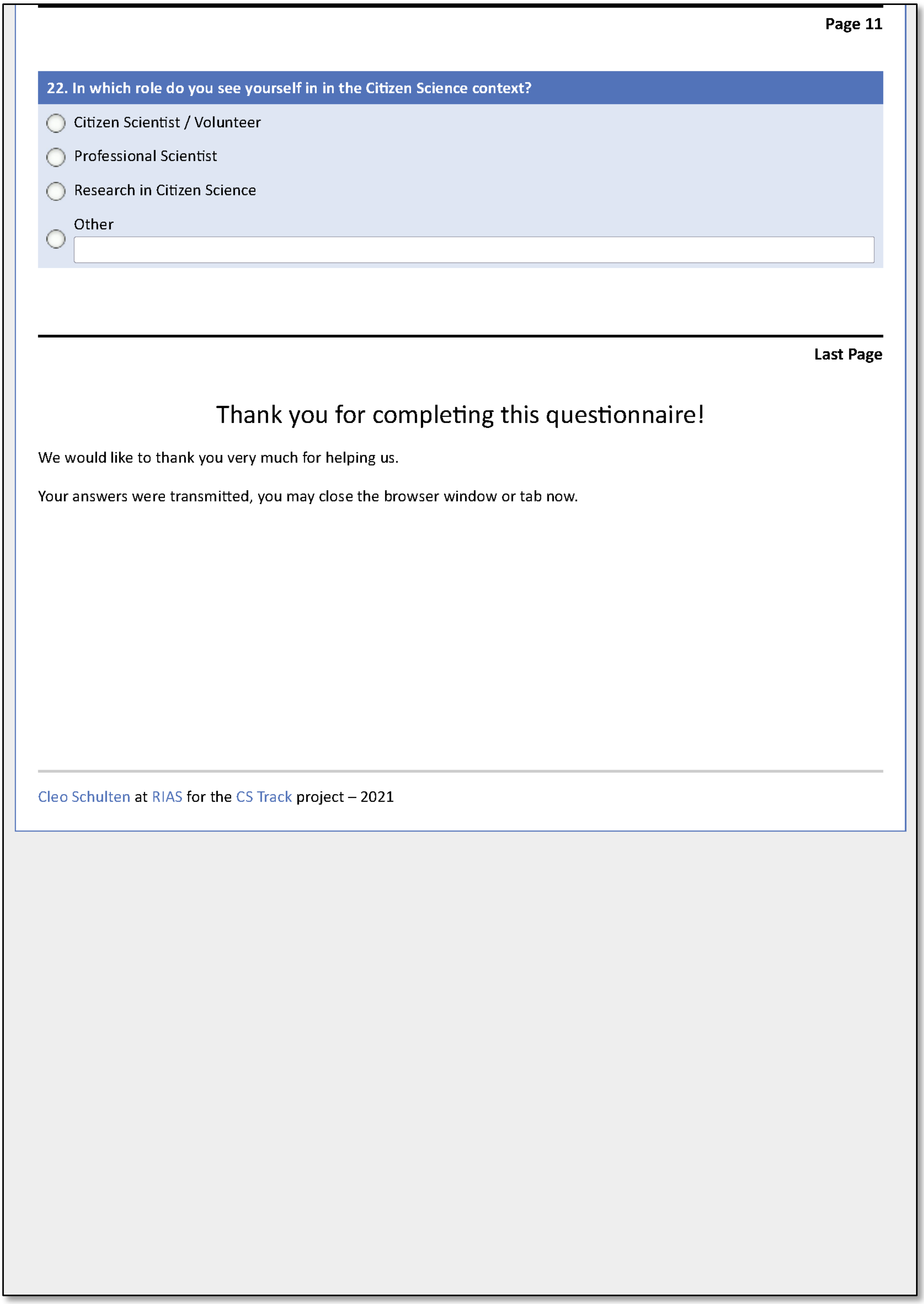
18. Did the Analytics Workbench meet your expectations? Did anything surprise you?

19. What else would you have expected? Or which features would you wish were included?

20. Please rate the following features in terms of their helpfulness

	not helpful	somewhat helpful	helpful	very helpful
assigned Research Areas per project	○	○	○	○
assigned SDGs per project	○	○	○	○
identified Named Entities per project	○	○	○	○
bar chart of Research Areas	○	○	○	○
bar chart of SDGs	○	○	○	○
bar chart of Named Entities	○	○	○	○
Network View	○	○	○	○
Recommendations	○	○	○	○

21. Room to elaborate on the last question (e.g. What would have helped?)



Page 11

22. In which role do you see yourself in in the Citizen Science context?

○ Citizen Scientist / Volunteer

○ Professional Scientist

○ Research in Citizen Science

○ Other ______________________

Last Page

Thank you for completing this questionnaire!

We would like to thank you very much for helping us.

Your answers were transmitted, you may close the browser window or tab now.

Cleo Schulten at RIAS for the CS Track project – 2021

Annex 2: Software licenses for third-party libraries used

Name	License
alabaster	BSD License
Babel	BSD-3-Clause
bertopic	MIT License
blis	MIT License (BSD)
Brotli	MIT License
catalogue	MIT License
certifi	Mozilla Public License 2.0 (MPL 2.0) (MPL-2.0)
charset-normalizer	MIT License (MIT)
click	BSD-3-Clause
cmake	BSD-3-Clause
cycler	BSD License (BSD)
cymem	MIT License (MIT)
Cython	Apache Software License (Apache)
dash	MIT License (MIT)
dash-bootstrap-components	Apache Software License (Apache Software License)
dash-core-components	MIT License
dash-html-components	??
dash-table	MIT
docutils	BSD License, GNU General Public License (GPL), Python Software Foundation License, Public Domain (public domain, Python, 2-Clause BSD, GPL 3 (see COPYING.txt))
emoji	BSD License (New BSD)
express	MIT License
filelock	Public Domain (Unlicense)
Flask	BSD-3-Clause
Flask-Compress	MIT License (MIT)
future	MIT License (MIT)
gensim	GNU LGPLv2.1 license
hdbscan	BSD-3-Clause
huggingface-hub	Apache Software License (Apache)

Name	License
idna	BSD-3-Clause
imagesize	MIT License (MIT)
itsdangerous	BSD-3-Clause
Jinja2	BSD-3-Clause
joblib	BSD License
kiwisolver	BSD License
llvmlite	BSD License
MarkupSafe	BSD-3-Clause
matplotlib	Python Software Foundation License (PSF)
mercury-parser	Apache License, Version 2.0 MIT license
mlxtend	BSD-3-Clause
murmurhash	MIT License (MIT)
networkit	MIT License (MIT)
NetworkX	BSD-3-Clause
nltk	Apache License Version 2.0
numba	BSD License
Numpy	BSD-3-Clause
packaging	Apache Software License, BSD License (BSD-2-Clause or Apache-2.0)
pandas	BSD-3-Clause
pathy	Apache Software License (Apache 2.0)
patsy	BSD License (2-clause BSD)
Pillow	Historical Permission Notice and Disclaimer (HPND) (HPND)
plotly	MIT
polyglot	GNU General Public License v3 or later (GPLv3+) (GPLv3)
preshed	MIT
pydantic	MIT License (MIT)
Pygments	BSD License (BSD License)
PyMongo	Apache Software License (Apache License, Version 2.0)
PyMySQL	MIT License

Name	License
pynndescent	OSI Approved (BSD)
pyparsing	MIT License (MIT)
python-dateutil	Apache Software License, BSD License (Dual License)
python-dotenv	BSD License (BSD-3-Clause)
pytz	MIT License (MIT)
pyvis	BSD-3-Clause
PyYAML	MIT License (MIT)
regex	Apache Software License (Apache Software License)
requests	Apache Software License (Apache 2.0)
retrying	Apache Software License (Apache 2.0)
sacremoses	MIT License (MIT)
scikit-learn	BSD-3-Clause
scipy	BSD License (BSD)
seaborn	BSD-3-Clause
sentencepiece	Apache Software License (Apache)
sentence-transformers	Apache Software License (Apache License 2.0)
six	MIT License (MIT)
smart-open	MIT License (MIT)
snowballstemmer	BSD-3-Clause
spacy	MIT License
spacy	MIT License
spacy-legacy	MIT License (MIT)
Sphinx	BSD License (BSD)
sphinxcontrib-applehelp	BSD License (BSD)
sphinxcontrib-devhelp	BSD License (BSD)
sphinxcontrib-htmlhelp	BSD License (BSD)
sphinxcontrib-jsmath	BSD License (BSD)
sphinxcontrib-qthelp	BSD License (BSD)
sphinxcontrib-serializinghtml	BSD License (BSD)
srsly	MIT License (MIT)

Name	License
statsmodels	BSD License (BSD)
thinc	MIT License (MIT)
threadpoolctl	BSD-3-Clause
tokenizers	Apache Software License (Apache License 2.0)
torch	BSD License (BSD-3)
torchvision	BSD
tqdm	MIT License, Mozilla Public License 2.0 (MPL 2.0) (MPLv2.0, MIT Licences)
transformers	Apache Software License (Apache)
typer	MIT License
typing-extensions	Python Software Foundation License
umap-learn	OSI Approved (BSD)
urllib3	MIT License (MIT)
vaderSentiment	MIT License
wasabi	MIT
webweb	GNU General Public License v3 or later (GPLv3+)
Werkzeug	BSD-3-Clause
wordcloud	MIT
xlrd	BSD-3-Clause
spaCy English language model	MIT
Wikipedia database dumps	CC0 1.0 Universal (CC0 1.0) Public Domain Dedication