CS TRACK
Investigating Citizen Science

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

Specification of Web-based Analytics Methods and Tools D3.1

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

Title of project	CS Track
Full title of project	Expanding our knowledge on Citizen Science through analytics and analysis
Title of this document	Specification of Web-based Analytics Methods and Tools
Number of this document	D3.1
Dissemination level	Public
Due date	August 1st, 2020
Actual delivery	August 7, 2020
Versioning history	Previous versions elaborated over the last two months are captured in a shared online document.
Authors	Laia Albó (UPF), Miriam Calvera Isabal (UPF), Sven Manske (RIAS), David Roldán Álvarez (URJC), H. Ulrich Hoppe (RIAS) – editor Additional contributions by Davinia Hernández-Leo (UPF), Nils Malzahn (RIAS), Raul Drachman (MOFET), Julia Lorke (WID), Sally Reynolds (ATIT)
Executive summary	One of the central characteristics of the CS Track approach to investigating citizen science projects and activities is to focus on underlying types of participation, collaboration and knowledge creation. This perspective comprises semantic and pragmatic aspects of scientific discourse as well as more structural analyses that focus on connections and networking inside the projects, between different projects as well as the information flow between CS projects and other institutions in society (e.g. public media or education). Many relevant activities are materialised in online digital media such as web pages, forums or contributions to social media channels. WP3 aims at extracting and further processing such pieces of information from web and social media sources. For doing so, the consortium draws on experience with a variety of representations and processing techniques with emphasis on text mining and semantic modelling as well as (social) network analysis. This deliverable specifies the WP3 approach in terms of the selection and description or pertinent methods and their relevance in the citizen science context. In addition to the general rationale, this is illustrated by an example application of network analysis techniques to a Zooniverse case. The final section explains the intended usage of the methodology regarding the interplay with other work packages and information provided to the outside.

Table of Contents

Section 1: Orientation and General Approach

A cornerstone in CS Track's approach to investigating how citizen science (CS) activities develop and work is the use of computational analysis techniques to characterize and analyse these activities in terms of their interaction among each other, with society and with "official" science using digital sources and traces. The activities to be analysed are selected based on the collection provided by WP 2. The results of the analysis will be fed into the broader analysis (WP 4) that also includes standard approaches from the field of social and organizational studies. An important field of reference from which corresponding techniques are borrowed and adapted is the study of collaboration in educational and organizational settings.

The computational methods envisaged in this context have been developed in the research fields of Social Network Analysis or SNA (Wasserman & Faust, 1994; Borgatti et al., 2009) and Learning Analytics (Shum & Ferguson, 2012; Hoppe, 2017). According to Hoppe (2017), we can distinguish three fundamentally different types of computational approaches to analysing learning and knowledge building communities based on their activity traces and emerging products. This "trinity of methods" includes analytics of 1) *network structures* including actor–actor (social) networks but also actor–artefact networks, 2) *processes* using methods of sequence analysis, and 3) *content* using text mining or other techniques of artefact analysis. In this context, the actors are participants of learning communities or communities of practice and the artefacts are products generated by these communities (e.g., an astronomy club may produce photographs and spectra based on their own observations and make these available on a website). Regarding this triangle of analytics approaches (see Fig. 1), we would particularly focus on network structures (1) and content (3). Process-centric analytics methods (2) would typically require fine-grained activity logs, e.g. from work activities in the projects themselves. We cannot expect to have such data for our objects of study. This rules out the possibility of applying process analysis techniques. However, both network content analysis can and will be used in such a way as to account also for changes over time.

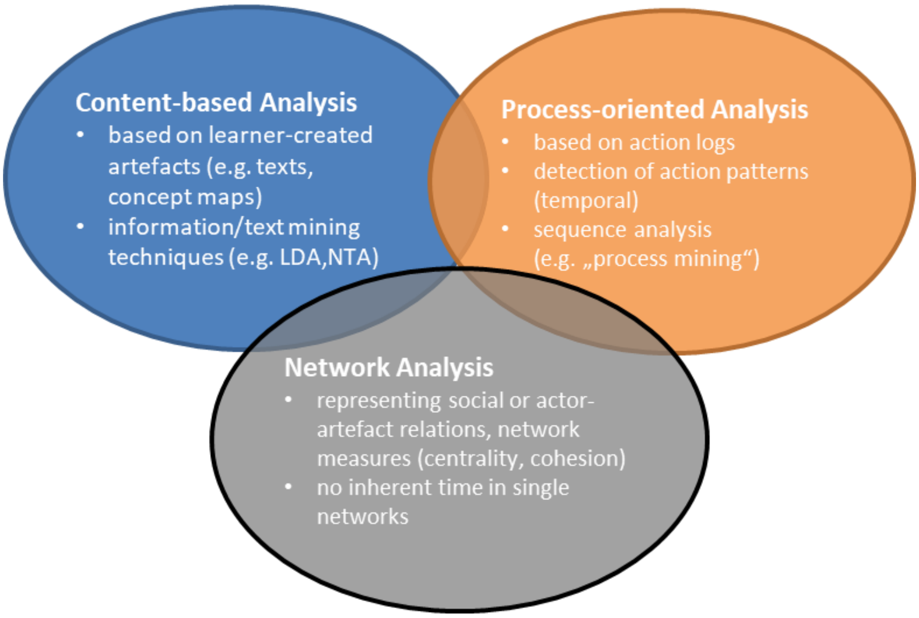

Content-based Analysis
• based on learner-created artefacts (e.g. texts, concept maps)
• information/text mining techniques (e.g. LDA,NTA)
Process-oriented Analysis
• based on action logs
• detection of action patterns (temporal)
• sequence analysis (e.g. „process mining")
Network Analysis
• representing social or actor-artefact relations, network measures (centrality, cohesion)
• no inherent time in single networks

Figure 1: The "trinity" of computational analysis methods according to Hoppe (2017)

Content analyses will be based on textual data retrieved from the public pages (web, social media) of specific projects/initiatives that we want to investigate. We are technically prepared also to include videos in this analysis. The envisaged outcome would in first place be a characterization of the given object of study (CS project) in terms of topics or themes. Certain analysis techniques will also allow us to extract relational structures in the form of semantic networks or concept maps. On a next level of analysis, such relational structures would be compared to similar constructs from “official” science activities which would allow us to pinpoint differences in the discourse structures between CS and official science in the same or similar domains. This will allow for judging scientific orientation and richness of the activities. Secondly, newly emerging relationships and topics may be considered as a source of information for revising and adapting science curricula in the light of new developments.

Network analysis techniques will be used to study the impact and inter-connection of projects in terms of information exchange through web-based and other media (“information diffusion”, cf. Hecking et al., 2019). If we see projects/initiatives as nodes in a network linked through information/ communication relations (inter-linking between websites, mentions of Y on pages created by X, Twitter connections through retweets of follower relationships etc.), we can apply network-based measures of relevance in terms of different types of centrality and we can identify certain levels of cohesion or inter-connectedness in larger group of projects. This gives us measures of “influence” that can be interpreted as indicators of impact or success.

In addition to these specific approaches addressing network structures and discourse characteristics, we will also use basic techniques to find relevant web pages and extract content. This involves techniques of web crawling and scraping as well as information mining from text sources. At the other end of the processing chain, we find information visualization techniques to display statistical findings, networks or to present data in a geo-mapping context. These general techniques will be explained in more detail in the next section.

Regarding the added value and benefits originating from web analytics, we would particularly mention the following targets:

- Automatic extraction of basic information from project web pages (named entity recognition, keyword extraction in combination with crawling and scraping of web pages);
- Assessment of disciplinary orientation as well as type of scientific discourse;
- Inter-relation and connection between different CS projects and activities;
- Assessment of public visibility of CS activities and projects in digital media.

The following sections will first elaborate on specific groups of methods, moving from general web-analytics over lexico-semantic mining to network analysis techniques. This exposition of methods is followed by a worked-out example that particularly relies on network analysis and finally wrapped up with concluding remarks and recommendations.

Section 2: General Web Analytics Methods and Tools

Digital media with user-generated content are broadly used by current citizen science (CS) projects not only for communication and information dissemination but also as working tools to produce results. Web pages, forums as well as social media channels are available as important sources of data for the study of CS projects and initiatives. The inspection and analysis of these data can provide insights into the ways in which professional researchers as well as volunteers participate in these projects and how institutions and people from different projects and countries are related to each other. This section is devoted to the different general web analytics methods and tools that can be used to extract data from web spaces and social media.

2.1 Current Extraction Tools

As of today, web scraping has been broadly applied in many different application fields. For example, it has been widely used by companies to keep in check the prices of their competitors. Since there are so many tools available, it is necessary to carry out a study to determine how data gathering and analysis will be done in CS Track. In this section we will present a selection of current data extraction tools, both complete tools and modules available for different programming languages.

2.1.1 Web data extraction

The automated gathering of data from the Internet is nearly as old as the Internet itself. Nowadays, this data gathering is known as "web scraping", which is a traditional technique to retrieve web content at scale. In theory, web scraping is the practice of gathering data through any means other than a program interacting with an API. In practice, it combines different programming techniques and technologies, such as data analysis and information security (Lawson, 2015).

As seen in Table 1, a multitude of configurable ready-to-use scraping tools exist (Glez-Peña et al., 2014). However, these scraping tools are not flexible enough to provide support for mining project data from different websites and do not provide the mechanisms necessary to adapt them to the needs that CS Track may have.

Table 1: Web scraping ready-to-use tools.

Name	Software type	Link
scrapingbot	Free/Commercial	https://www.scraping-bot.io/
Scraper API	Commercial	https://www.scraperapi.com/
Scrapinghub	Commercial	https://scrapinghub.com/?
Octoparse	Free/Commercial	https://www.octoparse.com/
Import.io	Commercial	https://import.io
Mozenda	Commercial	https://www.mozenda.com/
Fminer	Commercial	https://www.fminer.com

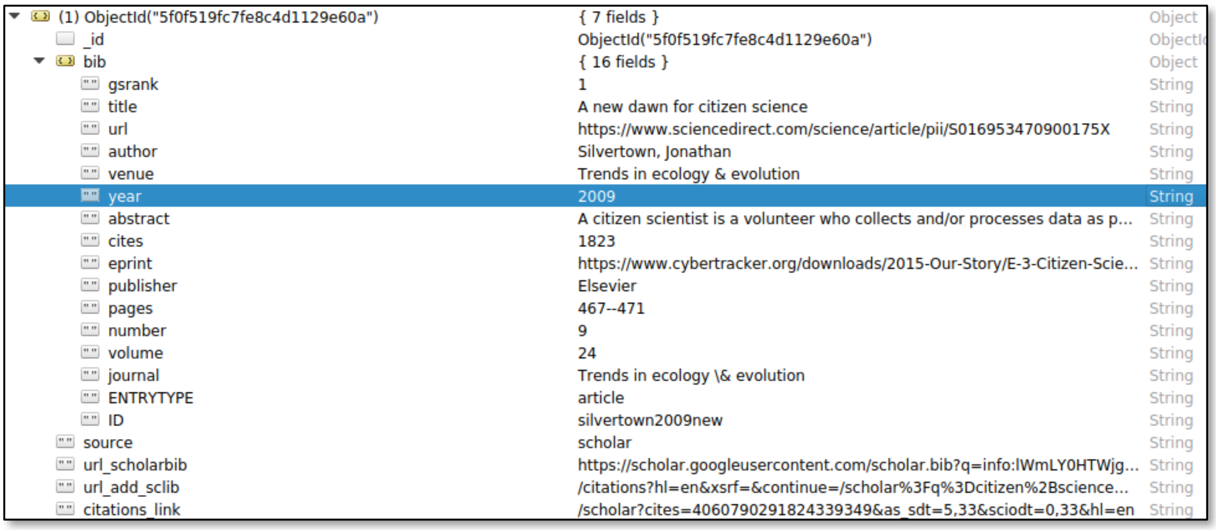
Therefore, in CS Track we will focus on the use of customizable frameworks to both gather the data and analyse it. In this regard, we can find several free programming frameworks (See table 2) that are fully customizable and that will let us achieve the proposed goals.

Table 2: Web scraping modules and frameworks.

Name	**Programming language**	**Type**
Scrapy	Python	Framework
MechanicalSoup	Python	Module
Jaunt	Java	Framework
Jauntium	Java	Framework
Storm Crawler	Java	Framework
Norconex	Several	Framework
Apify	NodeJs	Framework
Kimurai	Ruby	Framework
Colly	Golang	Framework
BeautifulSoup	Python	Module
Selenium	Several	Framework

After reviewing the tools, frameworks and modules available, we decided to base our data extraction and analysis methods in Python 3, since this language has been developed under an OSI-approved open source license, which makes it free to use and distribute. In addition, there is a full growth community that provides numerous third-party modules that make Python capable of interacting with most of the other languages and platforms. This is a key factor due to the interaction of WP3 with WP1, WP2, WP4 and WP5 and Python is suitable for providing tools, methods and data that is easily interchangeable with other Work Packages.

Figure 2 shows an example of data extracted by using Python3 and BeautifulSoup. We scraped Google Scholar to obtain articles related to citizen science projects so we could establish connections between researchers and institutions. In addition, information about the role of the authors could be gathered in order to evaluate if citizen scientists spend time publishing their results.



```
▼ (1) ObjectId("5f0f519fc7fe8c4d1129e60a")   { 7 fields }                                                          Object
      _id                                    ObjectId("5f0f519fc7fe8c4d1129e60a")                                  ObjectI
  ▼ bib                                      { 16 fields }                                                         Object
        gsrank                               1                                                                     String
        title                                A new dawn for citizen science                                        String
        url                                  https://www.sciencedirect.com/science/article/pii/S016953470900175X   String
        author                               Silvertown, Jonathan                                                  String
        venue                                Trends in ecology & evolution                                         String
        year                                 2009                                                                  String
        abstract                             A citizen scientist is a volunteer who collects and/or processes data as p...  String
        cites                                1823                                                                  String
        eprint                               https://www.cybertracker.org/downloads/2015-Our-Story/E-3-Citizen-Scie...  String
        publisher                            Elsevier                                                              String
        pages                                467--471                                                              String
        number                               9                                                                     String
        volume                               24                                                                    String
        journal                              Trends in ecology \& evolution                                        String
        ENTRYTYPE                            article                                                               String
        ID                                   silvertown2009new                                                     String
      source                                 scholar                                                               String
      url_scholarbib                         https://scholar.googleusercontent.com/scholar.bib?q=info:lWmLY0HTWjg...  String
      url_add_sclib                          /citations?hl=en&xsrf=&continue=/scholar%3Fq%3Dcitizen%2Bscience...   String
      citations_link                         /scholar?cites=4060790291824339349&as_sdt=5,33&sciodt=0,33&hl=en      String
```

Figure 2: Data extracted from Google Scholar

2.1.2 Data extraction from social networks

Within the range of social media, online social networks are web-based services that allow individuals to construct a public or semi-public profile within a bounded system, articulate a list of other users with whom they share a connection, and view and traverse their list of connections and those made by others within the system (Boyd & Ellison, 2008). Gathering data from social networks is particularly relevant for CS since it helps to understand how social relationships influence people's learning, thinking and acting. Accordingly, these are appropriate sources for understanding why and how people participate in CS activities. Social networks shared through the Internet may increase participation in CS by creating communities of practice that increase awareness of others' participation, quickly spread new ideas, and allow individuals to create or maintain relationships that influence each other's knowledge and behaviours (Triezenberg, Knuth, Yuan & Dickinson, 2012).

Twitter is becoming the preferred social network for data collection. In this network users can post short messages in a real-time manner, which may contain images, geographic locations, URL references and videos. In this sense, Twitter seems an appropriate source for querying and collecting large volumes of data in a short time. However, Twitter's API has a rate limitation that renders large data extraction difficult. In addition, you can only access tweets that are three weeks old after registering an enterprise API. To access old (and new) tweets without being rate limited, there are several webpages, such as https://brand24.com or https://www.trackmyhashtag.com/, which by paying a monthly fee allow you to track *hashtags.* In their most complete version, they also provide analytic tools. However, it is not that clear if they let the user download the tweets, so it could not be possible to perform analyses beyond those provided by the service.

In this area, the "Instituto de Ingeniería del Conocimiento" (Madrid, Spain) offers an online tool (Lynguo) that gathers news for a huge number of digital newspapers, blogs, and tweets in real time. With this tool the users have access to quantitative and qualitative analysis, providing visualizations, being able to track at any time the evolution of a conversation around the selected topic. In addition, Lynguo does not only work with *hashtags,* but it also allows the creation of custom searches to extract data about them. Furthermore, end-users can download and analyse raw datasets by themselves, which will allow the members of the project to create their own analytics and visualizations. In Figure 3 it is shown the number of tweets with the #citizenscience hashtag, dividing the mentions of profiles within the European Union and the total mentions.

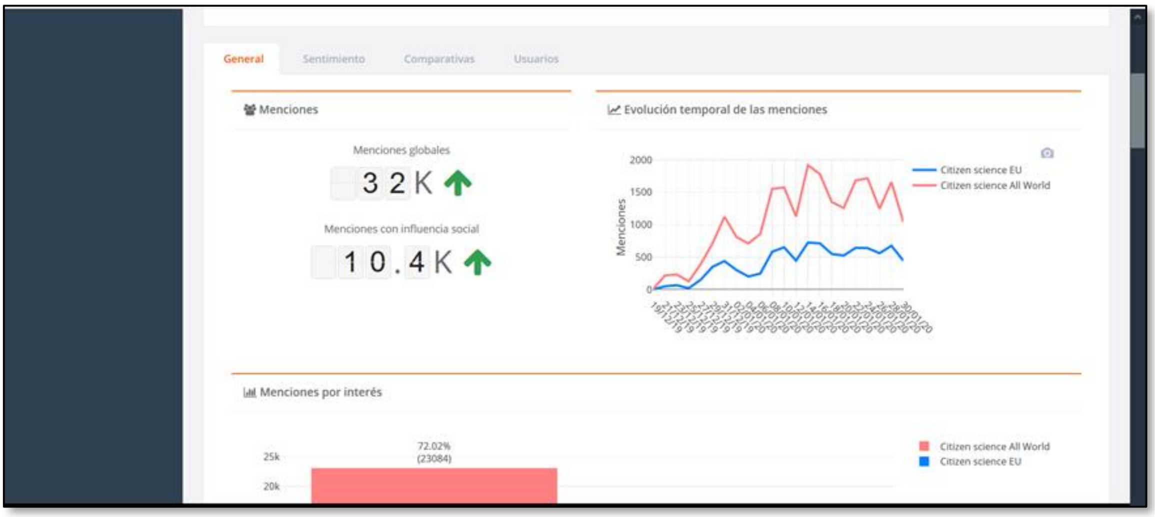


Figure 3: Lynguo

2.2 Web Analytics and Visualization

The extraction of data from web and social media sources is usually combined with different types of analytics and further on with graphical visualisations that help human stakeholders to survey and interpret the data. Lynguo is an example of a tool that supports such an information processing chain (cf. Fig. 3). In the sequel, we characterise different techniques used in such a workflow. More detailed and explanations regarding semantic and network-analytic methods will be elaborated in sections 3 and 4.

2.2.1 Text-based analysis techniques

Unstructured textual data can be very noisy. To achieve high quality web scraping, it is necessary to conduct data cleansing: spell checking, removing duplicates, finding and replacing text, removing spaces, fixing dates, consider missing data and incorrect data.

Once the data has been cleaned, there are several text analysis techniques that can be applied to it to extract meaningful information:

- Text analytics: It involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging, information extraction and predictive analytics.
- Natural language processing (NLP): It is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and human languages. Specifically, it is the computational process of extracting meaningful information from natural language input and/or producing natural language output. NLTK and spacy are the most popular NLP libraries for working with raw text in Python 3.
- Network (text) analysis: It is a method for encoding the relationships between words in a text and constructing a network of the linked words. It can be used to detect the structure of organizations. More details about network text analysis are presented in section 3.4.
- Opinion mining: This is an area of research that attempts to make automatic systems to determine human opinion from text written in natural language.

- Sentiment analysis: It refers to the application of NLP, computational linguistics and text analytics to identify and extract subjective information in source materials.

2.2.2 Data storage

WP2 has been working on the development of a database for the storage of data from the platforms and projects related to citizen science. This database is of great importance since it will allow everyone in the project to work easily with the data stored by any WP.

Due to the unstructured nature of data that we are dealing with in CS Track, WP2 decided to deploy a MongoDB system. MongoDB is a NoSQL database with high scalability which allows to store all the information generated through CS Track in a flexible way.

This database operates as a connector between different WPs. For instance, it allows WP3 to get citizen science projects and scrap through their webpages to gather information about the project's participants. Afterwards, WP3 would analyse the data and provide the corresponding visualizations to WP5 where, through the community platform, those visualizations will be shown to stakeholders, citizen scientists and other users that will read it.

2.2.3 Visualizations

Due to the large amount of data we will be dealing with, it is necessary to provide visualization tools that allow the user to digest the information in an easier way. Since the end users of our tools will have different profiles (stakeholders, citizen scientists, professional scientists…) different statistical visualizations will be provided, from simple charts and time series visualizations to complex graphs and networks. Several criteria for an information visualization have been proposed:

- The data are external, that is, they were not generated by an algorithm within the visualization program.
- The source data are not an image itself.
- The graphics must be readable.

In terms of intended aim two modes can be identified: exploratory and expository aim of use. If the visualization is used to explore the dataset, that is find new hypotheses, then the visualization should display the full dataset. However, if the visualization is going to be used to expose a certain issue, then only part of the data could be represented. The graphic must catch the reader's attention and it should facilitate reading of the data and enable the detection of underlying patterns and trends.

One of the main goals of CS Track is the observation and monitoring of citizen science activities and mirroring back the analytic findings to the initiative and a broader public. Considering the observational mode that this project takes, it is a key factor to create readable visualizations to shed light on the relationships around citizen science projects. In this regard, there is a need to analyse the different types of tools that will help CS Track to provide adequate visualizations. Since one of the ways of communicating the results is going to be through the community platform developed in WP5, we will work with online tools and libraries that can be used in a webserver environment to deploy the visualizations.

We considered to use an existing tool like Cognos Analytics[1], Power BI[2], MongoDB Atlas[3] and Metabase[4]. All these tools serve to create interactive dashboards through which to display the information and which are usually automatically refreshed with the information that is added to the

[1] IBM, Cognos Analytics (2019). https://www.ibm.com/products/cognos-analytics. Accessed: 2020-08-06.

[2] Microsoft, Power BI (2020). https://powerbi.microsoft.com/en-us/. Accessed: 2020-08-06.

[3] MongoDB, Inc., MongoDB Atlas (2020). https://www.mongodb.com/cloud/atlas. Accessed: 2020-08-06.

[4] Metabase, Metabase (2020). https://www.metabase.com/. Accessed: 2020-08-06.

database. Cognos Analytics and Power BI are paid tools that can connect to a large number of databases to provide the visualizations. They also have desktop versions which could also be used if you do not want to depend on an Internet connection. Figure 4 shows an example of the Cognos Analytics dashboard, where we can see different graphs with sample data showing current projects per country in a map, the role of the participants within a project or a tag cloud to represent how many CS projects each country has.

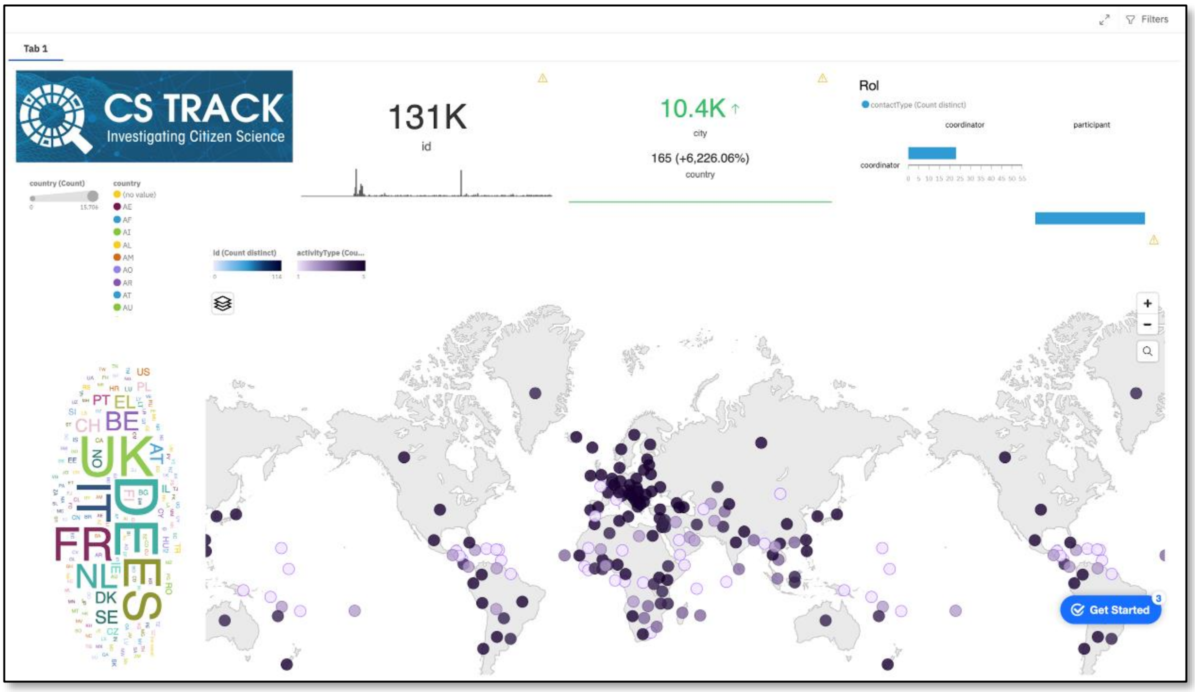


Figure 4: Cognos Analytics dashboard.

Figure 5 shows an example of Power BI dashboard. In this dashboard there are some graphs that represent the number of articles about citizen science written in different departments of URJC. There is also a pie chart representing the mean duration of projects and several counts of projects mentioned in publications from two different projects.

On the other hand, MongoDB Atlas is a free of charge tool focused on MongoDB databases. It allows the creation of a cluster containing a MongoDB database with its administration tools and access to the MongoDB Charts application, which allows the creation of dashboards for the visualization of the information contained in the said database. However, a premium account is needed in order to connect the database with other visualization tools. In addition, the MongoDB Charts version provided in MongoDB Atlas only allows the visualization of databases created in the cluster. However, MongoDB charts can be installed as a webservice through docker in any machine, which enables to be connected to any MongoDB database installed anywhere, maintaining the functionalities of its cloud version. In Figure 6 we show some examples of the types of graphs MongoDB Charts can generate. At this point, is relevant to mention that, in general, most of the tools found provide the same type of visualizations with small differences in the presentation styles.

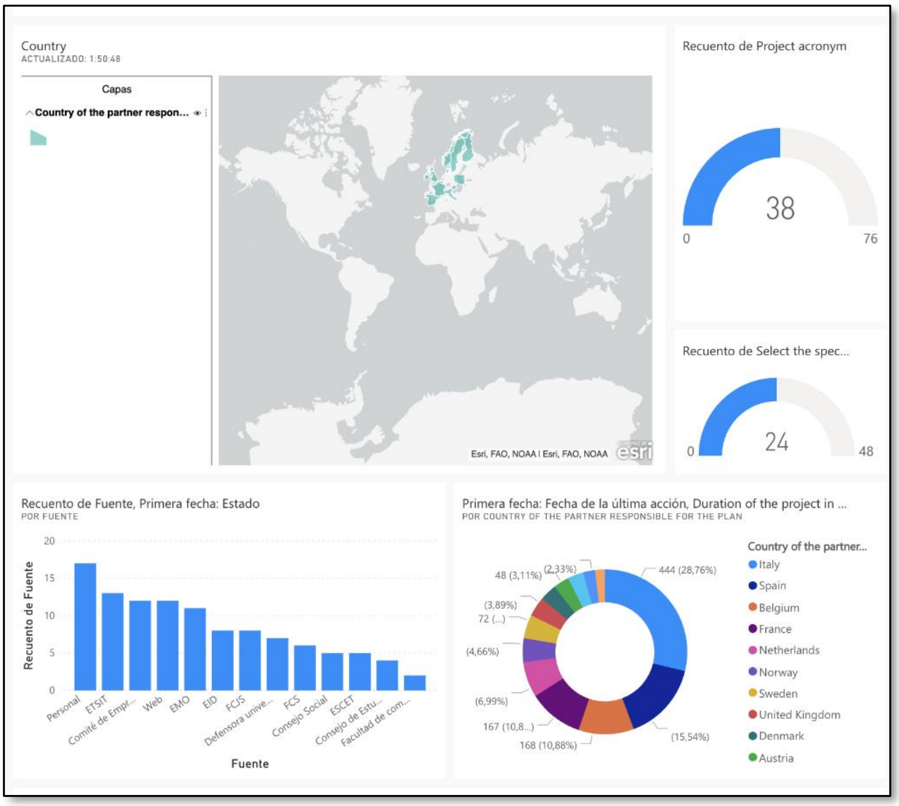
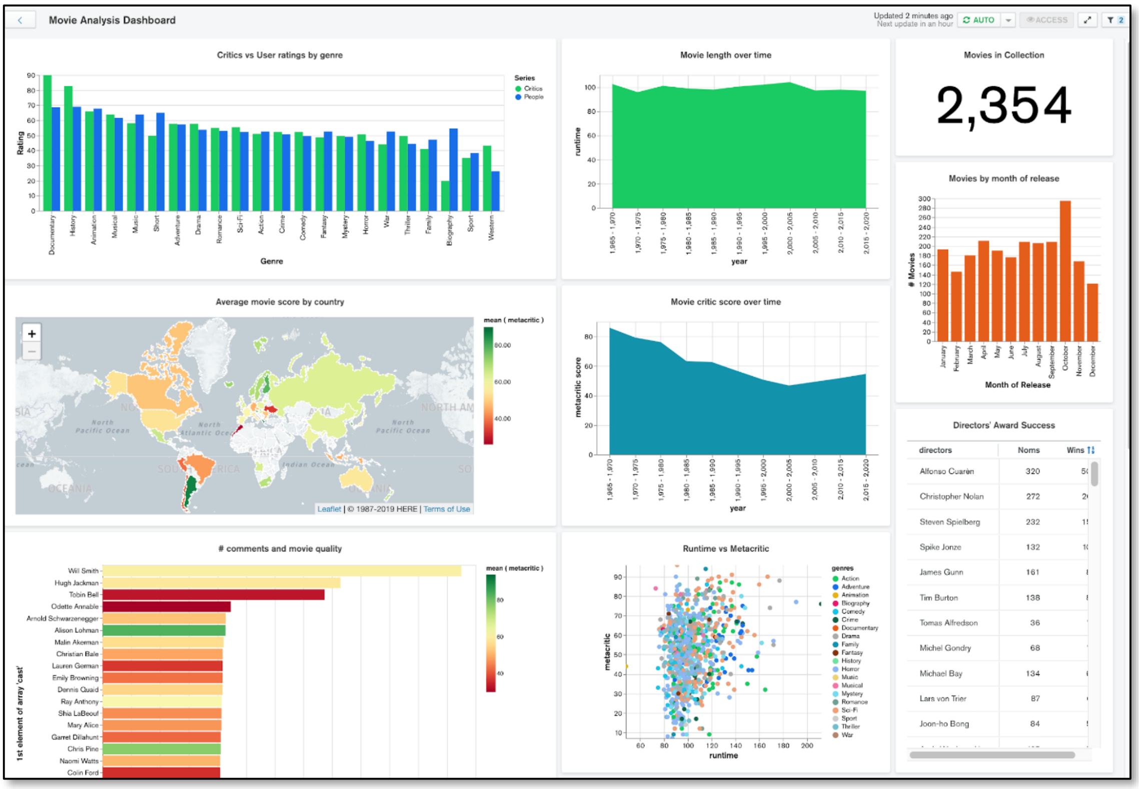


Figure 5: Power BI dashboard.



Directors' Award Success

directors	Noms	Wins
Alfonso Cuarén	320	[illegible]
Christopher Nolan	272	[illegible]
Steven Spielberg	232	[illegible]
Spike Jonze	132	[illegible]
James Gunn	161	[illegible]
Tim Burton	138	[illegible]
Tomas Alfredson	36	[illegible]
Michel Gondry	68	[illegible]
Michael Bay	134	[illegible]
Lars von Trier	87	[illegible]
Joon-ho Bong	84	[illegible]

Figure 6: MongoDB Charts

Lastly, Metabase is another tool that provides service for many of the databases available in the market (as Power BI and Cognos Analytics) and it is free of charge (as MongoDB Atlas). It can be installed through docker but also by using an installation file. Its flexibility and its combination of the advantages of the tools mentioned above, makes it a suitable choice at this stage of the project. Even if the type of database in which the information is stored was to change, there would be no need to change this visualization tool. In Figure 7 we show an example of the visualizations we can create with Metabase. To the left, we see a pie chart that represents the number of CS platforms available in each country. To the right, we see a bar graph with the number of CS projects of each country.

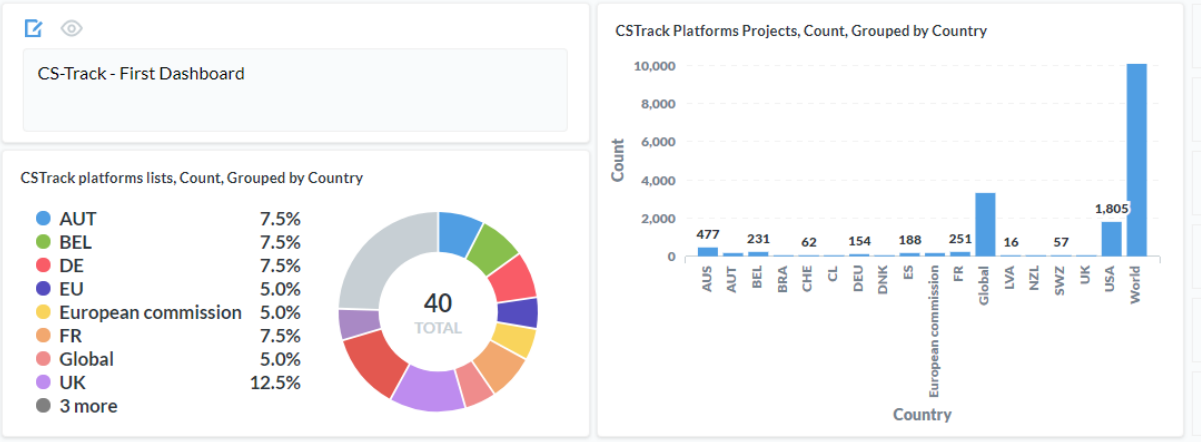


Figure 7: Metabase

2.3 Summary and Rationale

The examples of general web analytics and visualisation tools presented above reflect the requirements and preferences as a result of discussions between WP3, WP2 and WP 5. A dedicated task force from these WPs has particularly explored visualisation techniques and tools in connection with the set of CS projects collected in the WP2m database. The general tools serve as analysis components that provide statistical analyses, basic network and text mining in connection with different types of visualization components including diagrammatic representations and geo-mapping. This set of tools will be made available for all CS Track partners as part of the technical infrastructure of the project through server installations and provision of access to services.

The application of analytics techniques to different CS projects is facilitated by the provision of basic data on projects through the WP2 database (using MongoDB). On the other hand, results generated through analytics will be fed back to this database to enrich the description of projects and to provide indicators that characterise these projects.

In addition, WP3 will feed the computational analysis results into the community platform (WP5) and the broader analysis will be technically facilitated through adequate exchange of formats and interfaces. Here, particularly the visualisation techniques will be used to make project results available also to external stakeholders.

Section 3: Information Extraction and Semantic Modelling

This section discusses several approaches and methods from the fields of information extraction and semantic analysis that are of interest for the CS Track project. It incorporates state-of-the-art approaches in information extraction and text analytics as well as pragmatic considerations and recommendations from recent research that can guide the adoption in our project. The goal is not to advance the methods itself, but to combine good practices and well-received computational methods to characterize and analyse CS activities. This particularly involves techniques on the level of content analysis to get deeper insights into the knowledge building in CS projects based on the knowledge artefacts produced.

Section 3 starts with an overview of relevant concepts and measures from the field of information retrieval that underlie any quality characterisation of information extraction and search. An important aspect of information extraction from texts is the detection and recognition of named entities (entities with proper names such as persons, places, products etc.), abbreviated as NER. This is a common task and cornerstone for semantic technologies to transform unstructured texts into structured data with links to real-world entities. We characterise the core challenges and methods of NER in section 3.2. In the following section 3.3, we give an overview of different semantic technologies used for the extraction of semantic information from texts. We focus on common technologies for linked data to enrich the data corpus of CS Track (DBpedia and DBpedia Spotlight). In CS Track, we are for instance interested in detecting the association of CS projects to specific scientific disciplines. Such associations can be derived from a comparison of project descriptions with texts in pre-classified public knowledge bases such as Wikipedia. The method of Explicit Semantic Analysis (ESA) is an implementation of this approach and therefore also discussed in section 3.3. Section 3.4 deals with Network Text Analysis (NTA), an approach that has a long tradition in content analysis and can be used to extract relational semantic networks (or concept maps) from source texts. We reflect on recent advancements in the field of NTA that also incorporate semantic methods such as ESA. Similar to NTA, Epistemic Network Analysis (ENA, section 3.5) also extracts semantic networks form texts. However, inspired by ethnographic research, it focuses on a small number of pre-defined “codes”. This allows to characterise entities such as authors or projects in terms of “epistemic frames”. This is a basis for profiling and comparing types of discourse, e.g. between CS and other areas of science.

3.1 Information Retrieval Measures

To evaluate the performance of information extraction (IE), several measures are considered for typical IE tasks (e.g. for named entity recognition, cf. section 3.2). In relation to an ideal system response (“gold standard”), the measures precision and recall are used to assess the quality of text analysis approaches and extractors. In addition, the f-measure is a weighted harmonic mean of both. The definitions of the measures are widely used and commonly recognized from the field of information retrieval (Baeza-Yates & Ribeiro-Neto, 1999).

Precision quantifies the positive predictive value (‘true positive accuracy’), which is the fraction of relevant entities among the extracted entities.

$$precision = \frac{|\{\text{relevant documents}\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

Recall, sometimes called sensitivity or true positive rate, is the fraction of relevant concepts extracted compared to all relevant concepts.

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

The F-measure is a weighted harmonic mean of precision and recall which combines both aspects. The weighting factor is called β, where β=1 is the common case of the F-measure. This measure is usually employed to evaluate text classification tasks such as named entity recognition, as stated out by Derczynski (2016).

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

However, in a context of computer linguistics and the evaluation of semantic extraction, recall is often seen as more relevant for benchmarks as it highlights the correctness of extraction methods by quantifying the rate of true positives in all positives. Therefore, the F_2 measure is commonly used, which places more emphasis von the false negatives.

The relevant documents are derived from the gold standard, which assumes that the evaluator has the knowledge about all relevant entities to be extracted, for example, the named entities to be extracted properly. As a general remark to the notion of the mentioned formulae from the literature of information retrieval, the typical unit is "document". Therefore, the measures employed are adapted to the field of text analysis, where the unit of extraction is for example the named entity (cf. 3.2) instead of a document.

The presented measures precision, recall and F-measures do not take true negatives into account. For text analytics, true negatives are both computationally and conceptually difficult to grasp. This can yield to any arbitrary sequence of tokens or symbols from the analysed text. This notion of true negatives for text extraction leads to many combinations even for smaller texts. While this is plausible for text extraction, other approaches to adapt true negatives for the evaluation of information extraction performance exist in other fields of information extraction (Esuli & Sebastiani, 2010).

3.2 Named Entity Recognition

Named Entity Recognition is one of the core information extraction tasks in natural language processing (Nadeau & Sekine, 2007). Transforming unstructured text into structured and computational representations of text yields to the identification of so-called "Named Entities". According to Nadeau & Sekine (2007) the term named entity recognition (NER) was later coined in 1996 at the Message Understanding Conference, where much of the research focused on the extraction of information such as company names, person names or activities from unstructured information sources such as newspapers. The main information units extracted through NER are names (locations, organizations, persons, activities, …), numeric expressions (money, percent values, quantities, …) and temporal information (dates, time spans). During the development of applications for NER, certain domain-specific entity types have been considered. The following example illustrates the task of NER for the sentence "Sven bought 12 Beyond Meat burgers last week":

[Sven]$_{person}$ bought [12]$_{quantity}$ [Beyond Meat]$_{organization}$ burgers [last week]$_{time}$.

In this example, the sentence has been annotated using the specific entity types *person*, *quantity*, *organization,* and *time*. NER is the information extraction task to automatically annotate the recognized entities. However, the concrete information extracted varies on the domain and language. Even in the example above, other words can be annotated, or the annotated types might vary in the degree of specificity (being more concrete vs. being more generic). A general conceptual model for this task can be concluded as identifying entities first and classifying the types second (Farmakiotou et al., 2000). From the perspective of machine learning, the named entity recognition is a classification task and therefore often recognized as named entity recognition and classification ("NERC"). However, some definitions of the specific tasks differentiate named entity extraction (NEE), recognition (NER) and classification (NEC) in concept and implementation (Carreras, Màrquez & Padró, 2003).

While early approaches for NER are rule-based, more sophisticated solutions facilitate advanced machine learning approaches or combine the two approaches by automatically inducing rules from supervised or unsupervised learning methods. Recent approaches facilitate neural architectures and deep learning approaches for NER (Lample, Ballesteros, Subramanian, Kawakami & Dyer, 2016). However, there is a differing prevalence regarding the concrete approach in dependence on morphological character of a language and corpora for the NER. For example, a high precision can be achieved in Arabic with rule-based approaches, morphological tagging and well-defined domain models (Aboaoga & Ab Aziz, 2013). Without domain-specific corpora, Hidden Markov Models have been applied successfully to produce an accurate NER system for Indian languages (Morwal, Jahan & Chopra, 2012). The authors point out that there is a divergence between research on NER for English and other languages.

In addition to the difficulties implied by the specific character of the languages and the research focus, ambiguity is one of the grand challenges for NER. Particularly for the disambiguation, semantic technologies such as DBpedia (cf. 3.3.1) are helpful. The NERSO tool is a NER system that uses semantic open data and graph-based centrality measures to disambiguate entities (Hakimov, S., Oto, S. A., & Dogdu, E. (2012). The ontologies that are given by already existing semantic open data such as DBpedia can serve as naming services for entities. On top of the ontologies, frameworks utilize those services to perform a named entity recognition. DBpedia Spotlight is presented as one of the frameworks in section 3.3.2.

To evaluate the performance of NER, precision, recall and F-measures (cf. 3.1) can be considered. Early, but well-proven approaches from the MUC-7 conference in 1997 have a precision of 95% (recall 92%; F-Measure 93.39%), whereas more recent domain-specific approaches reported a precision and recall of more than 99% (Marsh & Perzanowski, 1998; Segura-Bedmar, Martínez & Segura-Bedmar, 2008). Schmitt et al. (2019) compared five popular software systems for NER in a comprehensive and reproducible experiment. As a result, *StanfordNLP* outperformed the compared systems *NLTK*, *OpenNLP*, *SpaCy* and *GATE*. As a remark, Schmitt et al. (2019) pointed out that many of the reviewed studies have differences in the reported performance evaluations due to different corpora or other characteristics that haven't been described clearly (e.g. the type of classifier used). Also, tagging modules that have been built on a certain corpus, tend to have lower performance on other corpora (Nothman, Murphy & Curran, 2009).

Named entity recognition is a basic task in information extraction, particularly for the extraction of structured information from texts, and thus of interest for the CS Track project. However, NER cannot be perceived as a concrete method, but as a task for information extraction. Specific semantic methods and tools such as DBpedia spotlight implement methods for NER, as presented in the following section.

3.3 Semantic Methods

3.3.1 DBpedia

The DBpedia project aims to extract structured information from web sites and interlink the resulting knowledge base with other open datasets. Initially, this comprises Wikipedia data as a main information source. Later, other data sources and ontologies such as Yago (Suchanek, Kasneci &

Weikum, 2008) have been integrated into the DBpedia knowledge base. With such an interlinking of open data, the project aims to “serve as a nucleus for an emerging Web of open data” (Auer et al., 2007). However, the core of the DBpedia is still the ontology that has been created within the project, which uses multilingual data extracted from Wikipedia:

> *"The DBpedia project extracts structured information from Wikipedia editions in 97 different languages and combines this information into a large multilingual knowledge base covering many specific domains and general world knowledge. The knowledge base contains textual descriptions (titles and abstracts) of concepts in up to 97 languages."* (Mendes et al. 2012)

For every Wikipedia page, DBpedia creates a URI for the correspondence between entity and Wikipedia page. The URIs are enriched by properties that are extracted through DBpedia and stored as RDF triples. Such triples are used to model semantic data using subject–predicate–object expressions, for example "Konrad Zuse is a Person". As RDF in DBpedia this would be modelled as the following, whereas each element of the triple refers to a URI:

```
<http://dbpedia.org/resource/Konrad_Zuse>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Person> .
```

This example clarifies the heavy use of URIs to link and disambiguate data correctly. All the URIs corresponding to the three used entities have different name spaces. The first element points to the resource "Konrad Zuse" as the subject, where the predicate classifies this as a "type"-relation with the object "Person". The URIs can be accessed via API or the web in order to browse the linked data such as other persons or other attributes of "Konrad Zuse". Data set version 3.8 of DBpedia consists of 1.89 billion RDF triples in 111 languages[5].

In addition to the extraction of information, accessibility and linking of data have become the most relevant aspects of DBpedia. RDF data can be accessed from DBpedia by using SPARQL, a graph-based query language. Due to the concept of linking data, RDF and the underlying database engines are highly relational. The SPARQL endpoint uses the Virtuoso SPARQL engine, which handles the linked open data cloud anchored by DBpedia. This engine is part of the Virtuoso Universal Server, a middleware and database engine to harmonize disparate data[6]. The Virtuoso SPARQL service implements the SPARQL Protocol for RDF and the SPARQL 1.1 specification[7]. The linked open data cloud[8] visualizes the interlinking of data between open data sources and illustrates the central role of DBpedia for the semantic web nowadays. DBpedia links to 30 external sources such as Yago[9] or Census data with a total of 31 million data triples which are extracted from the linked data sources.

The following query in RDF uses the DBpedia ontology to ask if the Amazon river is longer than the Nile. The query can be tested using the DBpedia SPARQL engine: http://dbpedia.org/sparql

```
PREFIX prop: <http://dbpedia.org/property/>
ASK
{
  <http://dbpedia.org/resource/Amazon_River> prop:length ?amazon .
  <http://dbpedia.org/resource/Nile> prop:length ?nile .
  FILTER(?amazon > ?nile) .
}
```

[5] DBpedia, Data Set 3.8 (2015). https://wiki.dbpedia.org/data-set-38. Retrieved: 2020-04-06.

[6] OpenLink Software, Virtuoso (2020). https://virtuoso.openlinksw.com/. Retrieved: 2020-07-06.

[7] W3C Recommendation 21 March 2013, SPARQL 1.1 (2013). https://www.w3.org/TR/sparql11-overview/. Retrieved: 2020-07-06.

[8] Linked Open Data Cloud, https://lod-cloud.net/#diagram. Retrieved: 2020-07-06.

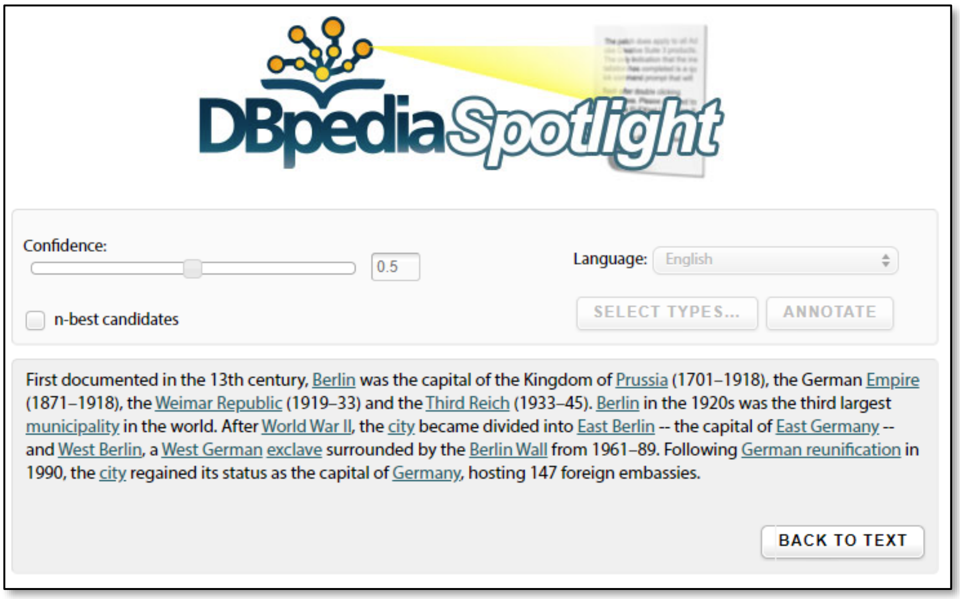
[9] Max Planck Institut, Yago (2018). https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/linking. Retrieved: 2020-07-29.

DBpedia is also of special interest for Named Entity Recognition (cf. section 3.2), where the DBpedia ontology serves as a dictionary for the entities to be recognized. Therefore, unstructured texts need to be processed and specific parts need to be matched to DBpedia entities. A framework that performs this task is DBpedia Spotlight, which is presented in the following section 3.3.2. In addition, DBpedia provides a rich ontology that links multilingual data and thus is of specific interest for the CS Track project. Extracted information from the data extraction (cf. section 2.1) can be enriched and inter-linked using the DBpedia ontology, even for descriptions in multiple languages.

3.3.2 DBpedia Spotlight

DBpedia Spotlight is a semantic technology to analyse a text and to extract entities from it (Mendes, Jakob, García-Silva & Bizer, 2011; Daiber, Jakob, Hokamp & Mendes, 2013). This is based on the DBpedia ontology and facilitates disambiguation using the context of the linked data. When we refer to concept extraction and respectively to the results of the extraction as concepts, we refer to the surface form of the corresponding URI. Therefore, we define concept extraction in the context of this work as the retrieval of relevant URIs which characterize a certain source text. Then, the URIs are projected to their surface forms for the output of concepts for our algorithms. However, the results of this extraction are then limited to concepts that have a corresponding Wikipedia (respectively DBpedia) entry. Consequently, the concepts that have been extracted typically represent domain knowledge or declarative knowledge. Procedural knowledge items can hardly be represented through this.

DBpedia spotlight comprises several APIs to access the services as a degree of freedom, the types of entities to be recognized can be selected from attached data sources (e.g., DBpedia, Freebase, Schema.org or custom). The service can be further parametrized by the language, the confidence for the recognition or the number n of the n-best candidate selection. Figure 8 shows a demo service with an already annotated text. The URIs that represent the DBpedia entities are automatically linked.



Confidence: 0.5

Language: English

□ n-best candidates

First documented in the 13th century, Berlin was the capital of the Kingdom of Prussia (1701–1918), the German Empire (1871–1918), the Weimar Republic (1919–33) and the Third Reich (1933–45). Berlin in the 1920s was the third largest municipality in the world. After World War II, the city became divided into East Berlin -- the capital of East Germany -- and West Berlin, a West German exclave surrounded by the Berlin Wall from 1961–89. Following German reunification in 1990, the city regained its status as the capital of Germany, hosting 147 foreign embassies.

Figure 8: Demo of DBpedia Spotlight. The English text has been annotated and the DBpedia entities are linked in the text using the service.

3.3.3 ESA

Explicit Semantic Analysis (ESA) was introduced in 2006 by Gabrilovich and Markovitch as an alternative to WordNet as an approach to calculate semantic relatedness (Gabrilovich & Markovitch, 2006; Gabrilovich & Markovitch, 2007). In this relatively novel approach, statistical models are combined with background knowledge to derive a vector space model from Wikipedia data. This vector space is called *concept space,* where each Wikipedia concept represents an article as a weighted vector of words that occur in this article. Within this vector space model, it is possible to compare terms or documents regarding their semantic relation. The similarity of two documents is then defined as the cosine similarity (cf. Salton & McGill, 1983) of the concept vectors. Although originally the weights have been assigned to the words using tf-idf[10] to quantify the association between words and concepts, more general variants have been applied successfully (Gottron, Anderka & Stein, 2011).

Because of this association, the approach of ESA is of interest for the CS Track project, for example, to automatically assign scientific disciplines to project descriptions. Following this method, for each discipline, concept vectors based on specific corpora are extracted and compared to a certain (textual) project description. The ESA induced similarity measure can be facilitated to assign multiple disciplines to each description, if the measures are above a threshold. An important component of ESA is the weighted inverted index that is created from the concept vectors. Thus, it is possible to link back from each word or term to the concept associated with it in the concept space. With this inverted index, any word can be interpreted as a sparse vector in a high-dimensional space spanned by Wikipedia articles (Egozi, Markovitch & Gabrilovich, 2011). However, this word-concept-matrix needs to be kept in memory for the similarity calculations and is a bottleneck in the whole processing. Apart from this limitation, the word-concept matrix (or the weighted inverted index) can be used as a mathematical foundation to text analysis tasks for arbitrary words as inputs.

Gottron, Anderka & Stein (2011) argue that ESA still achieves a good performance for documents from different corpora and even for randomly concatenated Wikipedia articles. In contrast to a regular vector space model, ESA induces a semantic similarity that can be facilitated to deal with synonyms by using the term's context to disambiguate.

In contrast to latent models such as Latent Dirichlet Allocation (Blei, Ng & Jordan, 2003), each dimension stands explicitly for a concept (Cimiano et al., 2009). Such an explicit notion of concepts in the model make this human-interpretable, whereas for LDA concepts need to be retrieved ex-post, which can be performed in extensions such as Probabilistic Latent Semantic Indexing (Hofmann, 1999). Although latent models, particularly Latent Dirichlet Allocation (LDA), are commonly used, there is a general debate about the validity and usefulness for research. LDA poses a probabilistic approach, where each document is characterized by a distribution of topics. Grimmer and Stewart (2013) point out that the method relies on a fixed set of assumptions because the model is limited to a single family of probability distributions. Further, the abstract notion of topic distributions makes it difficult for humans to interpret the results or label concrete documents with respect to the topic. Hecking and Leydesdorff (2019) point out several limitations of LDA. In their work, they reproduced and validated topic models based on previous works. As one of the results from the study, LDA is sensitive to small variations in the document corpus, which is particularly a challenge for changing corpora. They conclude that "topic models should not be used as the basis for decision making or intellectual delineations of domains in scholarly works, but for statistical purposes."

[10] Tf-idf is an abbreviation for "term frequency – inverse document frequency", a function that assigns values representing the importance of a certain term within a document. See Manning, Raghavan, & Schütze (2008).

3.4 Network Text Analysis

Network Text Analysis (NTA) is a method that is intended to extract conceptual networks that represent mental models from texts. Such networks are characterized by a set of concepts and pairwise relations (Carley, 1997). A concept is an “ideational kernel—a single idea”, which corresponds to the earlier concept of entities in NER (see section 3.2) or URIs in DBpedia (section 3.3). As an implication of this definition, it is not desirable to have surface forms with ambiguities as a representation, as it does not satisfy the condition of a single idea. Concepts are the nodes of the text network. An edge is defined as a *relationship* between two concepts. The attributes of such an edge can be the direction and the weight (“strength”). Two concepts and the relation between them is called a “*statement*” in such conceptual networks. Analogously, this can be related to RDF triples, which also capture semantic statements in subject–predicate–object expressions.

The approach presented originally by Carley (1997) has been picked up in several research projects to create automated and semi-automated frameworks for NTA in the field of text mining. Diesner and Carley (2008) developed *AutoMap* as a part of a dynamic networks project to perform a semi-automatic construction of text networks. AutoMap features many NLP routines such as stemming, reduction, normalization, part of speech tagging, but also more sophisticated mechanisms such as a proximity-based extraction of relational data, NER or anaphora resolution to resolve personal pronouns. In the NER stage, manual feedback is needed to provide thesauri for the entities to be recognized. Alternatives and applications are presented in the next section.

The extraction mechanisms and processing chains vary depending on the specific implementation. However, many of the tools use a certain kind of proximity analysis to identify relations between entities. This can be done either by a syntactical-oriented approach (textual proximity and co-occurrences in a “sliding window”) or semantic measures from ontologies such as WordNet or DBpedia. Popping (2000) provides an overview about different methods and approaches of text analysis to construct text networks in his book “Computer-assisted Text Analysis”.

As one of the benefits of text networks, network measures known from the field of social network analysis and graph theory can be applied, for example to identify key concepts in texts. Similar mechanisms have been applied in the domain of concept map evaluation, where a combination of structural and semantic measures have been used to assess concept map quality (Hoppe, Engler & Weinbrenner, 2012). Another selling point of text networks and conceptual networks is the ease of interpreting the visualizations. Therefore, text networks are a common tool in qualitative data analysis (Weitzman & Miles, 1995). These aspects are of specific interest for the CS Track project, particularly representing project data and relations between concepts and projects as interpretable network models. Furthermore, such representations and network models can be quantified using measures from the field of (social) network analysis and graph theory (cf. section 4).

3.4.1 Applications and Implementations of Network Text Analysis

In addition to the *AutoMap* software to create text networks that has been presented previously, several alternatives and applications are listed in this section.

NodusLabs created the commercial software *InfraNodus* (open source, non-commercial free use), which has been used in an interesting application to visualize the relations in the leaks of the panama papers. Paranyushkin (2019) published a whitepaper about the tool and the respective processing chain. While the construction of the network is quite rudimentary, most of the processing is performed in the visualization stage. This is probably due to the fact that the application case of the panama paper leaks demands the user to reinspect the original texts.

The R package *textnets* was motivated by the analysis of the newspaper articles (Bail, 2016; Rule, Cointet & Bearman, 2015). Bail (2016) presented an analysis of the discourse on autism spectrum disorder and the relation to vaccination in newspapers using computational methods of network text analysis and social network analysis measures. The *textnets* package features methods to convert texts into two-mode (bipartite) networks that connect two different entity types with each other (e.g. concepts with affiliations). According to Bail (2016), the implemented mechanisms are better suited for short texts such as social media messages.

In the *JuxtaLearn* project a dedicated learning flow has been created, where the learning takes place by creating videos (Malzahn, Hartnett, Llinás & Hoppe, 2016). The JuxtaLearn process is divided into eight steps such as planning, editing, and sharing creative videos in the CLIPIT environment. By creating videos, the learners overcome specific stumbling blocks in so-called "tricky topics" (Cruz, Lencastre, Coutinho, José, Clough & Adams 2017). These topics usually relate to science topics, particularly in the STEM fields and are predefined by instructors. Although an automatic video content analysis was not foreseen, research around the project context facilitated semi-automatic methods for the content analysis of the created learning videos (Erkens, Daems & Hoppe 2014). Network Text Analysis has been applied to relate signal words and domain concepts to indicate missing pre-knowledge and misconceptions. This incorporated both learning videos and context, particularly the comments that were given on the videos. Such network-based approaches ("NTA") are often using a sliding window approach, which quantifies proximity of words in a sentence to establish connections between nodes (Hecking & Hoppe 2015).

Text2Network is a web service that employs network text analysis for the extraction of key concepts (Hecking and Hoppe 2015). The service uses a technical implementation of the NTA approach using the DKPro framework (Gurevych et al. 2007) and Apache UIMA 4. Hecking, Dimitrova, Mitrovic & Hoppe (2017) applied NTA to characterize the engagement of learners within the active video watching project. In the context of CS Track, such approaches can be used similarly for the characterization and comparison of CS projects regarding engagement and participation of volunteers in the landscape of citizen science.

In the implemented approach, concepts are extracted using the Stanford part-of-speech tagger (Manning et al. 2014) and chunking[11] to create meaningful noun phrases from the output of the tagger. An entity resolution step helps to identify similar concepts, based on text similarity. For a threshold of 0.7, similar nodes are merged. A relation between two concepts is established if both words co-occur in a sliding window of a certain size.

However, this approach has some limitations. First, a plain NTA approach is not able to identify compound terms without the existence of an external knowledge source for co-occurrences, e.g., in a specific corpus. The sliding window approach splits, for example, Caesar cipher into both terms Caesar and cipher, because the part-of speech tagger does not have the knowledge about this unique compound concept. Second, the relations between nodes are based on proximity in a sentence, which is error-prone to using relative and demonstrative pronouns. A disambiguation is not part of the process chain and is even competing with the entity resolution, which is based on string similarity. To achieve the desirable result of spotting compound terms, a dictionary-based approach has been implemented for the Text2Network service. This can be used to bridge ontologies using dynamically created dictionaries into the service. This has been done in an application that facilitates the semantic technology as a service (Manske, 2020). Following this approach, the dictionary has been enriched by synonyms to improve the precision. Such a method helps to bridge science-related concepts that are externalized in knowledge sources like DBpedia to text networks.

[11] Apache, OpenNLP Chunker (2019). https://opennlp.apache.org/docs/1.7.1/apidocs/opennlp-tools/opennlp/tools/chunker/Chunker.html, retrieved 2019-06-01.

3.4.2 Improving NTA using semantic methods

Although co-occurrences of words are easy to spot using a text-proximity, such approaches need a manual feedback, dictionary, or external knowledge source in order to spot compound terms. It is desirable to spot a compound term such as “semipermeable membrane” and treat it as a single node instead of handling “Semipermeable” and “membrane” separately. While unsupervised approaches for key phrase extraction solve this to some extent, Hasan and Ng (2014) state out that such algorithms still need to incorporate background knowledge in order to improve their performance and accuracy. This can be achieved by adding external knowledge representations such as Wikipedia or DBpedia-related knowledge sources. The ESA-T2N approach is such an extension to network-text analysis that incorporates explicit semantic analysis (“ESA”) in order to bridge science-related concepts to text networks (Taskin, Hecking & Hoppe 2019). Thus, such approaches convert a text to a network (“T2N”) structure which externalizes a mental model of the text modelling semantic similarity or other semantic measures of the connected concepts.

In contrast to pure network-based or linguistic approaches, it is possible to automatically link concepts by using Wikipedia or DBpedia-related data (Auer et al., 2007). The DBpedia project aims to extract structured information from Wikipedia as RDF triples. These triples are interrelated to external open datasets to enrich the data, for example with synonyms, translations, or geo-location. Each entry in DBpedia or Wikipedia has a dedicated URI representing a unique concept. When two different surface forms have the same URI, they are mapped to the same concept. For example, this is the case for “chemical reactions” and “reaction type”, where both have the same URI and a corresponding entity “Chemical reaction”. This mechanism is facilitated by DBpedia Spotlight (Mendes, Jakob, García-Silva, & Bizer 2011), which analyses texts and spots DBpedia entries in it (and therefore Wikipedia entries as well). The spotting uses common techniques from NLP as mentioned in section 3.3.2, which outputs multiple matchings allowing the spotting of compound terms. In addition, it provides a disambiguation based on a vector-space model to output the correct surface form depending on the context.

3.5 Epistemic Network Analysis (ENA)

Epistemic Network Analysis (ENA) is a set of cotemporal analytical techniques for quantifying, visualizing, and interpreting network data (Shaffer, Wesley, Collier and Ruis, 2016). This method of analysis bridges over between quantitative and computational methods of data analysis and network analysis and interpretative, ethnographic methods used in social studies (Shaffer, 2017). ENA utilises epistemic frames theory to analyse discourse data in individual and collaborative settings and it can be used to characterize and profile different contributions in a discourse or different types of discourse (Shaffer, 2004). The original theory was developed to model the patterns of association between skills, knowledge, identity, value, and epistemology (SKIVE), elements which constitute an epistemic frame (Cai et al., 2017; Shaffer, Wesley, Collier and Ruis, 2016). The epistemic frames theory models the ways of thinking, acting, and being in the world of some community of practice (Rohde and Shaffer, 2004; Shaffer and Ruis, 2017). However, ENA has evolved to a versatile method that supports meaning-making from any system that presents a complex network of dynamic relationships among a fixed set of elements (concepts) using pre-coded semantics (Shaffer and Ruis, 2017).

Within ENA, a network of relationships among different codes (concepts) is created for each unit of analysis (e.g., citizen science participant in a discussion forum). A main theoretical assumption of ENA is that repeated co-occurrences of two or more codes in the discourse can reveal epistemic networks which characterize an underlying discourse (Csanadi et al., 2017; Gee, 1999; Collier et al., 2016) - two codes are considered related if they co-occur in the same chunk of text, called stanza (or conversation). To identify a unit of analysis for calculating such co-occurrences, ENA provides an adaptable feature: the moving stanza window size (MSWS; Siebert-Evenstone et al., 2016). The term stanza window refers to a window or scope within which ENA is searching for connections. Moreover, within this graph-based technique, associative connections are established through relative weighting. Finally, statistical techniques are used to compare the most noticeable properties of networks generated in the context of the content of the network and traces of learning processes (Shaffer, Wesley, Collier and Ruis, 2016).

ENA provides an alternative to traditional frequency-based approaches (also known as "coding and counting") by examining the structure of these co-occurrence, or connections in the coded data. Thus, ENA is a novel method, since compared to other methodological approaches (e.g. sequential analysis), it allows (1) modelling whole networks of connections and (2) affording both quantitative and qualitative comparisons between different network models (Csanadi et al., 2017). As a result, the researcher can search connections not only within propositions (as in the case of "coding and counting" approaches) or between neighbouring propositions, but even between propositions that are two, three or more steps further from each other in the discourse (Csanadi et al., 2017).

ENA is of specific interest to CS Track because it allows for extracting semantic networks from the textual descriptions produced by the citizen science initiatives to be studied. Firstly, contrasting these semantic structures to the structures underlying our standard taxonomies of science or curricular structures will allow, for instance, for judging scientific quality and richness of the activities. Secondly, newly emerging relationships and topics may be considered as a source of information for revising and adapting science curricula in the light of new developments.

Epistemology has already been an object of study in several citizen science research projects. Watson and Floridi (2018), for instance, analysed the epistemological implications of crowdsourced citizen science projects within the context of the Zooniverse platform (the world's largest citizen science web portal, according to the authors). In the study, they showed how information and communication technologies enhance the reliability, scalability, and connectivity of crowdsourced e-research, giving online citizen science projects powerful epistemic advantages over more traditional modes of scientific investigation.

Citizen science projects are often characterized by using online collaboration between experts and amateurs on scientific research (sometimes called volunteers). In this vein, Kasperowski and Hillman (2018) have investigated the epistemic culture of a large-scale online citizen science project, the Galaxy Zoo (from Zooniverse platform), that turns to volunteers for the classification of images of galaxies. The same authors have studied the tensions that arise between the mobilizing values of a project and the epistemic cultures and subjects that are enacted by the volunteers.

Turning now to the relationship between epistemology and learning, Vallabh, Lotz-Sisitka, O'Donoghue and & Schudel (2016) examined the relationship between epistemic cultures in citizen science projects and learning potential related to matters of concern. Results of their study showed an iterative relationship between matters of fact and matters of concern across the projects. Moreover, they observed that the nexus of citizens' engagement in knowledge production activities varied. They concluded that the knowledge-production purposes informed and shaped the epistemic cultures of all the sampled citizen science projects, which in turn influenced the potential for learning within each project (Vallabh et al., 2016).

Whereas epistemological implications and the connections among the implied actors (e.g. expert-citizen relationships) have been studied in the citizen science context, there are less studies that incorporate and take advantage of ENA techniques to analyse these complex networks that are the citizen science communities. In the context of CS Track, the use of ENA can bring novelty and an added value to the expected research results. Moreover, it is expected that ENA will be combined with other analyses techniques to get full potential of analysis, e.g. combining ENA with social network analyses (SNA) following the SENS or ISENS approaches (Gašević, Joksimović, Eagan,& Shaffer, 2019; Swiecki & Shaffer, 2020).

Section 4: Network Analysis Techniques

Technically mediated communication, ranging from email exchange to interactions in social media channels, leaves traces that can be used to determine the relationships between the communicating persons. Some of these relations such as “A following B” (on Twitter) or “A and B being friends” (on Facebook) are even explicit in the original information environment, others can be derived from available trace data such as messages exchanged between A and B. Such relations can be modelled as networks (graphs), which allows for applying mathematically well understood in the analysis of such structures. Even before the massive usage of communication through digital media, the same techniques have been used in the analysis of relations between members of scientific communities as part of the methodological repertoire of “scientometrics”. Typical networks in this area are based on co-authorship relations (between scientists) or citations graphs (between publications). Although these methods are often subsumed under the notion of “Social Network Analysis” (SNA), they are not limited to social (person-person) relationships in a narrow sense. Citation graphs are an evident example, but also mixed networks that connect actors (e.g. authors) to artefacts (e.g. publications) can be analysed using corresponding techniques. The basic claim is that what has been beneficial in the structural-relational analysis of scientific activities should also be relevant to the analysis of CS activities. Different from classical scientometrics, we will however not only and not even primarily rely on analysing official scientific publications but rather rely on project web pages, forums and social media channels as sources. This corresponds to the less formal nature of many CS activities.

After a short introduction to the origins of network analysis and the over-arching idea of “network science” (section 4.1), we will elaborate on different notions of “centrality” in networks and on the detection “cohesive subgroups” or “subcommunities” in section 4.2. Centrality measures are node attributes that are typically interpreted as a certain type of influence, whereas subcommunities are divisions of the whole network in which the members are more closely connected than the average. Meanwhile these techniques are quite well known and accepted in many fields of research on team collaboration (e.g. in software development or Wikipedia authoring) as well as collaborative learning and knowledge building. There is big unexploited potential for applying such methods to CS activities, too. In addition, we propose to also consider less widely used approaches that allow for analysing hybrid actor-artefact networks (section 4.3) and knowledge evolution in knowledge building and sharing communities (section 4.4). This is an exploratory but promising part of our agenda. Section summarised certain requirements related to data formats used in both content and network-related analysis components. Some techniques (such as NTA and ENA) already integrate both perspectives (content/network), but also in general it is important to support data integration.

A worked-out example of what we can expect from applying network analysis techniques to CS projects is given in the following section 5.

4.1 Network Science and Network Analysis

“Network science” is an interdisciplinary paradigm that generalizes studies of the structure and evolution of networks of various nature, such as technical networks and their applications (including WWW and Internet), biological networks as well as social networks and linked communities. Network science provides models that simulate and possibly explain the emergence of certain structures in networked communities based on relations involving actors and artefacts. It also provides mathematically well understood methods of analysis to detect such structures in existing networks. In this sense it resonates with social theories such as Actor-Network Theory (Latour, 2005). A well-established application of network analysis techniques is the analysis of scientific cooperation and production in the field of “scientometrics” (see, e.g., Leydesdorff, 2001). Network analysis has the potential of providing a general formal-analytic underpinning for the study of collaboration in networked communities.

Reflecting seminal pieces of work in the progression of network analysis, Newman, Barabási and Watts (2006) summarize several modelling approaches and analytic results that led to the notions of “scale free” or “small world” networks. These findings have challenged and modified the original assumptions about the evolution of dynamic networks as “random graphs” (Erdős & Rényi, 1959). Studies of citation networks (de Solla Price, 1965) made social networks a popular theme in scientometrics. More recently, many web-based or online communities have become social networks in their own right (Twitter, Facebook etc.).

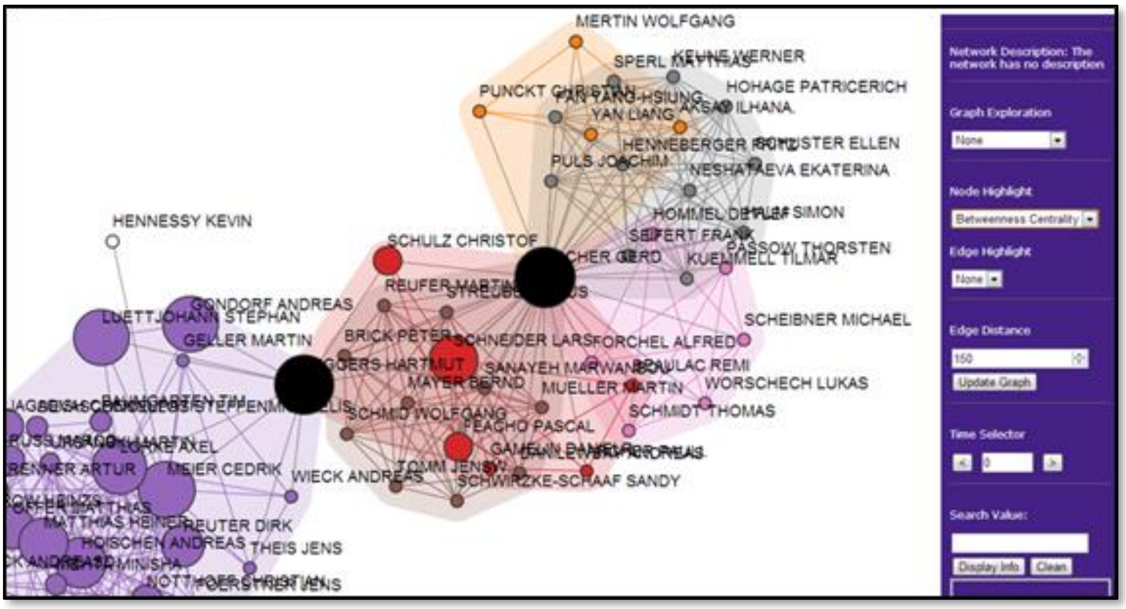
Social Network Analysis or SNA (cf. Wasserman & Faust, 1994; Borgatti et al., 2009) is characterized by taking a relational perspective on actors as parts of network structures. In this sense, a network consists of a set of actors together with a set of connections or ties between pairs of actors (Wasserman & Faust, 1994). Examples of different kinds of ties are affiliation, friendship, professional, behavioural interaction, or information flows. Together with the evolution of network analysis techniques also the visualization of such network structures has emerged as a related subarea (Krempel, 2005). A well-known inherent limitation of SNA is that the target representation, i.e. the network or sociogram, does no longer represent temporal characteristics but aggregates data over a given time window. It has been shown that the size of the time window has a systematic influence on certain network characteristics such as subcommunity structures (Hecking et al., 2013). Of course, the dynamic evolution of networks is also of interest. To explicitly address time-dependent effects, SNA techniques have been extended to analysing time series of networks in dynamic approaches.

4.2 SNA Measures - Centralities and Cohesive Subcommunities

SNA methods can be used to identify important or also marginal or isolated actors in networks both for research purposes (i.e. understanding the nature of a collaboration network) as well as for regulative feedback and reflection on the part of the actors (i.e. supporting the network during interaction). Means for this are centrality measures, group detection or positional analyses within a network. Among the most common usage types of SNA methods in community scenarios are the identification of most central actors and isolated actors. A popular approach for this end is the computation of centrality measures (Wassermann & Faust, 1994), such as the degree centrality that represents how many links an actor has to other community members. In directed networks, we must distinguish “indegree” and “outdegree”. E.g., in social network based on a “following” relationship, a high indegree (also called “prestige”) would characterize actors by their popularity in the community whereas a high outdegree would characterise networking effort in terms of establishing many links. Refinements of the standard degree count are measures that consider not only the sheer number of neighbours but also their weight. Measures of this type are “Eigenvector centrality”, “Katz centrality” and the well-known “Page Rank” measure (Franceschet, 2011). Recent studies show the similarities and differences between such measures (Rosa et al. 2018; Was & Skibski, 2018). Centrality measures such as “closeness” of “betweenness” take into account the specific of position of nodes in the overall network, e.g. nodes close to the periphery cannot have a high closeness centrality even though they may have a high number of (local) neighbours. High betweenness values are characteristic of nodes representing actors that bridge over between different more densely connected components of a network. Whereas plain “degree” is a local property and easy to calculate, the other measures require knowledge of the whole network and come with a higher computational cost.

Social networks tend to develop cohesive substructures (also called “subcommunities”). There are many different operationalisations and underlying models for the detection of such subcommunities. A comprehensive overview of such methods is given by Fortunato (2010). The different methods coincide in the point that a subcommunity must be more densely connected than the whole network on average. Subcommunity detection is particularly relevant for the analysis of collaborating communities because it allows for identifying groups of actors who tend to interact more closely (or intensively) than the rest. Concerning the diffusion of information in a social network, subcommunities can be considered as components in which information is circulated internally with limited access from outside, which may lead to undesired effects such as “filter bubbles”.

Like centrality measures, methods to calculate cohesive subcommunities in larger networks have also gained considerable popularity in different application fields. The simplest of such methods are based on elementary graph-theoretic constructs such as cliques. In this category, clans, cores, or k-plexes are relaxed variants of cliques as completely connected subgraphs with relaxations on distance or number of neighbours. The more sophisticated clique-percolation method (Palla et al., 2005) allows for detecting potentially overlapping subcommunities, whereas the Girvan-Newman approach is based on continuous splitting and partitioning (i.e. no overlaps) controlled by the modularity as a numerical measure of cohesion in a subnetwork (Girvan & Newman, 2002). The analysis of subcommunity structures is a particularly challenging problem if combined with the dynamics or evolution of the underlying networks. The additional challenge has to do with identifying continuity or discontinuity (e.g. split, decay or emergence) of the cohesive subgroups. Palla, Barabasi, and Vicsek (2007) have introduced a basic conceptualization and approach to deal with these phenomena.



Network Description: The network has no description

Graph Exploration: None

Node Highlight: Betweenness Centrality

Edge Highlight: None

Edge Distance: 150

Time Selector: 0

Search Value:

Figure 9: Network visualization of a co-publication network (based on bibliographic data).

Figure 9 shows an example graph based on a collection of publications from a research cluster in the area of nanotechnology. The underlying analysis was conducted in the context of the EU project SISOB (“An Observatorium for Science in Society based in Social Models”, 2010-13) using the SISOB network workbench (Göhnert et al., 2013). In this visualization, the size of the nodes is proportional to the betweenness centrality value, which indicates the characteristic of a certain author to “bridge over” between different groups of authors. Regarding the subcommunity structure, every underlying single publication induces a completely connected subgraph (i.e. a “clique”) among all its co-authors. The clique percolation method that has been applied here extends such basic cliques to form bigger subcommunities if certain connectivity requirements are fulfilled. The basic links represent co-authorship relations on specific articles. These subcommunities are visualized as coloured clouds. The one in the lower left edge is an example that comprises more than one clique of co-authors. The nodes which are coloured black represent authors that belong to several cohesive clusters. Apparently, these are also characterized by a high betweenness centrality (node size).

Computer-supported collaboration is typically reflected in terms of several distinct relations between the actors, such as direct communication using chat accompanied by knowledge co-construction using wiki tools. The analysis of relational patterns between these different relations can make use of methods for role analysis (Wasserman & Faust, 1994) or "positional analysis" (Doreian et al., 2002). The resulting "block models" characterise actors by the similarity (presence or absence) of their relations with other actors. In contrast to a grouping based on cohesion, the corresponding aggregation is along positional or role aspects and does not mean that actors in the same position are connected to each other (e.g., in an organization, secretaries have similar roles to other actors without necessarily being densely connected within their group). Although this type of analysis can create very interesting insights, it is often misunderstood, especially confounded with subcommunity detection, and still not much used outside the expert community in network analysis.

4.3 Hybrid (Bi-partite) Networks Involving Actors and Artefacts

E-mail exchange or direct messaging are types of digital interactions that can be directly used as indicators of actor-actor relationships. However, in the absence of direct communication channels such as chats or mailing lists it is still possible to infer indirect relations between actors mediated by artefacts. For example, based on the log protocols of actors and the resources they accessed or edited it is possible to derive bipartite, or "two-mode" networks where actors are affiliated to "knowledge artefacts" created and made available in the community. In learning environments, this has been interpreted as a kind of social-thematic navigation through the sharing of learner-created "emerging learning objects". Such relations between groups of actors (or specifically learners) are induced by thematically related objects, which may indicate a common interest (Hoppe et al., 2005). Interaction can then take place indirectly mediated by those objects without necessarily implying a person-to-person communication. Typical examples for bipartite networks in collaboration scenarios are users and forum topics, researchers and their affiliations to conferences, or wiki editors and articles they modified.

Technically, the relation between actors (or authors) and the artefacts or products can be regarded as another basic relationship, which can be modelled as so-called two-mode-networks. In the context of SNA, such two-mode-networks are called "affiliation networks" (Wasserman & Faust, 1994). In pure form, these networks are assumed to be bi-partite, i.e. only alternating links actor-artefact (relation "created/modified") or artefact-actor (relation "created-by/modified-by") would be allowed. Using simple matrix operations such bi-partite two-mode-networks can be "folded" into homogeneous (one-mode) networks of either only actors or only artefacts. Here, e.g., two actors would be associated if they have acted upon the same artefact. We would then say that the relation between the actors was mediated by the artefact. A typical example of such a transformation is found in co-publication networks based on co-authorship. Similarly, we can derive relationships between artefacts by considering agents (engaged in the creation of two different artefacts) as mediators. The folding of a two-mode network into a one-node network leads to information loss. Therefore, it is preferable to maintain the richer representation unless the specific "projection" is needed.

Certain algorithms for the detection of cohesive subgroups or "subcommunities" have been generalized to also work with bi-partite networks. One example is the "bi-clique communities" method (Lehmann, Schwartz, & Hansen, 2008). In analogy to the original clique percolation method (by Palla et al., 2005), it allows for detecting potentially overlapping subcommunities. It can be directly applied to affiliation networks. Hecking, Ziebarth, and Hoppe (2014) have applied this technique in the analysis of networks made up from learner-resource interactions, specifically students accessing learning materials on a learning platform. In two case studies, one with a small blended learning course on interactive learning and teaching technologies and a second one with a large online lecture on computer mediated communication the bipartite clustering approach could give surprising insights into the patterns of resource usage. As a result, students and resources were grouped into mixed and possibly overlapping clusters. Such a cluster can be interpreted as indicating a specific common interest of the involved group of students in the connected learning resources, i.e. as a kind of specific interest group. Being part of such an interest group does necessarily imply social connections.

A typical example cluster is depicted in Figure 10. The bipartite representation allows for detecting clusters with a much better "resolution" as compared to a cohesive subgroup analysis after a folding or projection operation.

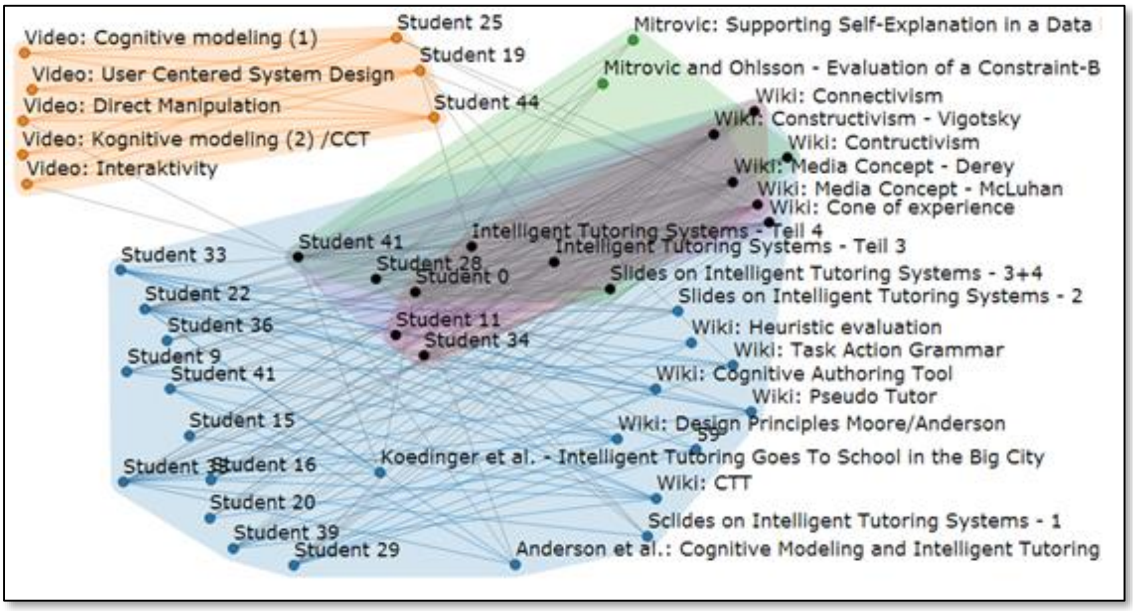

Video: Cognitive modeling (1)
Student 25
Student 19
Video: User Centered System Design
Video: Direct Manipulation
Student 44
Video: Kognitive modeling (2) /CCT
Video: Interaktivity
Mitrovic: Supporting Self-Explanation in a Data
Mitrovic and Ohlsson - Evaluation of a Constraint-B
Wiki: Connectivism
Wiki: Constructivism - Vigotsky
Wiki: Contructivism
Wiki: Media Concept - Derey
Wiki: Media Concept - McLuhan
Wiki: Cone of experience
Intelligent Tutoring Systems - Teil 4
Intelligent Tutoring Systems - Teil 3
Student 33
Student 41
Student 28
Student 0
Slides on Intelligent Tutoring Systems - 3+4
Slides on Intelligent Tutoring Systems - 2
Student 22
Student 36
Student 11
Student 34
Wiki: Heuristic evaluation
Wiki: Task Action Grammar
Student 9
Student 41
Wiki: Cognitive Authoring Tool
Wiki: Pseudo Tutor
Student 15
Wiki: Design Principles Moore/Anderson
59
Student 38
Student 16
Koedinger et al. - Intelligent Tutoring Goes To School in the Big City
Wiki: CTT
Student 20
Sclides on Intelligent Tutoring Systems - 1
Student 39
Student 29
Anderson et al.: Cognitive Modeling and Intelligent Tutoring

Figure 10: Bipartite clusters of students and learning resources.
(Black nodes belong to more than one cluster.)

4.4 Network Models of Knowledge Evolution in Science

Scientific production is a prototypical case of knowledge building in communities. For example, the knowledge building pedagogy introduced by Scardamalia and Bereiter (1994) is essentially based on this analogy. Accordingly, methods that have been developed to analyse scientific production ("scientometric methods") can plausibly also be used to analyse other types of knowledge building in networked communities. Scientometric methods are tailored to the analysis of the inter-relation between actors (authors) and knowledge (most prominently publications).

Under the label of "main path analysis" (MPA), Hummon and Doreian (1989) have proposed a network-analytic method to detect the flow and evolution of scientific ideas based on citation graphs. The original paper is based on an example corpus of publications on DNA in the field of biology ranging over several decades. The original MPA method relies on the acyclic nature of citation graphs. If the direction of the relations between the documents is modelled according to the flow of information, namely from the cited to the citing publication, there will be information sources (nodes with no ingoing edges) and information sinks (nodes with no outgoing edges). These "open ends" are connected to only one source and one sink, which closes the whole graph. Now, the idea of MPA is to find the path of the maximum information flow from this source to the sink. One common method to find these edges is the "search path count" or SPC method (Batagelj, 2003). All sources in the network are connected to a single artificial source and all sinks to a single artificial sink. SPC assigns a weight to an edge according to the number of paths from the source to the sink on which the edge occurs. The main path can then be found by traversing the graph from the source to the sink by using the edges with highest weight as depicted in Figure 11.

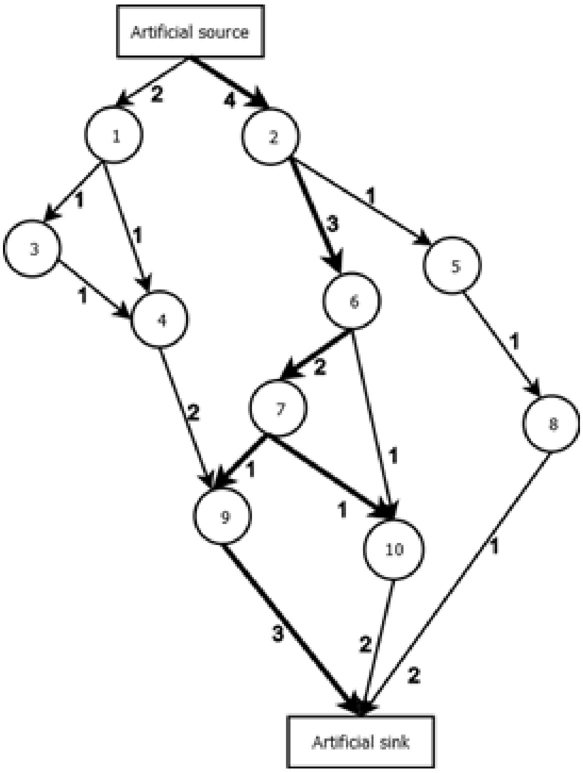


Figure 11: Example network - edge weights were calculated according to the SPC method (with thick edges indicating the main path)

Halatchliyski et al. (2014) have applied this method to analyse the flow of ideas in a networked knowledge building and learning community ("Wikiversity"). This community can be conceived as similar to Wikipedia in terms of building on use of user-created content with an additional focus on supporting learning and teaching among the community members. In such wiki environments, the equivalent of citations are hyperlinks between articles. However, MPA cannot be directly applied to this case because the underlying link structure is typically not acyclic (two articles may even be interlinked in both directions). Since the content of articles in a wiki is dynamic, hyperlinks between two articles do also not induce a temporal order. To cope with this issue, the approach introduced additional revision links with multiple copies of the article pages (as network nodes). Whereas a wiki article may change in content over time, such revisions are knowledge artefacts with a fixed content and an ordered time structure. The resulting graph is directed, time-compliant and acyclic and fulfils the requirements of the MPA method. Figure 12 shows an example from the medical area of Wikiversity. It is reduced to parallel main paths and shows revisions (e.g., a chain of revisions in the upper left, originating from user 9357) and the inter-linking of different pages (e.g. "Gynaecological History" and "Menorrhagia" by user 15539).

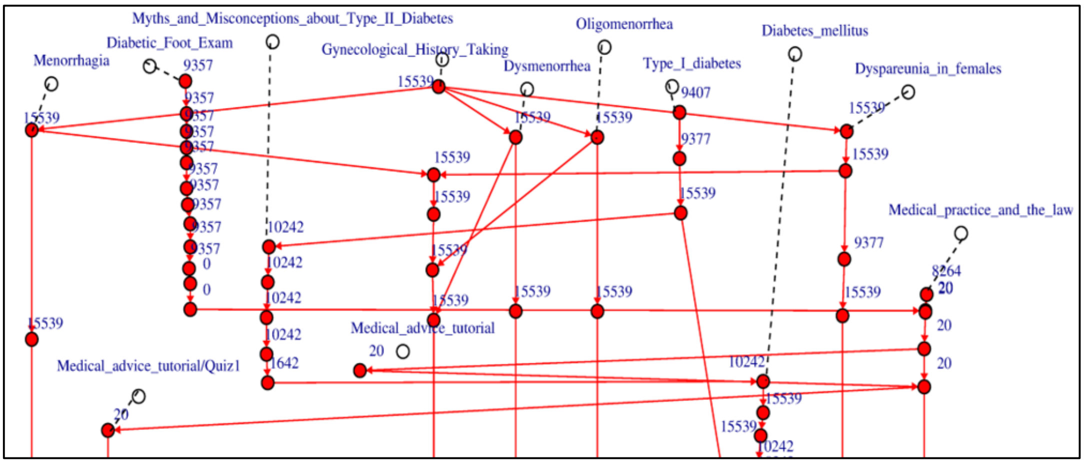
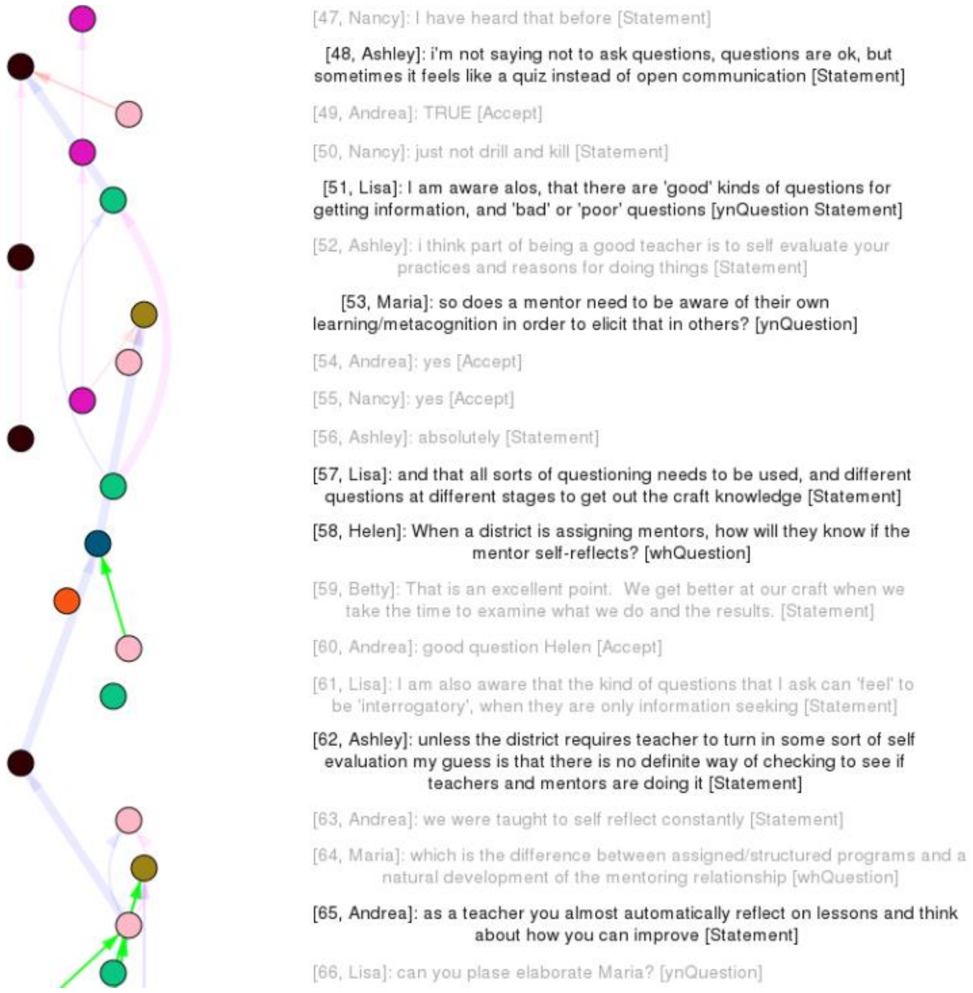


Figure 12: Creating revisions and inter-linking in Wikiversity.

In the Wikiversity study, the results of MPA were used to distinguish contributions that were on a main path from others considered less important. This distinction could also be used as a weighting factor in a comparative analysis of user activities.

Another application study (Hoppe et al., 2014) combined MPA with existing approaches to the analysis of chat interactions in a group of educational experts and decision makers. Chat is a synchronous communication medium that has a temporal precedence structure, however, possibly includes parallel threading. There is no explicit and "given" equivalent to citation links because postings that follow each other in a chat session are not necessarily directly related, which does necessarily imply that they could be "linked" to each other. This implies that such links have first to be established based on certain clearly defined criteria. This was achieved by taking the "contingency analysis" approach (Suthers et al., 2010) as a reference and starting point. Although initially based on human judgement, contingency analysis comes with quite clearly defined rules for establishing links. Adding methods developed in the context of "dialogue act tagging" (Wu et al., 2005) allowed for operationalising these rules for automatic processing.

Applied to the available corpus of the educational expert chat, the automatic results could be compared to the ones that were previously gained through hand-coding (Suthers and Desiato, 2012). The automatically generated contingencies reached an F-score similarity of 83% to 97%, which is comparable to the pairwise F-score similarity of results from different human coders. Figure 13 shows a fragment of the chat with an overlay of contingency links and contributions on a main path in boldface. As can be seen in this figure, MPA allows for filtering the discourse for main threads of argument.



[47, Nancy]: I have heard that before [Statement]

[48, Ashley]: i'm not saying not to ask questions, questions are ok, but sometimes it feels like a quiz instead of open communication [Statement]

[49, Andrea]: TRUE [Accept]

[50, Nancy]: just not drill and kill [Statement]

[51, Lisa]: I am aware alos, that there are 'good' kinds of questions for getting information, and 'bad' or 'poor' questions [ynQuestion Statement]

[52, Ashley]: i think part of being a good teacher is to self evaluate your practices and reasons for doing things [Statement]

[53, Maria]: so does a mentor need to be aware of their own learning/metacognition in order to elicit that in others? [ynQuestion]

[54, Andrea]: yes [Accept]

[55, Nancy]: yes [Accept]

[56, Ashley]: absolutely [Statement]

[57, Lisa]: and that all sorts of questioning needs to be used, and different questions at different stages to get out the craft knowledge [Statement]

[58, Helen]: When a district is assigning mentors, how will they know if the mentor self-reflects? [whQuestion]

[59, Betty]: That is an excellent point. We get better at our craft when we take the time to examine what we do and the results. [Statement]

[60, Andrea]: good question Helen [Accept]

[61, Lisa]: I am also aware that the kind of questions that I ask can 'feel' to be 'interrogatory', when they are only information seeking [Statement]

[62, Ashley]: unless the district requires teacher to turn in some sort of self evaluation my guess is that there is no definite way of checking to see if teachers and mentors are doing it [Statement]

[63, Andrea]: we were taught to self reflect constantly [Statement]

[64, Maria]: which is the difference between assigned/structured programs and a natural development of the mentoring relationship [whQuestion]

[65, Andrea]: as a teacher you almost automatically reflect on lessons and think about how you can improve [Statement]

[66, Lisa]: can you plase elaborate Maria? [ynQuestion]

Figure 13: Fragment of a chat protocol with inferred contingency links (main path contributions and links indicated in bold).

Here again, the contributions that lie on a main path information should be interpreted as particularly relevant for the evolution and progress of the overall discussion. The percentage of contributions on a main path (%MainPath) can in turn be interpreted as an indicator of influence of a participant in the overall discussion. These values were found to be quite highly correlated (r=0.82) with the PageRank centrality of actors in a sociogram derived from the original contingency graph (hand-coded version, not using MPA).

4.5 Data Formats

The analytical methods and tools impose requirements for specific data structures and formats. The different techniques can be categorized into two different data types: (1) unstructured textual data, for example textual project descriptions; (2) structured network data such as extracted text networks. Although some of the tools use proprietary formats, we outline the rationale for selecting data formats and structures in the following.

4.5.1 Network and graph data (SNA)

GML (Himsolt, 1997) and GraphML (GraphML Working Group, 2009) are widely known formats for representing graphs and (social) network data. Such networks are initially constructed from a data collection and afterwards analysed and visualized using certain software tools. This multi-layered process involves several data exchanges between processing layers, for example visualizing network measures that have been previously computed in the analytics step. This might affect the structure, topology, shape, geometry and the rendering of graphs (Brandes, Marshall & North, 2000).

However, altering the network usually goes along with a transformation of the network structures. Therefore, another approach of representing graphs is the use of a decorator-pattern. The SiSOB Graph Format (SGF) approached this by providing analytical tools to annotate the network while keeping the original file along with the annotations (Göhnert, Harrer, Hecking & Hoppe, 2013). Thus, it is easy to add network measures such as centralities to the graph without changing the original structure or format. Although, there no de facto standard for graph formats exists, many of the already existing analytical tools that are intended to be used within the project scope such as Gephi[12] for the visualization or NetworkX[13] as a common library for SNA are using widely recognized and common formats such as GML or GraphML. Due to the variety of already existing tools in this scope, a pragmatic approach for integrating and developing analytical tools is the choice of a widely recognized interchange format.

4.5.2 Textual data (text analytics)

For the analysis of textual data, i.e., text analytics, text mining and natural language processing, no common exchange format exists. Usually, unstructured text is transformed into quantitative or relational data. Particularly for encoding knowledge, a variety of generic formats (e.g. RDF) exist. However, tools such as DBpedia (cf. section 3.3) use this internally for their knowledge representations, but not for the input of artefacts or the representation of processed output. Commonly, proprietary formats are used to construct the output of text analytics tools. Although ENA (cf. section 3.5) constructs networks, the input is provided in a proprietary stanza format (Shaffer, Collier & Ruis, 2016). In natural language processing, it is desirable to create workflows to process inputs in multiple steps, for example to compose processing chains that consist of stemming, term weighting, lemmatization, and named entity recognition. Thus, it is a natural requirement to produce internal formats that allow for data exchange. One of the common frameworks for implementing text processing pipelines is Apache UIMA[14], which uses an annotation approach similarly to the decoration pattern in the SiSOB Graph Format.

In summary, there is no widely recognised format for all the tasks. The pattern to decorate or annotate the original file is useful in order to preserve the original structure while it is possible to add and enrich the data with analytical tools. However, many of the already existing tools use proprietary formats or limited formats such as GML that are not capable of decorations. Therefore, the choice of certain data formats is rather induced by the tool.

[12] Gephi.org, Gephi (2017): https://gephi.org/.

[13] NetworkX developers, NetworkX (2020): https://networkx.github.io/.

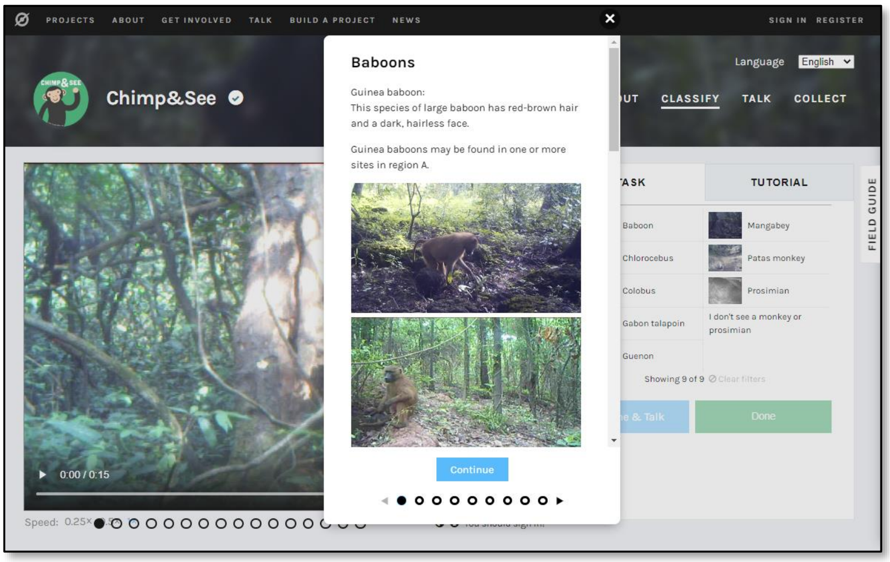
[14] The Apache Software Foundation, Apache UIMA (2013). https://uima.apache.org/. Retrieved: 2020-07-30.

Section 5: Case Study - Applying SNA to Zooniverse Forums

5.1 The Chimp & See Project

Zooniverse[15] is one of the larger citizen science web portals. It has been launched in 2009 by the Citizen Science Alliance (CSA), which has board members from various institutions such as the Adler Planetarium or the Johns Hopkins University. Until now, the platform has more than 1.6 million registered volunteers participating in citizen science activities. In June 2020, the platform registered 99 active, 122 paused and 46 finished projects[16]. The projects encompass crowdsourced CS activities with *active participation*, where the volunteers annotate and classify entities among other types of activity. A whole description of the typology of the CS activities in Zooniverse has been published by Michalak (2015).

Chimp & See, has been started in 2015 by the Max Planck Institute for Evolutionary Anthropology as one of the projects on the Zooniverse platform. The goal of the project is to gain a better understanding of chimp culture, population size and demographics in specific regions of Africa. The type of activity as an annotation and classification task, where the volunteers identify species on video. The web-based platform features a video player with interactive tools to annotate parts of the video, for example to highlight certain behavioural patterns of chimpanzees. The volunteers are not required to have a certain knowledge in the field. They receive instructions for the annotation in an interactive web tutorial (Figure 14).



Baboons

Guinea baboon:
This species of large baboon has red-brown hair and a dark, hairless face.

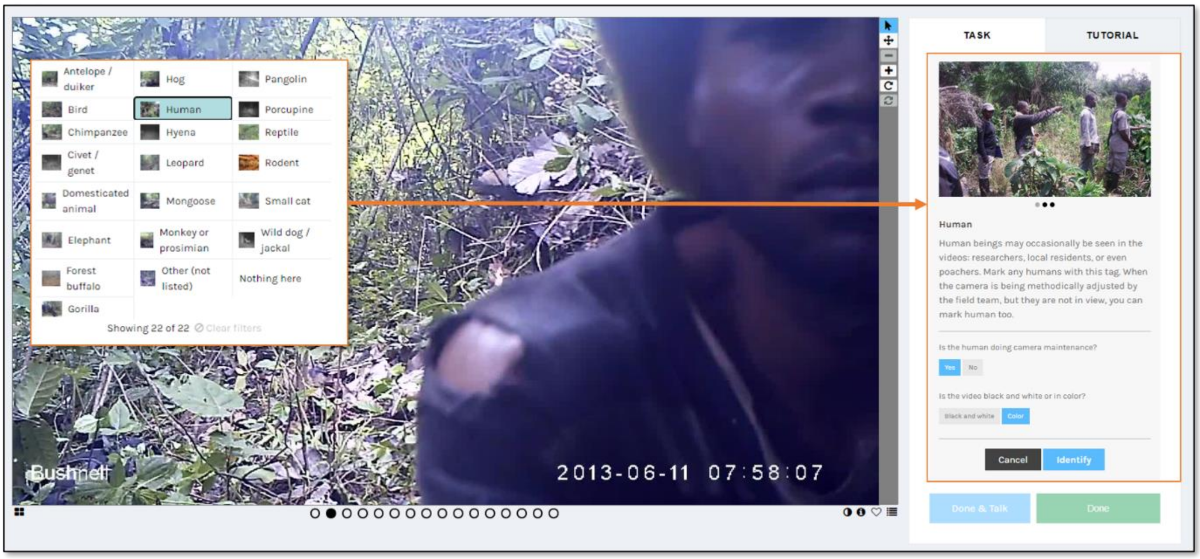
Guinea baboons may be found in one or more sites in region A.

Figure 14: Web tutorial with instructions to train volunteers to conduct the annotation and classification task in Chimp & See.

The annotation task is then performed on video material in a web-based video player. The volunteer watches a video sequence and decides to tag it regarding if something could be observed and if, which species. Therefore, the user might select from 22 predefined labels for different species (including the label "Nothing here"). Figure 15 shows one of the video sequences, where a member of the field team who maintains the technical equipment could be observed in the video. Therefore, the label "Human"

[15] Citizen Science Alliance, Zooniverse (2020): https://www.zooniverse.org/. Retrieved: 2020-07-08.
[16] CSA, Zooniverse - Projects: https://www.zooniverse.org/projects. Retrieved: 2020-06-20.

has been selected. After a selection has been made, the user interface further queries the volunteer to clarify certain things, such as the maintenance of the spotted worker or even the colour of the video (black and white or colour).



Human

Human beings may occasionally be seen in the videos: researchers, local residents, or even poachers. Mark any humans with this tag. When the camera is being methodically adjusted by the field team, but they are not in view, you can mark human too.

Is the human doing camera maintenance?

Yes No

Is the video black and white or in color?

Black and white Color

Figure 15: User interface for the annotation and classification task in the Chimp & See project. The volunteer confirms the annotations and receives further queries.

In addition to the web tutorial, a discussion forum ("talk pages") can be facilitated for clarifications, discussions, guides, help seeking, but also for community building and announcements. Within the project, 3 different roles of users interact with each other. The roles are prescribed by the system, for example moderators and scientists, which are members of the (official) project team. The moderators are usually assigned by the project lead, for instance, the Max Planck Institute. Volunteer is the third role, which is assigned by default to the users who volunteer in the citizen science activity.

5.2 Goals and Indicators

The Zooniverse projects are per se crowdsourced activities with active participation (Haklay, 2013). From an external perspective, it is neither obvious nor trivial, how the given roles correspond to the actual roles the volunteers take in the discourse or communication in the forum (talk pages). The behavioural patterns and communication structures that can be observed might give some clues about the general structure and how the analysis of the discourse can be attributed as a characteristic of a citizen science project. The goal of this work is to create an initial and exploratory study to characterize a CS project to infer more general mechanisms that can be applied to a variety of projects.

To characterize the communication structure and particularly the role of certain users in the discourse, we use centrality measures such as (weighted) In-Degree, (weighted) outdegree, and eigenvector centrality to measure different types of importance (see section 4). Additionally, descriptive statistics about the distribution across the different roles give important insights about the communication structure, particularly who initiates communication and who replies to enquiries.

5.3 Data Collection and Processing

The forum data of the Chimp & See talk pages[17] have been processed to create a dataset for the analysis. For this purpose, 3218 forum threads with 24531 individual posts have been processed. The forum involved a total of 575 unique user accounts, which represents 10.1% of all the active

[17] CSA, Zooniverse - Chimp & See Talk Pages: https://talk.chimpandsee.org/#/boards. Retrieved: 2020-06-20.

volunteers of the Chimp & See project[18]. The number of accounts splits up in the following (system) roles: 8 moderators, 25 scientists, 542 volunteers. The time window of the forum discussion that has been processed was from 2015-04-03 until 2019-07-05. Three sub forums have been analysed: *help*, *science,* and *chat* ("community building"). The average length of a discussion thread is 6.5 posts, with a variation depending on the specific forum (help: 5,7; chat: 5; science: 8,7). The overview page that contains all sub forums served as a seed for the crawler.

For the technical implementation of the data collection and processing, the following pipeline with the techniques and tools described in section 2.1.1 has been created using the Python programming language. The crawler uses Selenium[19] with a headless browser to access page content from the forum and BeautifulSoup[20] to extract relevant data (paging for multi-page threads included). Table 3 shows all the fields extracted using the crawler.

Afterwards, the NetworkX library (cf. section 4.5) has been used to extract the social network from the retrieved forum data. For an ex-post analysis (centrality measures) and visualisation of the extracted network with dynamic graph layouting, Gephi has been used. The network is created as a directed graph, where nodes represent the users, and the colouring of nodes indicates the prescribed role (moderator, scientist, volunteer). Edges between users are established either when a user replies to a post, or when a user mentions someone using the designated @ character. A weight is assigned to each edge representing the number of replies and mentions. The weight of a node is the outdegree. To characterize the dynamics, time slices are selected for each year.

Table 3: Data and descriptors of the extracted forum posts.

Data	Description
Board Category	Sub-Forum
Post Number	Sequential post id
Time stamp	Date of the post
Title	Title of the post
User Type/Role	Moderator | Scientist | Volunteer
Username	Username
Response To	Direct response
References to Users	Mentioning (@)
References to Objects/Tags	Hashtag / Object references
Post Content	Textual representation of the post
URL	URL of the thread
Raw HTML	Raw HTML

[18] CSA, Zooniverse. The Chimp & See project has 5686 active users according to https://www.zooniverse.org/projects/sassydumbledore/chimp-and-see. Retrieved: 2020-07-09.

[19] Baiju Muthukadan, Selenium (2018): https://selenium-python.readthedocs.io/.

[20] Leonard Richardson, BeautifulSoup (2020): https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

5.4 Results

To address the dynamic character of forum discussions, we investigate the number of forum posts per time slice. The extracted posts are aggregated per year to gain an overview about the posts over time. Figure 16 shows that the number of posts decreases over time, in total and for every sub forum. This might have several reasons. For example, a decreased post activity in the help forum might be justified with a growing knowledge base of already answered questions that new users search before posting. This might also affect the science forum, as this encompasses announcements for competitions and contests, which generate a lot of traffic. Potentially, more recruitment and retainment efforts from the project team have been made from the start of the project. However, methods of content analysis can be useful in future work to depict this issue.

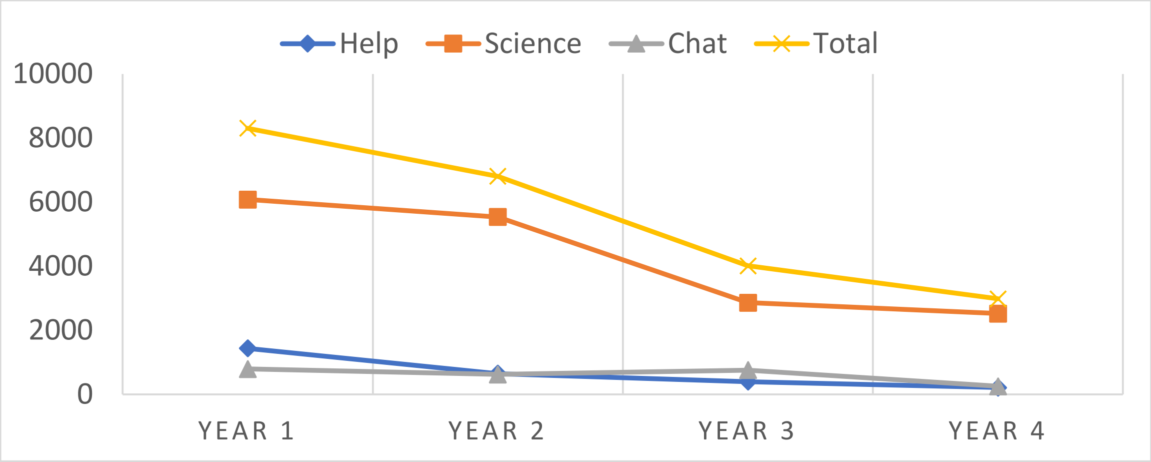


Figure 16: Posts over time per role and in total.

The decreasing number of posts might also indicate a decreased momentum in the project. In this dataset, 41% of the posts are by moderators, 34% by scientists and 25% by volunteers. “Volunteer” is the role that is dedicated to citizen scientists who are not part of the official project team. However, the distribution of posts is not equal. As expected in such communities, a few forum users contribute the most to the discussion. The top 5% (in total 28) of the users that contribute the most are 9 scientists, 8 moderators and 11 volunteers. Figure 17 illustrates this long tail effect in a histogram, where all users are ordered according to their post contribution on the domain axis. Those top 5% contribute to around 87% of the forum posts, nearly 80% of the in-degrees and approximately 50% of the outdegrees. These findings are in line with the results from other research about crowdsourcing in the context of citizen science programmes, where a few superusers are responsible for most of the work (Herodotou, Aristeidou, Miller, Ballard & Robinson, 2020).

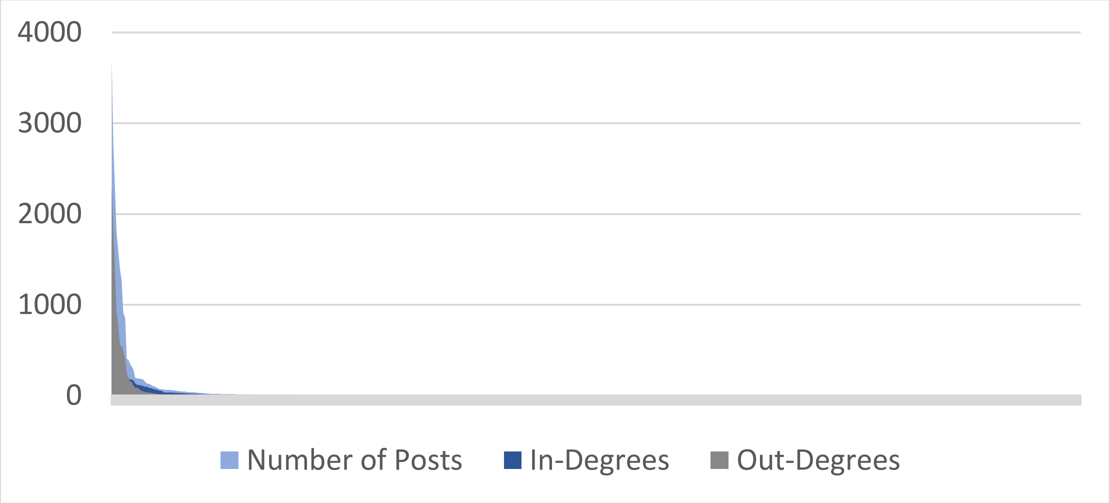


Figure 17: Distribution of posts and degree-centralities over the users. The domain axis contains the users, ranked according to the number of posts.

The distribution of roles between the top users is equal (8 moderators, 9 scientists and 11 volunteers). It is remarkable that there are also volunteers who are highly motivated to contribute to the discussions. For future analyses, it may be interesting to investigate the incentives for volunteers to participate in such an extent. Following the analogy of PageRank or eigenvector centrality, it might induce a feeling of importance to communicate with people of high reputation. Therefore, we investigate the direction of communication, in particular with respect to who references whom in terms of the affiliated forum role. The following analysis considers the whole communication structure and is not restricted to the superusers.

Figure 18 shows the relative amount of references, normalized by the total number of references over all user roles in the specific forum. A reference is either a direct mentioning of a user (with the '@' symbol) or a post reply in the thread structure of the discussion forum. In the help forum, most references are made by moderators. When volunteers seek for help, they typically do not know whom to address, whereas moderators might point to scientists and / or mention the user who asked a question. Further analysis of communication patterns is needed to ground this. Volunteers typically reference moderators to say "thanks" regarding the prior reply to the help seeking. In the chat board, the references are similar, except that there is less need for moderators to direct to scientists, which explains the lower bar for this reference. Volunteers are mentioned quite often in this forum, usually because the moderators and scientists welcome them. The chat forum is sometimes used by users to introduce themselves. This can serve for further analyses to deepen the understanding of the incentives and backgrounds of volunteers, and particularly their motivation to participate in CS activities.

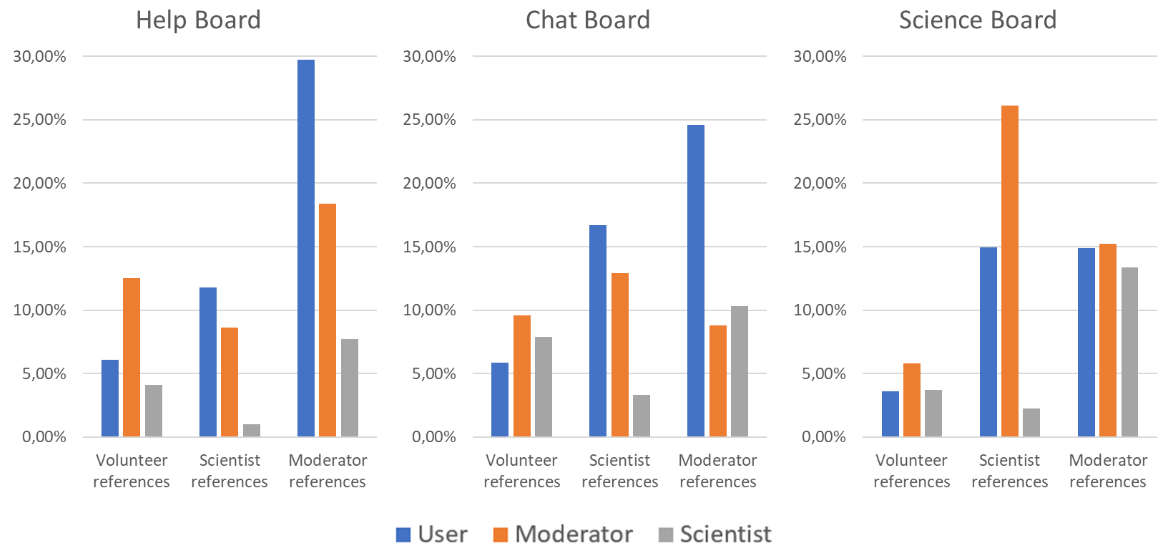


Figure 18: Who references whom depending on the specific board (sub forum).

These distributions reveal that particularly moderators play an important role in the mediation of citizen science activities. Figure 19 shows the average eigenvector centrality per role. Compared to moderators, volunteers have a very low centrality on average. Eigenvector centrality is a measure of importance or influence of nodes in a network (cf. section 4.2). Nodes that are connected to other high-scoring nodes, also have a high eigenvector centrality. For directed graphs like the underlying network, both in- and out-edges are incorporated into the calculations. Therefore, such a high centrality score for the moderators (on average) indicates that they are important to the whole communication, both referencing and being referenced.

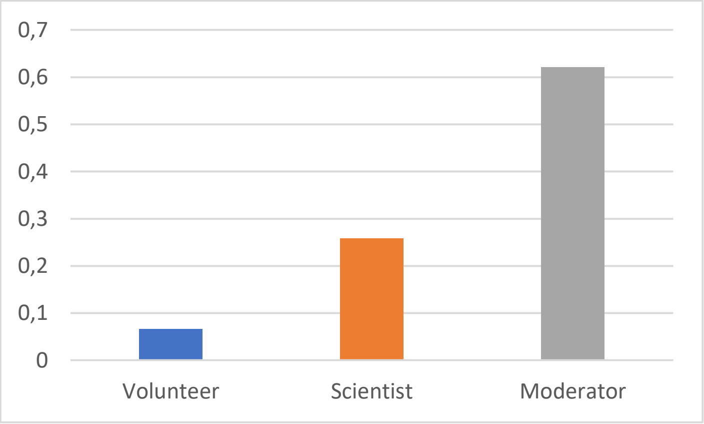


Figure 19: Average (normalized) eigenvector centrality grouped by role.

In general, the moderators have the highest centrality values in the Chimp & See community. Where the top 3 moderators (*AnLand*, *Boleyn* and *ksigler*) have eigenvector centralities ranging from 0.8 to 1.0, the top 2 and very exceptional volunteers are still below (*Snorticus:* 0.74; *Batfan:* 0.61). Scientists such as *kristinahaverkamp* reach a high centrality value because they announced contests in the forum, which obviously lead to a communication structure that is beneficial for the score. Figure 20 shows a part of the extracted network with the different roles. The node size indicates the outdegree (cf. section 4.2) quantifying a branching factor of the node, particularly how many other users are referenced or mentioned.

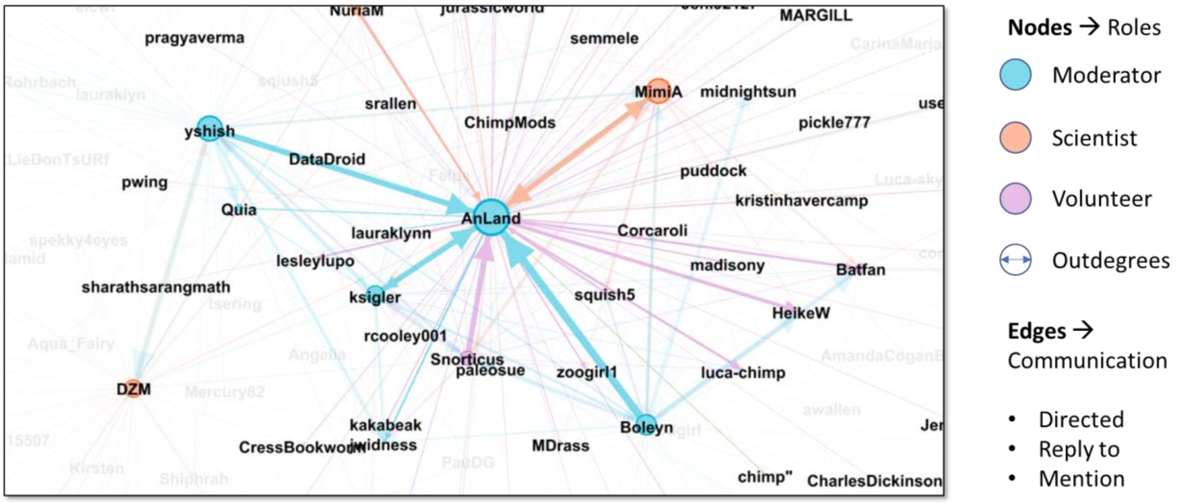


Figure 20: Social Network constructed using the previously mentioned approaches. Node size indicates outdegree, edges connect users when they replied or mention someone.

The previously outlined example illustrates a first case study about how to assess communication structures in online communities of citizen science projects. Using the methods of SNA shows the importance of certain roles in the mediation of citizen science activities. As a singularity, some of the researchers have a higher outdegree, which could be identified when investigating their actual contributions. Contest announcements trigger that many users reply to the post or mention the scientist who announced the contest, which implies a higher outdegree.

To better understand the specific roles and the contributions the users and volunteers make, method of text analysis (see section 3) might help to gain more insights about knowledge-building, artefact creation and the way in which the members of the online community interact, particularly regarding the role (volunteers vs. professional scientists vs. moderators). Semantic technologies help to better understand the artefacts that are created within the communities, both in the discourse and the core activity itself. However, different online communities might address certain issues in another way. To effectively analyse and compare the discourse across the different citizen science projects, epistemic network analysis (cf. section 3.5) might produce more and deeper insights.

Section 6: Conclusion and Recommendations

The various examples interspersed with the description of general methods and computational analysis techniques as well as the longer example elaborated in section 5 illustrate what kinds of insights, we can expect from applying such techniques to sources and traces from CS project activities. In the prior exposition of methods, most of the applications mentioned were related to collaborative knowledge building in educational settings or in the context of scientific research activities. This is consistent with the general orientation of CS Track that includes looking at CS activities in the light of understanding the underlying types of collaboration and knowledge creation. This perspective comprises semantic and pragmatic aspects of scientific discourse as well as more structural analyses that focus on connections and networking inside the projects, between different projects as well as the information flow between CS projects and other institutions in society (e.g. public media or education).

This perspective implies that we will work on different levels of scale and granularity: Information mining together with web crawling and scraping techniques can address a larger number of projects, for instance to assess their overlap with certain scientific disciplines. Also, Twitter-based analyses relying on follower or retweet relations would be on macro or meso level of granularity addressing a number of projects and their interaction with social actors.

The content level analysis ranges from named entity recognition (persons, places, institutions) over the extraction of keywords indicating a disciplinary reference until the identification of themes in a discourse, possibly in relation to different types of actors within the projects (e.g. comparing volunteers to professional scientists). This latter point requires an integration of semantic and structural methods. “Epistemic Network Analysis” (ENA) is a promising candidate for such in-depth analyses. Here, we have to consider that ENA works with predefined “codes” indicating specific themes. We may, however, also want to allow for exploring and “digging out” new themes which can be supported by topical analyses, e.g. based on LDA.

Computational analytics (WP3) together with the database-related activities in WP2 from the core of technology-oriented elements of CS Track. In its first versions, the WP2 database will serve as an aggregated catalogue that assembles basic information associated with CS projects such as type of actors, topics of interest, regional distribution, etc. All that is usually required to start analysing a specific project is the project-specific URL. In the other direction, however, we expect the analytics to feed into the database certain descriptors that are not explicitly represented in the projects’ web pages. In this sense, our analytics tools and methods will gradually enrich the database. The design of the database has considered extensibility as an important feature, also for this purpose. One of the descriptors that we expect to be instantiated in this way is the relation of projects to different scientific disciplines. Here, we do not assume that this will be a unique value per project but that most projects will be related to different reference disciplines. Another descriptor to be filled would be the “outreach” as measured by their mentions in classical public media or in micro-blogging channels (including Twitter) and news feeds.

The envisaged computational analyses usually rely on public (web-based) sources. However, we are aware that there are still potential ethical issues related to the use of this information. It is well known that personal information can be found in public social media sources, and that this type of information can be used for individual profiling. Within our ethics framework, we exclude this kind of usage, i.e. we would not aim at profiling individuals based on their CS-related web traces. Instead, our perspective will focus projects or even aggregate over several projects, e.g. by region or application area. This does not exclude that individuals would be mentioned in a specific leading or representative role especially in the project database. However, we would not search specifically for more information about such individuals in the sense of personal attributes in accordance with the statement about data protection of personal data from deliverable D8.2 (page 7).

In the overall picture of CS Track, analytics methods are only one ingredient in the analysis and evaluation of CS activities. The analytics results will be inter-related and integrated with empirical-interpretative approaches originating from social studies practices in a triangulation approach. This integration happens in WP4, where also the empirical-interpretative perspective with quantitative and qualitative approaches is elaborated. The triangulation takes place before results are transformed into strategic recommendations or decision support for stakeholders. Selected results of analytics (WP3) and further analysis (WP4) will be made available to external stakeholders on the community platform and in the e-magazine (WP5). From a research perspective, analytics results as well as the experience with applying and adapting analytics methods to CS projects activities will be the source of scientific contributions in the context of research on electronic communities and collaboration with technologies with venues such as ACM Group or CollabTech.

We are convinced that applying analytics to CS activities will deepen our understanding of interactions and collaborations in CS, providing valuable insights into individual journeys throughout the CS project lifetime and identifying factors that hinder or support participation and interaction. The results will have implications for CS practitioners and help their work by generating recommendations on how to consider those factors in the design and facilitation of CS activities to best support volunteers in their citizen science experience.

References

Aboaoga, M., & Ab Aziz, M. J. (2013). Arabic person names recognition by using a rule based approach. Journal of Computer Science, 9(7), 922.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, 722-735. Springer, Berlin, Heidelberg.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). New York: ACM press.

Bail, C. A. (2016). Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences*, *113*(42), 11823-11828.

Batagelj, V. (2003). Efficient algorithms for citation network analysis. arXiv preprint cs/0309023.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network Analysis in the Social Sciences. Science, 323(5916), 892-895.

Boyd, D. M., & Ellison, N. B. (2008). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210-230.

Brandes, U., Marshall, M. S., & North, S. C. (2000). Graph data format workshop report. In International Symposium on Graph Drawing (pp. 407-409). Springer, Berlin, Heidelberg.

Cai, Z., Eagan, B., Dowell, N. M., Pennebaker, J. W., Shaffer, D. W., & Graesser, A. C. (2017). Epistemic network analysis and topic modeling for chat data from collaborative learning environment. *Proceedings of the 10th International Conference on Educational Data Mining*, EDM 2017, 104–111.

Carley, K. M. (1997). Network text analysis: The network position of concepts. *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*, *4*, 79-100.

Carreras, X., Màrquez, L., & Padró, L. (2003). A simple named entity extractor using AdaBoost. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 (pp. 152-155).

Cimiano, P., Schultz, A., Sizov, S., Sorg, P., & Staab, S. (2009). Explicit Versus Latent Concept Models for Cross-Language Information Retrieval. In IJCAI (Vol. 9, pp. 1513-1518).

Collier, W., Ruis, A. R., & Shaffer, D. W. (2016). Local versus global connection making in discourse. In C. K. Looi, J. L. Polman, U. Cress, & P. Reimann (Eds.), *Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS)* Vol. 1, 426-433. Singapore: International Society of the Learning Sciences.

Cruz, S. M. A., Lencastre, J. A., Coutinho, C. P., José, R., Clough, G., & Adams, A. (2017). The JuxtaLearn process in the learning of maths' tricky topics: Practices, results and teacher's perceptions.

Csanadi, A., Eagan, B., Shaffer, D., Kollar, I., & Fischer, F. (2017). Collaborative and individual scientific reasoning of pre-service teachers: New insights through epistemic network analysis (ena). *Computer-Supported Collaborative Learning Conference, CSCL*, 1, 215–222.

Daiber, J., Jakob, M., Hokamp, C., & Mendes, P. N. (2013, September). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, 121-124.

Derczynski, L. (2016). Complementarity, F-score, and NLP Evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), 261-266.

de Solla Price, D. J. (1965). Networks of scientific papers. Science, 149 (3683), 510–515.

Doreian, P., Batagelj, V., & Ferligoj, A. (2002). *Positional analyses of sociometric data*. University of Ljubljana, Institute of Mathematics, Physics and Mechanics, Department of Mathematics.

Diesner, J., & Carley, K. M. (2008). From Texts to Networks. *CASOS Summer Institute 2008*.

Erdős, P., & Rényi, A. (1959). On Random Graphs. *Publicationes Mathematicae*. 6, 290–297.

Erkens, M., Daems, O., & Hoppe, H. U. (2014). Artifact analysis around video creation in collaborative STEM learning scenarios. In *2014 IEEE 14th International Conference on Advanced Learning Technologies* (pp. 388-392). IEEE.

Esuli, A., & Sebastiani, F. (2010). Evaluating information extraction. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 100-111. Springer, Berlin, Heidelberg.

Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, *29*(2), 1-34.

Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000). Rule-based named entity recognition for Greek financial texts. In Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000) (pp. 75-78).

Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3), 75-174.

Franceschet, M. (2011). PageRank: Standing on the shoulders of giants. Communications of the ACM, 54(6), 92-101.

Gabrilovich, E., & Markovitch, S. (2006). Computing semantic relatedness of words and texts in Wikipedia-derived semantic space (No. CS Technion report CIS-2006-04). Computer Science Department, Technion.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In IJcAI (Vol. 7, pp. 1606-1611).

Gašević, D., Joksimović, S., Eagan, B. R., & Shaffer, D. W. (2019). SENS: Network analytics to combine social and cognitive perspectives of collaborative learning. *Computers in Human Behavior*, *92*(July 2018), 562–577. https://doi.org/10.1016/j.chb.2018.07.003

Gee, J. P. (1999). *An introduction to discourse analysis*. Routledge, New York.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12), 7821-7826.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. *Briefings in bioinformatics*, *15*(5), 788-797. http://doi.org/10.1093/bib/bbt026.

Göhnert, T., Harrer, A., Hecking, T., & Hoppe, H. U. (2013). A workbench to construct and re-use network analysis workflows: concept, implementation, and example case. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 1464-1466). ACM.

Gottron, T., Anderka, M., & Stein, B. (2011). Insights into explicit semantic analysis. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 1961-1964).

GraphML Working Group. (2009). GraphML Specification.
http://graphml.graphdrawing.org/specification.html. Retrieved July 30, 2020.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political analysis, 21(3), 267-297.

Gurevych, I., Mühlhäuser, M., Müller, C., Steimle, J., Weimer, M., & Zesch, T. (2007). Darmstadt knowledge processing repository based on uima. In *Proceedings of the First Workshop on Unstructured Information Management Architecture at Biannual Conference of the Society for Computational Linguistics and Language Technology*, Tübingen, Germany, page 89.

Hakimov, S., Oto, S. A., & Dogdu, E. (2012). Named entity recognition and disambiguation using linked data and graph-based centrality scoring. In Proceedings of the 4th international workshop on semantic web information management (pp. 1-7).

Haklay, M. (2013). Citizen science and volunteered geographic information: Overview and typology of participation. In Crowdsourcing geographic knowledge (pp. 105-122). Springer, Dordrecht.

Halatchliyski, I., Hecking, T., Göhnert, T., & Hoppe, H. U. (2014). Analyzing the path of ideas and activity of contributors in an open learning community. *Journal of Learning Analytics*, JLA, 1(2), 72-93.

Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 1262-1273.

Hecking, T., Göhnert, T., Zeini, S., & Hoppe, H. U. (2013). Task and time aware community detection in dynamically evolving social networks. *Proceedings of the International Conference on Computational Science*, Barcelona, Spain, 5-7 June 2013.

Hecking, T., Ziebarth, S., & Hoppe, H. U. (2014). Analysis of dynamic resource access patterns in a blended learning course. *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge*, Indianapolis, Indiana. 173-182.

Hecking, T., & Hoppe, H. U. (2015). A Network Based Approach for the Visualization and Analysis of Collaboratively Edited Texts. In *VISLA@ LAK*, 19–23.

Hecking, T., Dimitrova, V., Mitrovic, A., & Hoppe, H. U. (2017). Using network-text analysis to characterise learner engagement in active video watching. In *ICCE 2017 Main Conference Proceedings*, 326-335. Asia-Pacific Society for Computers in Education.

Hecking, T., & Leydesdorff, L. (2019). Can topic models be used in research evaluations? Reproducibility, validity, and reliability when compared with semantic maps. *Research Evaluation*, 28(3), 263-272.

Herodotou, C., Aristeidou, M., Miller, G., Ballard, H., & Robinson, L. (2020). What Do We Know about Young Volunteers? An Exploratory Study of Participation in Zooniverse. Citizen Science: Theory and Practice, 5(1).

Himsolt, M. (1997). GML: A portable graph file format (p. 35). Technical report, University Passau.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57).

Hoppe, H. U., Engler, J., & Weinbrenner, S. (2012). The impact of structural characteristics of concept maps on automatic quality measurement. In *Proceedings of the 10th international conference of the learning sciences* (ICLS). Sydney, Australia: ISLS.

Hoppe, H. U., Göhnert, T., Steinert, L., & Charles, C. (2014). A web-based tool for communication flow analysis of online chats. LAK 2014 Workshop Proceedings.

Hoppe, H. U. (2017). Computational methods for the analysis of learning and knowledge building communities. *Handbook of learning analytics*, 23-33. SOLAR.

Hummon, N. P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. Social networks, 11(1), 39-63

Kasperowski, D., & Hillman, T. (2018). The epistemic culture in an online citizen science project: Programs, antiprograms and epistemic subjects. *Social Studies of Science*, *48*(4), 564–588. https://doi.org/10.1177/0306312718778806.

Krempel, L. (2005). Visualisierung komplexer Strukturen: Grundlagen der Darstellung mehrdimensionaler Netzwerke. Campus Verlag.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In Proceedings of NAACL-HLT (pp. 260-270).

Latour, B. (2005). Reassembling the Social - An Introduction to Actor-network-theory. Oxford University Press.

Lawson, R. (2015). Web scraping with Python. Packt Publishing Ltd.

Lehmann, S., Schwartz, M., & Hansen, L. K. (2008). Biclique communities. *Physical review. E, Statistical, nonlinear, and soft matter physics*, *78*(1 Pt 2), 016108. https://doi.org/10.1103/PhysRevE.78.016108.

Leydesdorff, L. (2001). The challenge of scientometrics: The development, measurement, and self-organization of scientific communications. Universal-Publishers.

Malzahn, N., Hartnett, E., Llinás, P., & Hoppe, H. U. (2016). A smart environment supporting the creation of juxtaposed videos for learning. In *State-of-the-Art and Future Directions of Smart Learning*, 461-470. Springer, Singapore.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Scoring, term weighting and the vector space model. Introduction to information retrieval, 100, 2-4.

Manske, S. (2020). Managing Knowledge Diversity in Computer-Supported Inquiry-Based Science Education. https://doi.org/10.17185/duepublico/71585

Marsh, E., & Perzanowski, D. (1998). MUC-7 evaluation of IE technology: Overview of results. In Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998.

Mendes, P. N., Jakob, M., & Bizer, C. (2012). DBpedia: A multilingual cross-domain knowledge base. In *LREC*, 1813–1817. Citeseer.

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems* (pp. 1-8).

Michalak, K. (2015). Online localization of Zooniverse citizen science projects–on the use of translation platforms as tools for translator education. Teaching English with Technology, 15(3), 61-70.

Morwal, S., Jahan, N., & Chopra, D. (2012). Named entity recognition using hidden Markov model (HMM). International Journal on Natural Language Computing (IJNLC), 1(4), 15-23.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3-26.

Newman, M. E. J., Barabási, A.-L., & Watts, D. J. (2006). The Structure and Dynamics of Networks. Princeton Studies in Complexity. Princeton University Press.

Nothman, J., Murphy, T., & Curran, J. R. (2009). Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp. 612-620).

Palla, G., Barabasi, A. L., & Vicsek, T. (2007). Quantifying social group evolution. Nature, 446(7136), 664-667.

Palla, G., Derenyi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435(7043), 814-818.

Paranyushkin, D. (2019). InfraNodus: Generating insight using text network analysis. In *The World Wide Web Conference*, 3584-3589.

Popping, R. (2000). Computer-assisted text analysis. Sage.

Rohde, M., & Shaffer, D. W. (2004). Us, ourselves and we: Thoughts about social (self-) categorization. *Association for Computing Machinery (ACM) SigGROUP Bulletin*, *24*(3), 19-24.

Rosa, H., Carvalho, J. P., Astudillo, R., & Batista, F. (2018). Page rank versus katz: is the centrality algorithm choice relevant to measure user influence in Twitter? In Interactions Between Computational Intelligence and Mathematics (pp. 1-9). Springer, Cham.

Rule, A., Cointet, J. P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, *112*(35), 10837-10844.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. Auckland, New Zealand: McGraw-Hill.

Scardamalia, M., & Bereiter, C. (1994). Computer support for knowledge-building communities. The Journal of the Learning Sciences, 3(3), 265-283.

Schmitt, X., Kubler, S., Robert, J., Papadakis, M., & LeTraon, Y. (2019). A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 338-343). IEEE.

Segura-Bedmar, I., Martínez, P., & Segura-Bedmar, M. (2008). Drug name recognition and classification in biomedical texts: a case study outlining approaches underpinning automated systems. Drug discovery today, 13(17-18), 816-823.

Shaffer, D. W. (2004). Epistemic frames and islands of expertise: Learning from infusion experiences. *Proceedings of the 6th International Conference of Learning Sciences (ICLS 2004): Embracing Diversity in the Learning Sciences*, 22-26 June 2004, Santa Monica, CA, USA, 473-480. Mahwah, NJ: Lawrence Erlbaum.

Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics*, *3*(3), 9-45.

Shaffer, D. W. (2017). *Quantitative ethnography*. Cathcart Press, Madison, Wisconsin, USA.

Shaffer, D. W., & Ruis, A. R. (2017). Epistemic Network Analysis: A Worked Example of Theory-Based Learning Analytics. *Handbook of Learning Analytics*, 175–187. https://doi.org/10.18608/hla17.015

Siebert-Evenstone, A. L., Arastoopour, G., Collier, W., Swiecki, Z., Ruis, A. R., & Shaffer, D. W. (2016). In search of conversational grain size: Modeling semantic structure using moving stanza windows. Paper presented at the 12th International Conference of the Learning Sciences, Singapore.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, *6*(3), 203-217.

Suthers, D. D., Dwyer, N., Medina, R., & Vatrapu, R. (2010). A framework for conceptualizing, representing, and analyzing distributed interaction. *International Journal of Computer-Supported Collaborative Learning*, *5*(1), 5-42.

Suthers, D. D., & Desiato, C. (2012). Exposing chat features through analysis of uptake between contributions. In *2012 45th Hawaii international conference on system sciences* (pp. 3368-3377). IEEE.

Swiecki, Z., & Shaffer, D. W. (2020). ISENS: An integrated approach to combining epistemic and social network analyses. *ACM International Conference Proceeding Series*, 305–313. https://doi.org/10.1145/3375462.3375505

Taskin, Y., Hecking, T., & Hoppe, H. U. (2019). ESA-T2N: A Novel Approach to Net-work-Text Analysis. In *International Conference on Complex Networks and Their Applications*, 129-139. Springer, Cham.

Triezenberg, H. A., Knuth, B. A., Yuan, Y. C., & Dickinson, J. L. (2012). Internet-based social networking and collective action models of citizen science. *Citizen science: Public participation in environmental research*, 214-225.

Vallabh, P., Lotz-Sisitka, H., O'Donoghue, R., & Schudel, I. (2016). Mapping epistemic cultures and learning potential of participants in citizen science projects. *Conservation Biology*, *30*(3), 540–549.

Was, T., & Skibski, O. (2018, April). An axiomatization of the eigenvector and Katz centralities. In Thirty-Second AAAI Conference on Artificial Intelligence.

Wasserman, S., & Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge.

Watson, D., & Floridi, L. (2018). Crowdsourced science: sociotechnical epistemology in the e-research paradigm. *Synthese*, *195*(2), 741-764.

Weitzman, E., & Miles, M. B. (1995). Computer programs for qualitative data analysis. Sage.

Wu, T., Khan, F. M., Fisher, T. A., Shuler, L. A., & Pottenger, W. M. (2005). Posting act tagging using transformation-based learning. In *Foundations of data mining and knowledge discovery* (pp. 319-331). Springer, Berlin, Heidelberg.