A comprehensive analysis of acknowledgement texts in Web of Science: a case study on four scientific domains

Smirnova, Nina; Mayr, Philipp

Check for
updates

A comprehensive analysis of acknowledgement texts in Web of Science: a case study on four scientific domains

Nina Smirnova[1] · Philipp Mayr[1]

Abstract

Analysis of acknowledgments is particularly interesting as acknowledgments may give information not only about funding, but they are also able to reveal hidden contributions to authorship and the researcher's collaboration patterns, context in which research was conducted, and specific aspects of the academic work. The focus of the present research is the analysis of a large sample of acknowledgement texts indexed in the Web of Science (WoS) Core Collection. Record types "article" and "review" from four different scientific domains, namely social sciences, economics, oceanography and computer science, published from 2014 to 2019 in a scientific journal in English were considered. Six types of acknowledged entities, i.e., funding agency, grant number, individuals, university, corporation and miscellaneous, were extracted from the acknowledgement texts using a named entity recognition tagger and subsequently examined. A general analysis of the acknowledgement texts showed that indexing of funding information in WoS is incomplete. The analysis of the automatically extracted entities revealed differences and distinct patterns in the distribution of acknowledged entities of different types between different scientific domains. A strong association was found between acknowledged entity and scientific domain, and acknowledged entity and entity type. Only negligible correlation was found between the number of citations and the number of acknowledged entities. Generally, the number of words in the acknowledgement texts positively correlates with the number of acknowledged funding organizations, universities, individuals and miscellaneous entities. At the same time, acknowledgement texts with the larger number of sentences have more acknowledged individuals and miscellaneous categories.

✉ Nina Smirnova
ninasmirnova@web.de

✉ Philipp Mayr
philipp.mayr@gesis.org

1 GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany

🖄 Springer

Introduction

Acknowledgments in scientific papers are short texts where the author(s) express "*gratitude towards different types of support received during the research process*"(Alvarez-Bornstein & Montesi, 2021). Kassierer and Angell see acknowledgements as an instrument to "*identify those who made special intellectual or technical contribution to a study that are not sufficient to qualify them for authorship*" (Kassirer & Angell, 1991, p. 1511). Cronin and Weaver (1995) ascribe an acknowledgment alongside authorship and citedness to measures of a researcher's scholarly performance. Together they build "*The Reward Triangle*": a feature that reflects the researcher's productivity and impact (Cronin & Weaver, 1995, p. 173). Diaz-Faes and Bordons argue, "*citations provide a measure of the underlying intellectual influence and foundations of research output*" (Diaz-Faes & Bordons, 2017, p. 577).

Acknowledgments as a rule contain information about technical, instrumental and financial support together with intellectual and conceptual support (Diaz-Faes & Bordons, 2017; Giles & Councill, 2004). The latter fall into the category "*peer interactive communication*" (PIC) (McCain, 2018). Giles and Councill (2004) argue that PIC acknowledgments, in the same way as citations, may be used as a metric to measure an individual's intellectual contribution to scientific work. Acknowledgements of financial support are interesting in terms of evaluating the influence of funding agencies on academic research. Acknowledgments of technical and instrumental support may reveal "*indirect contributions of research laboratories and universities to research activities*" (Giles & Councill, 2004, p. 17599). Thus, acknowledgments not only give information about funding, but they are also able to reveal hidden contributions to authorship and the researcher's collaboration patterns, context in which research was conducted, and specific aspects of the academic work.

In this paper, acknowledgement texts from four scientific domains (economics, social sciences, oceanography and computer science) were gathered from the Web of Science (WoS) database containing the collection of funding acknowledgments. The following acknowledged entities were extracted from the acknowledgement texts and distinguished into six categories: funding agencies (FUND), grant numbers (GRNB), corporations (COR), universities (UNI), individuals (IND), and miscellaneous (MISC), as Fig. 3 demonstrates. An *acknowledged entity* is an object of acknowledgment. Acknowledged entities could be names and surnames of individuals (also including abbreviations of the names), names of institutions and organizations, and numbers and identification numbers of grants, as Fig. 3 shows. A comprehensive analysis was performed on the acknowledgement texts and extracted acknowledged entities.

Most of the previous works on acknowledgement analysis were limited by the manual evaluation of data and therefore by the amount of processed data (Giles & Councill, 2004; Paul-Hus et al., 2017; Paul-Hus & Desrochers, 2019; Mccain, 2017). Furthermore, to our knowledge, previous works on automatic acknowledgment analysis were mostly concerned with the extraction and analysis of funding organizations and grant numbers (Alexandera & Vries, 2021; Kayal et al., 2017) or classification of acknowledgement texts (Song et al., 2020). Moreover, large bibliographic databases such as Web of Science (WoS)[1] and Scopus selectively index only funding information, i.e., names of funding organizations and grant identification numbers. Therefore, to our knowledge, there are no previous works on acknowledgement analysis, which examines a large quantity of acknowledged entities other than funding organizations and grant numbers (Smirnova & Mayr, 2022). In the present paper, we have used a

---

NER model, specifically trained to extract six types of entities described above and conducted an extensive analysis of the great amount of the obtained acknowledged entities. Acknowledged entities are potentially a great tool for the analysis of different aspects of the scientific society. Thus, Thomer and Weber (2014) argue that using named entities can benefit the process of manual document classification and evaluation of the data. Petrovich (2022) used the Acknowledged Entities Network (along with other acknowledgement related networks), containing funding agencies, congresses and conferences, institutions and persons, to develop a method for mapping scientic and scholarly social networks.

Research questions In the present paper, we will investigate two broad research directions: the general acknowledgment trends among different disciplines and the relationships between different variables: entity types, acknowledged entities, scientific domains, length of the acknowledgement text, number of citations.

The first research direction will focus on the research question (RQ) 1:

- **RQ1:** What are the general acknowledgment trends among different disciplines, i.e., which are the most acknowledged organizations, individuals, universities, or corporations? Do they vary between disciplines? How do acknowledgment patterns differ among disciplines and what do they have in common? Does the lenght of acknowledgement vary between scientific domains?

The second research direction will address the research questions 2 and 3:

- **RQ2:** Are there correlations between the acknowledged entities and the scientific domains?
- **RQ3:** Do highly-cited papers have more funding entities acknowledged than paper with fewer citations?

Previous works in acknowledgment analysis show that acknowledgement patterns diverge in different scientific domains (Paul-Hus et al., 2017; Diaz-Faes & Bordons, 2017; Baccini & Petrovich, 2021). In this paper, we expect to find correlations between the scientific domain and the type of acknowledged entity. We assume [similar to Paul-Hus et al. (2017); Diaz-Faes and Bordons (2017)] that social sciences and economics will show the highest number of acknowledgments of peer-interactive communication (acknowledged individuals). Computer science and oceanography are expected to have a high percentage of acknowledged entities carrying information about funders (names of funding organizations and projects, grant numbers) and affiliated universities. Generally, more entities are expected to fall into the FUND and GRNB categories, as WoS stores only acknowledgements containing funding information. It is also expected to find a correlation between acknowledged entity and label, i.e. precise entities will have a precise label, and consequently, there should be a correlation between acknowledged entity and scientific domain.

Background and related work

Currently, there are several major bibliographic databases (Singh et al., 2021). The following major bibliographic databases can be distinguished: Web of Science (WoS), SCOPUS and Dimensions. WoS provides subscription-based access to publisher-independent global

citation databases[2]. WoS contains publications from different scientific fields, and the WoS Core Collection consists of a number of databases.

From 2008 on, WoS started indexing funding information (mainly funding agencies and grant numbers) in its databases[3]. WoS uses information from different funding reporting systems such as researchfish[4], Medline[5] and others. The funding information indexed in WoS is distinguished into three different fields: "Funding Text" (FT), "Funding Agency" (FO) and "Grant Number" (FG). The FT-field contains the full text of the acknowledgement, the FO-field contains the names of organizations acknowledged for their funding contribution, and the FG-field contains grant numbers affiliated with the funding organizations identified in the FO-field (Paul-Hus et al., 2016, pp. 170-171). Figure 1 demonstrates an example of funding information indexed in WoS.

Research in the field of acknowledgements analysis has been carried out since the 1970s. The first typology of acknowledgements was proposed by Mackintosh (1972) (as cited in, Cronin, 1995). The acknowledgement texts found in American Sociological Review were examined and distinguished into three categories: facilities, access to data, help of individuals. McCain (1991) proposed a five-level typology based on the analysis of acknowledgement text from the genetics domain: research related information, secondary access to research-related information, specific research-related communication, general peer-communication, technical or clerical support. Cronin (1995) distinguished acknowledgements to the three broad categories: resource related, procedure-related and concept related. Mejia and Kajikawa (2018) analysed acknowledgement texts from the WoS Core Collection from the robotic research field. Aim of the analysis was to characterize funding organizations by the types of the sponsored research field. Mejia and Kajikawa (2018) developed a four-level classification of research fields: change maker, incremental, breakthrough and matured. Incremental and matured research fields showed the highest number of acknowledged funding organizations. The authors claimed that the classification of funding agencies may benefit the development of funding strategies of policy makers and funding organizations. Paul-Hus and Desrochers (2019) created an extended classification of acknowledgments stored in Web of Science (WoS). About four thousand acknowledgment sentences were manually coded into 13 categories. The analysis of coded data reveals three distinct axes: the contribution, the disclaimers, and the authorial voice (Paul-Hus & Desrochers, 2019, p. 5).

Lewison (1994) analysed WoS SCI papers funded by the Biotechnology Action Programme (BAP) in the period 1986–1993. The aim of the analysis was to investigate the level of multi-nationality of the publications compared to other European Community biotechnology programs. Tollison and Laband (2003) examined asterisk footnotes from three major journals of economics, to investigate acknowledging behaviour in economics. Analysis shows that the most famous economists are acknowledged the most. Giles and Councill (2004) proposed a method for the automated extraction of acknowledged entities from the text and analysis of acknowledgment texts. Research papers from the computer science domain deposited in the CiteSeer digital library were used as the data source. Extracted acknowledged entities were linked to the source articles and to the data from CiteSer's citation index for further analysis. Based on manual analysis of the most acknowledged

---

[2] https://clarivate.com/webofsciencegroup/solutions/web-of-science/.

[3] http://wokinfo.com/products_tools/multidisciplinary/webofscience/fundingsearch/.

[4] https://mrc.ukri.org/funding/guidance-for-mrc-award-holders/researchfish/.

[5] https://www.nlm.nih.gov/bsd/funding_support.html.

entities, Giles and Councill developed a four-level classification: funding agencies, corporations, universities, and individuals. Wang and Shapira (2011) investigated the connection between research funding and the development of science and technology. Articles from the field of nanotechnology published in the period 2008–2009 were included in the study. The analysis showed that "*most of nanotechnology funding is nationally-oriented*" (Wang & Shapira, 2011). International collaborations are mainly indicated by the funding of individual international researchers. Rigby and Julian (2014) studied WoS collection of acknowledgement texts containing funding from European Molecular Biology Organisation or the Human Frontier Science Program published in the period from 2008–2012. Analysis of the patterns indicating over-finance of the particular project, i.g., projects funded by more than one funding organization in a similar area, has greater citation impact. Paul-Hus et al. (2016) made an overview on the funding data stored in WoS. Data from all WoS Core Collection databases was used for the analysis: Science Citation Index Expanded (SCIE), Social Science Citation Index (SSCI), and Arts & Humanities Citation Index (AHCI). SCIE provides a multidisciplinary search across scientific journals, SSCI covers journals in social sciences disciplines, and AHCI contains publications in the fields of arts and humanities[6]. The aim of the research was to help other researchers understand the potential and limitations of funding information. The analysis was performed on all publications included in WoS for the period from 2005 to 2015 and comprised more than 43 million documents. The analysis of distribution of funding texts by year showed that the full coverage of funding text data starts from the year 2009, therefore the collection of funding texts published before 2009 is incomplete. Funding information is not included in every WoS core collection database. SCIE is mostly covered in terms of funding text data. Funding texts are also included in the SSCI starting from 2015, and the AHCI has no funding information indexation. Consequently, very little funding data was found for many humanities fields. Most funding information was found for publications written in English. Publications in other languages had an extremely low contribution of indexed acknowledgments texts. The main limitation of WoS data is that only acknowledgments containing funding information are included. It may bias the analysis of non-funding types of contribution. Acknowledgment patterns are argued to be domain-specific. Paul-Hus et al. (2017) conducted an analysis of acknowledgments from research articles and reviews stored in WoS. The dataset was restricted by articles published in 2015. The aims of the study were to distinguish types of acknowledged contributions, and to find out how they vary by discipline. The results matched the observations of previous research (Paul-Hus et al., 2016). The highest percentage of publications that included acknowledgments came from the biomedical and natural sciences, followed by clinical medicine. The lowest percentage of funding texts came from the social sciences. Thus, differences between frequencies of types of acknowledged entities across disciplines were found. For example, technical support was mostly acknowledged in publications in the natural sciences. Earth and space, professional fields, and social sciences tend to acknowledge peer contribution. Biomedical research showed a higher percentage of funding information. Acknowledgments in Biology are mainly focused on logistic and fieldwork-related tasks. Diaz-Faes and Bordons (2017) observed the prevalence of PIC acknowledgments in the humanities domain. Tian et al. (2021) analysed an acknowledgement network to estimate the influence of the acknowledged person. The acknowledgement network for the analysis was created using the WoS collection of acknowledgement texts from the field of wind energy

---

[6] https://clarivate.com/webofsciencegroup/solutions/web-of-science-core-collection/.

Fig. 1 An example of funding information indexed in WoS

published between 2008 and 2010. The study showed that the acknowledged individual's centrality positively moderates the citation count. On the opposite authors' centrality in the collaboration network is negatively associated with the relationship between acknowledged individual's centrality and citation count of a paper. Baccini and Petrovich (2021) studied acknowledging behaviour by analysing WoS articles from top-five journals of economics. Acknowledgements were assessed from two perspectives. According to the normative account, acknowledgements serve to recognize the contribution of other collaborators. According to the strategic account, acknowledging influential scholars increases the visibility and quality of the paper. The authors argue that both accounts "*should be conceived as partial accounts of the various motivations behind the acknowledging behaviour of researchers*" (Baccini & Petrovich, 2021, p. 34). Their study showed the variations of acknowledgement behaviour among different fields. Rose and Georg (2021) examined informal collaborations in academic research. The authors developed a dataset containing 5000 acknowledgements from 6 high-ranking journals in economics. The dataset analysis revealed generational and gender differences in informal collaborations. The authors argue that "*information derived from networks of informal collaboration allows us to predict academic impact of both researchers and papers even better than information from co-author networks.*"

Materials and methods

Acknowledgement texts from the WoS Core Collection were used for the following analysis. As WoS contains millions of metadata records, the data chosen for the present study was restricted by year and scientific domain. Records from four different scientific domains published from 2014 to 2019 were considered: two domains from the social sciences (social sciences and economics), oceanography and computer science[7]. Four different domains were selected to analyse the interdisciplinary differences in acknowledgement behaviour. Social sciences and economics are closely associated domains, while oceanography is a branch of natural science and computer science is a broad technical scientific domain. In the MinAck project[8], which is the context of this research, it is of interest to study the differences between closely and widely associated scientific domains. For the present study, the acknowledgments corpus was restricted to approximately 200,000 entries. Therefore, approximately 50,000 records were taken randomly from each scientific domain which resulted in the total amount of records in the acknowledgments corpus of 196,875 entries. Table 1 represents the total number of articles (research articles, reviews) in the corpus from each scientific domain restricted by these criteria.

---

[7] The full list of the disciplines selected for the present study can be found in the Appendix 1.
[8] https://kalawinka.github.io/minack/.

Extraction and classification of the acknowledged entities was performed using the Flair-NLP framework (Akbik et al., 2019). The reason to use our own software is mainly contingent on problems with WoS funding information indexing: only information about funders is included, i.e., individuals are not indexed. Furthermore, the existing indexing of funding organizations is incomplete as Fig. 2 demonstrates. Furthermore, indexed funding organizations are not divided into different entity types like universities, corporations, etc.

The choice of categories was inspired by Giles and Councill's (2004) classification: funding agencies (FUND), corporations (COR), universities (UNI), and individuals (IND). For the present study, this classification was enhanced with the MISC (miscellaneous) and grant numbers (GRNB) categories. The GRNB category was adopted from WoS funding information indexing. The entities in the miscellaneous category could provide useful information but cannot be ascribed to other categories, e.g., names of projects and names of conferences. Fig. 3 demonstrates an example of acknowledged entities of different types.

The Flair Named Entity Recognition (NER) model with Flair Embeddings[9] (Akbik et al., 2018) was used to build a NER tagger (Smirnova & Mayr, 2022). The Flair Embeddings model uses stacked embeddings, i.e., a combination of contextual string embeddings with a static embeddings model [in our case GloVe (Pennington et al., 2014)]. Contextual string embeddings is a character based contextual string embeddings method proposed by Akbik et al. (2018). This approach will generate different embeddings for the same word depending on its context. Our model was trained with a training corpus of 654 sentences extracted from the acknowledgement texts stored in WoS. Selection criteria for the training corpus were similar to the ones used for the acknowledgments corpus, i.e., records were restricted by year, scientific domain, article type, and language. The total accuracy[10] of the model is 0.77. The model is able to recognize six types of entities: FUND, GRNB, IND, COR, UNI and MISC. Some entity types (IND and GRNB) showed very good F1-Scores[11] over 0.9[12] (Smirnova and Mayr, 2022).

Data disambiguation

After reviewing the first analysis of entities retrieved from the acknowledgements corpus, we realized that acknowledged entities need to be disambiguated for a more plausible analysis. Some entities have more than one writing variant as Example 1 demonstrates. Ideally these variants should be reduced to one representative.

---

[9] The model can be tested via an online demo https://colab.research.google.com/drive/1Wz4ae5c65VDWan Y3Vo-fj__bFjn-loL4?usp=sharing.

[10] In our case accuracy computes as average of F1 scores of each entity type (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html).

[11] Precision, Recall, and F1-Score are metrics to evaluate the model's accuracy. Precision calculates the percentage of items that the model marked as positive that are actually positive, i.e., the percentage of labels assigned by the model that matches the gold standard. The gold standard consists of a set of human annotations. Recall measures how many items that should have been marked as positive were actually marked as positive. The F1-Score is "*a weighted harmonic mean of precision and recall*" (https://web.stanford.edu/~jurafsky/slp3/4.pdf).

[12] F1-Score can take values from 0 to 1, where 1 indicates the perfect values of precision and recall.

Table 1 Number of articles (research articles, reviews) selected for the analyzed acknowledgment corpus in each scientific domain

Scientific domain	Number of records in our acknowledgments corpus
Oceanography	49,420
Economics	49,487
Social sciences	49,211
Computer science	48,757

Support for this work was provided in part by the National Institute of Mental Health (R01 MH071589 to LP.) and a fellowship of the Japan Society for the Promotion of Science (to TY). We thank Dr. Steven Most for sharing pictures from his previous study. We also thank Brenton McMenamin and Jong Moon Choi for discussions of this work.

Fig. 2 An example of WoS indexing problems. Entities indexed in WoS are marked yellow and with a dotted underline. Funding agencies not indexed in WoS are marked blue and with a double underline. Individuals (also not indexed in WoS) are marked green and with a wavy underline

Fig. 3 An example of acknowledged entities. Each entity type is marked with a distinct color

IND : person   **UNI** : university
FUND : funding organization   **COR** : corporation
GRNB : grant number   **MISC** : miscellaneous

(1) Jan De Houwer is supported by Methusalem Grant BOF09/01M00209 of Ghent University and by the Interuniversity Attraction Poles Program initiated by the Belgian Science Policy Office (IUAPVII/33).
(2) Data on Anthem Blue Cross PPO enrollees were provided by Anthem, Inc.

Example 1

- National Science Foundation.
- NSF.
- National Science Foundation (NSF).

The second problem that arose was that some different entities have identical abbreviations as Example 2 demonstrates.

Example 2

- Australia Awards Scholarship   AAS
- African Academy of Sciences   AAS

Misspelling problems (Example 3) were existent for all entity types.

Example 3

- National Nature Science Foundation of China
- Natural National Science Foundation of China

To solve these problems, we created a disambiguation dataset containing frequent funding organizations and universities and their abbreviations. We analysed not-disambiguated data to find as much as possible variants of one entity, as Table 2 shows.

Furthermore, all entities in the model output dataset of FUND, UNI and MISC categories were compared to the entity names (column "text" in Tabel 2), and abbreviations (column "abbreviation" in Table 2) from the disambiguation datasets using the Levenshtein distance. The Levenshtein distance is a similarity measure that shows how similar two strings are[13]. We used the Python fuzz.ratio function[14] which calculates the Levenshtein distance similarity ratio between the two strings. The disambiguation corpus containing entities from the FUND category was used for the FUND and MISC categories. Meanwhile, the disambiguation corpus containing names of universities was used for the UNI category. For entity names, entries with the fuzz.ratio value more than 93 were replaced with the unified writing variant (the column "disambiguated form" in Tabel 2). Thus, for entities in Example 3, it would be "National Science Foundation (NSF)". For abbreviations, entries with fuzz.ratio value of more than 99 were considered as different variants of one abbreviation. As in this case, only entities with different upper- and lower-case writing should have been marked as one entity. Entities, for which no matching patterns were found, remained unaltered.

The problem of various writing variants of the same entity also occurred in the COR category. In this case, a different approach was used. Tests showed that all variants of one entity could be found the most successfully using the fuzz.partial_ratio function. The partial_ratio picks the shortest string from the two compared strings and matches it with all substrings of the same lengths from the second string. All the entities labeled COR were compared to each other using fuzz.partial_ratio. Entries with a partial ratio value greater than 96 were identified as one entry. That fuzz.partial_ratio value was determined by running tests on different writing variants of different COR entities.

The problem with abbreviations that are similar for different entities was solved during the creation of the disambiguation dataset. Duplicated abbreviations were excluded from the disambiguation dataset. That way, if only the abbreviation (e.g., AAS) was in the output of the NER model (without its full name), then it would not be altered and would be placed in the original format into a disambiguated corpus.

To solve the misspelling problem, all entities were compared to each other within their entity types (e.g., FUND with FUND, MISC with MISC, etc.) using Levenshtein distance. Entities with the Levenshtein distance of more than 90 were identified as the same entity. For the IND category, entities with the Levenshtein distance equal to 100 were identified as one category; as in this case, only entities which differ only in upper- and lower-case writing variants (e.g., *John Doe* vs. *john doe*) were considered as different writing variants of the same entity.

Additionally, the top 30 most frequent entities were manually reviewed. Entities mistakenly classified to the wrong categories were placed to the appropriate category in the whole dataset[15]. Furthermore, the IND category had some entities like *Drs.* or *J.*, which do not carry meaningful information, as Table 3 demonstrates. For that reason, individuals containing less than 4 characters (with spaces) were removed from the dataset. The same procedure was conducted with the grant numbers.

---

[13] https://en.wikipedia.org/wiki/Levenshtein_distance.

[14] https://github.com/seatgeek/thefuzz.

[15] For example entity World Bank in economics was distributed between COR and FUND categories. We moved all mentions of that category from the FUND category into the COR category.

Table 2 An example of the disambiguation dataset

Text	Abbreviation	Disambiguated form
Deutsches Klimarechenzentrum	DKRZ	German Climate Computer Center (DKRZ)
German Climate Computer Center	DKRZ	German Climate Computer Center (DKRZ)
German Computing Center	DKRZ	German Climate Computer Center (DKRZ)
UC Berkeley	UCB	University of California, Berkeley (UCB)
University of California, Berkeley	UCB	University of California, Berkeley (UCB)

Results

Prior to extracting data for the acknowledgement corpus, we performed a preliminary analysis of the entries stored in WoS[16]. Only articles from four scientific domains published between 2014 and 2019 in English were considered. As Fig. 4 demonstrates, not every WoS entry has an acknowledgement text. This corresponds to the findings of Paul-Hus et al. (2016). Thus, oceanography has the smallest amount of articles, but at the same time is the best covered area in terms of availability of acknowledgement texts: 75 % of all articles posses an acknowledgement text. Moreover 73 % of these articles have indexed funding information. The computer science domain includes the largest amount of entries, 42 % of which are in possession of an acknowledgment text. At the same time 84 % of these articles have an indexed funding information, which makes the computer science domain to the best covered domain in terms of funding information indexing. In both economics and social science only 28 % of all records include acknowledgement text and only 65 % of these records have indexed funding information.

Acknowledgment trends

Figure 5 shows the distribution of entities of different types between scientific domains. The distribution of entities demonstrates clear differences among scientific domains. Therefore, IND is the most frequent entity type in economics, while oceanography demonstrates a highest mean number of acknowledged individuals per acknowledgement text among the highest standard deviation (as Fig. 6A and Table 9 demonstrate), which means that the numbers of acknowledged individuals in some acknowledgement texts are significantly more or less as mean values. In general, all scientific domains (Fig. 6) show the highest standard deviation for the numbers of acknowledged individuals. In computer science as opposed to other disciplines the highest mean number of acknowledged entities belongs to grant numbers. Overall, all disciplines showed high dispersion of number of all acknowledged entities relating to the mean values. FUND is the most frequent in social science and oceanography, at the same time GRNB is the most frequent entity in computer science. However, social science and oceanography domains show similar frequency patterns of FUND, IND and GRNB categories: FUND is the most frequent category followed by IND and GRNB. COR is the most underrepresented category in all scientific domains. UNI and MISC have similar distributions in economics, social sciences and computer science: more entities fall into the UNI category than into the MISC category. On the other

---

[16] Analysis was performed on the disambiguated (cleaned) data.

Table 3 An example of disambiguated but not manually reviewed entities from the IND category

Oceanography		Social sciences		Economics		Computer science	
IND	Frequency	IND	Frequency	IND	Frequency	IND	Frequency
Drs.	89	Kathleen Mullan Harris	289	Jan Wallander	163	H. Zhang	27
J.	72	Richard Udry	158	Andrei Shleifer	115	X. Li	26
Kelly Cooper	68	Peter S. Bearman	149	Li	89	J.	25
Ki-Han Kim	45	Barbara Entwisle	120	Raj Chetty	85	P. Shi	23
Maria S. Merian	44	Ronald R. Rindfuss	117	Sushanta Mallick	83	J. Liu	23
M.	44	John D.	72	Andrew Karolyi	83	H. Wang	23
R.	35	Oliver Linton	60	Iftekhar Hasan	80	L. Wang	22
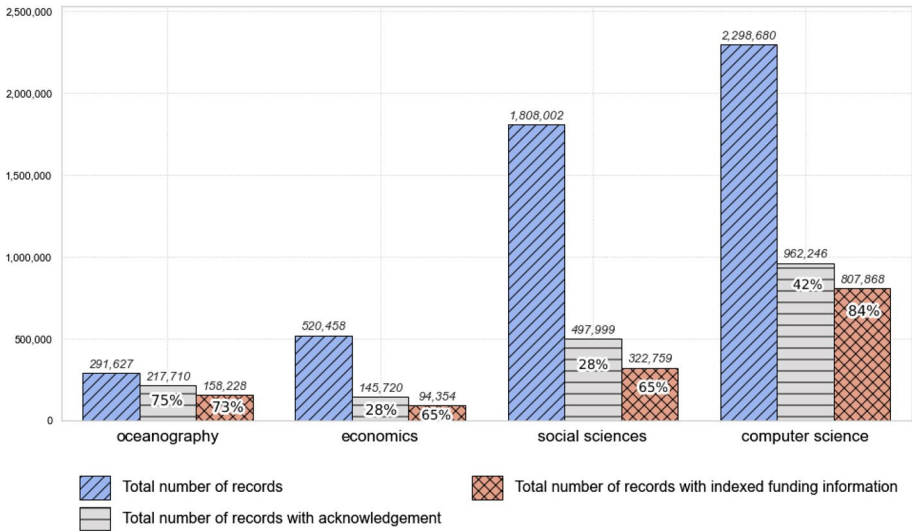
Fig. 4 Analysis of WoS entries from four scientific domains published between 2014 and 2019 (publications in English)
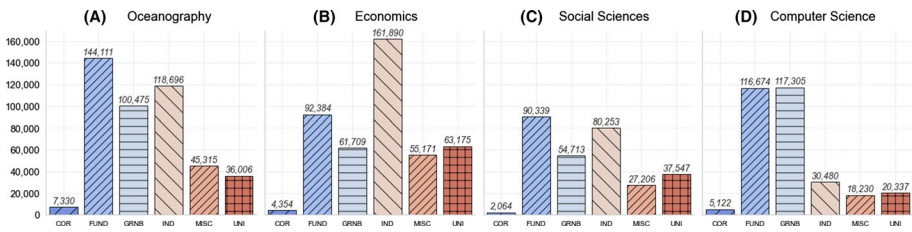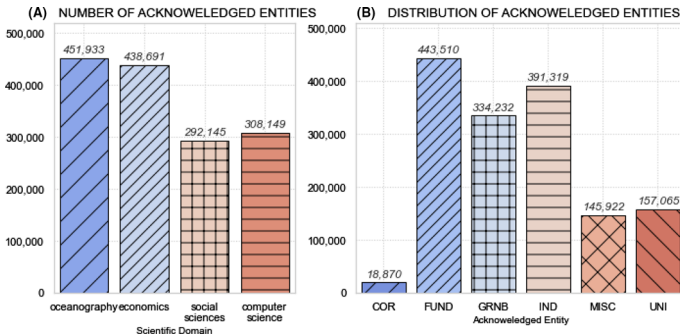

Fig. 5 The distribution of acknowledged entities between disciplines
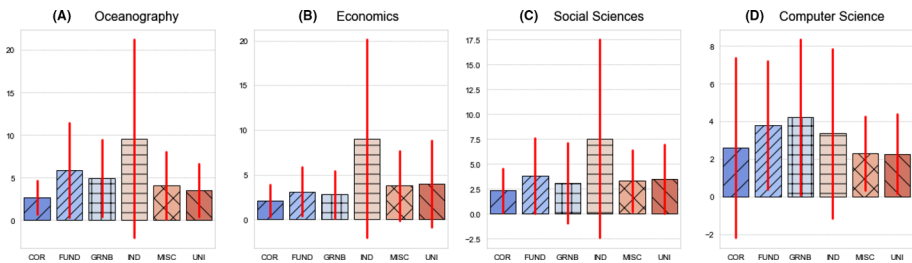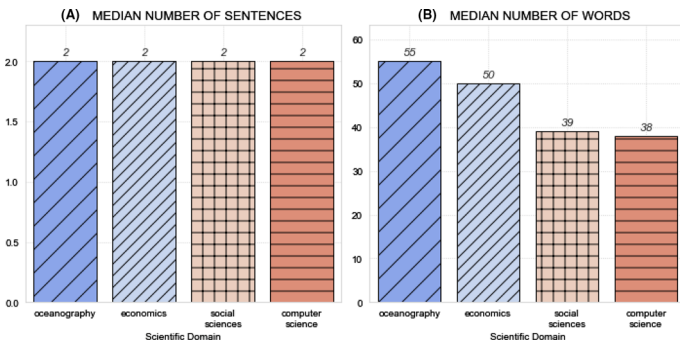

Fig. 6 Mean number of acknowledged entities (with the standard deviation) per paper in different disciplines. Note: the scale of the y-axis is non-uniform due to clearness

hand, oceanography shows the opposite pattern. From all scientific domains, computer science demonstrates the lowest number of acknowledged individuals and the highest number of grant numbers.

Fig. 7 The general distribution of acknowledged entities categories. **A** represents the total amount of acknowledged entities among disciplines and **B** represents the general distribution of acknowledged entities


Fig. 8 Median number of sentences (**A**) and median number of words (**B**) in acknowledgment texts in different scientific domains

Overall, oceanography showed the highest number of acknowledged entities, followed by economics, computer science and social sciences, as Fig. 7A demonstrates. Oceanography and economics, and social sciences and computer science have minor differences between the numbers of acknowledged entities, while there is a bigger gap between the two pairs. As Fig. 7B shows, FUND is the most frequent entity followed by IND, GRNB, UNI, MISC and COR. IND and GRNB are the second most frequent categories. The COR category has the lowest number of occurrences.

Analysis of the lengths of acknowledgement texts showed that in general in all four scientific domains an acknowledgement consists of two sentences, which comprise 38 (computer science), 39 (social sciences), 50 (economics), and 55 (oceanography) words[17]. Thus, as Fig. 8 demonstrates, the oceanography domain has in general longer acknowledgement texts with longer sentences followed by economics, social sciences and computer science.

Figure 9 depicts the development of the lengths of acknowledgement texts in four scientific domains from 2014 till 2019. Changes in the number of sentences were observed for all scientific domains. Thus, in economics (Fig. 9B) the number of sentences increased from 2 in 2014 to 3 in 2015 and then decreased to 2 again in the following years. In social

---

[17] Punctuation marks were excluded from the count.
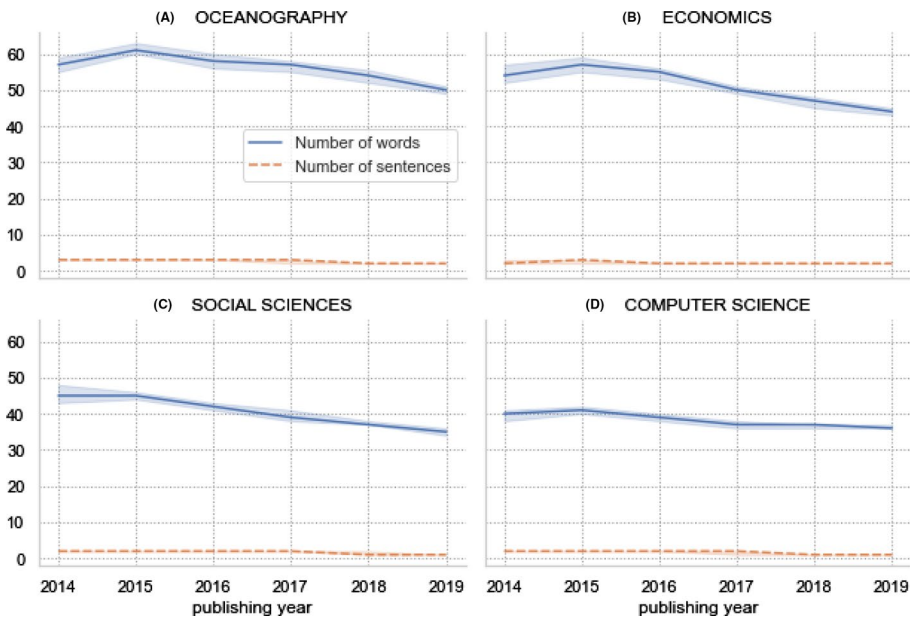
Fig. 9 Development of the lengths of the acknowledgement texts from four scientific domains in 5 years (2014–2019). The figure shows median and 95% confidence interval for each unit

sciences (Fig. 9C), oceanography (Fig. 9A) and computer science (Fig. 9D) the similar pattern can be observed: the number of sentences decreased by 50% in computer science and social sciences and by 33% in oceanography starting from the year 2018. Number of words also decreased in all scientific domains. Thus, the highest percentage decrease of 22% showed social sciences, followed by economics (19%) and oceanography (12%). The lowest value of decrease of 10% is in the computer science domain. The highest range in the median number of words was observed in economics and social sciences in 2014. The smallest amount of variations showed computer science. In general, from 2014 to 2019 the value range decreased in all disciplines.

Figures 10, 11, 12, 15, and 16 demonstrate the top 30 acknowledged entities of different types except GRNB. Some interesting trends can be observed as described further below.

Figure 10 shows the top 30 most frequent entities that fall under the FUND category. All scientific domains, except social sciences, have similar top first funding organizations: the National Natural Science Foundation of China (NSFC), whereas the top first funding organization in social sciences is the National Institutes of Health (NIH). The United States-Israel Binational Science Foundation (BSF) is the second most frequent funding organization in all disciplines, except computer science, which second most frequent funding organization is Fundamental Research Funds for the Central Universities and the United States-Israel Binational Science Foundation (BSF) took the third place. Mainly, the major funders such as the National Natural Science Foundation of China, National Institutes of Health, Deutsche Forschungsgemeinschaft (DFG), etc. can be found in the top 30 funding organizations list.

Figure 11 shows the top 30 most frequent entities which fall into the UNI category. The Academy of Finland is the first top university in social sciences and computer science. The Norwegian University of Science and Technology (NTNU) is in the second position for oceanography and in the third position in computer science. The University of California,
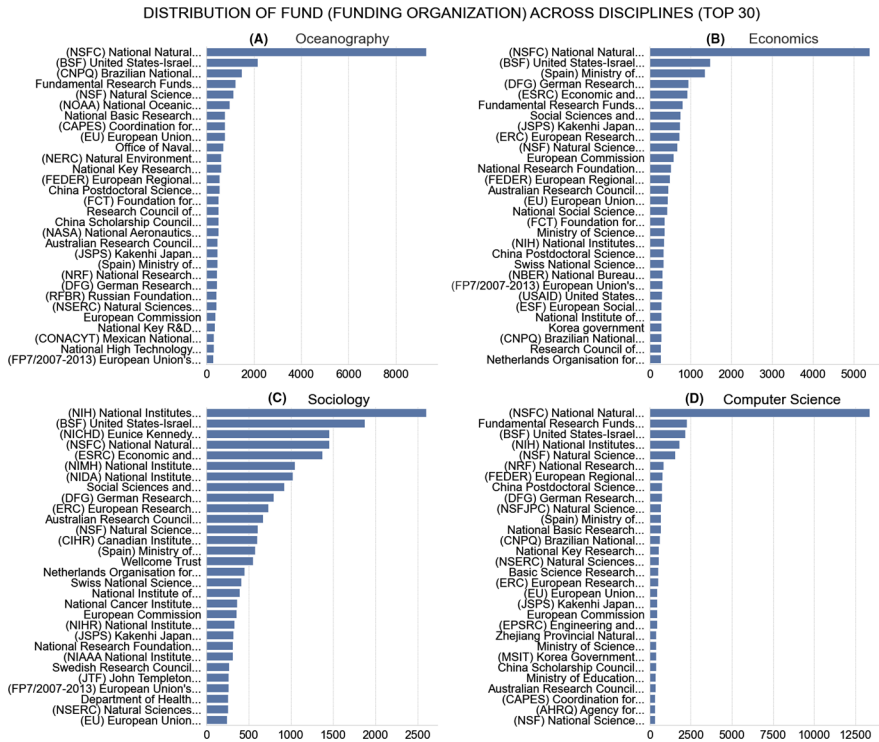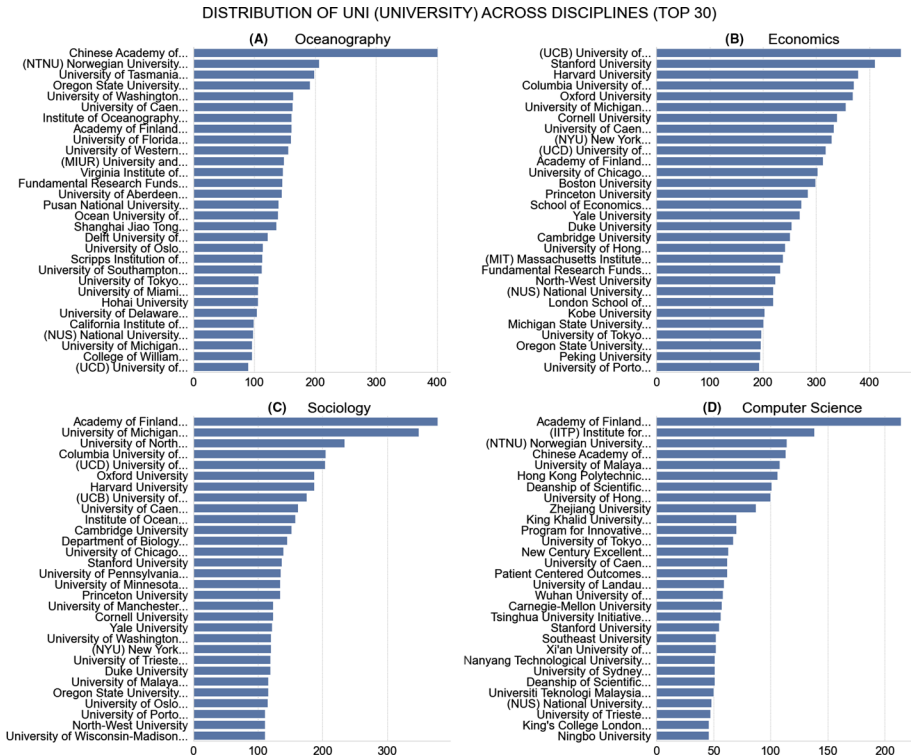
Fig. 10 Top 30 acknowledged entities which fall into the FUND (funding organization) category. **A** represents entities from oceanography, **B** from economics, **C** from social science, and **D** from computer science

Berkeley (UCB) is in the top position in economics and the Chinese Academy of Sciences in oceanography.

Figure 12 depicts the top 30 most frequent entities which fall into the COR category and entities from this category tend to be scientific domain specific. Thus, the top two most frequent corporations in computer science (NVIDIA Corporation and Google) are companies from the software and computing field. The top first and third entities in oceanography (Petrobras and Shell) are companies from the petroleum industries. Pfizer (a pharmaceutical and biotechnology corporation) made the top third in economics, computer science and social sciences. The top second in social sciences and the top third in economics is Novartis, a pharmaceutical company. The World Bank is the most frequent entity in economics and social sciences. The top third for social sciences is Merck & Co., Inc, which is also a pharmaceutical corporation.

Overall, the distribution of acknowledged entities in oceanography, economics and computer science has the following pattern: the top first most frequent entity has the greatest number of occurrences, other entities starting from the top second occur massively less frequently. On the other hand, the distribution of entities in social sciences is more uniform. That distribution pattern can be observed (more or less defined) almost for all entity types in all scientific domains and correspond with the Giles and Councill's (2004) findings and follows a power law, which postulates that "*few entities are named very frequently and a great many entities are named only rarely*" (Giles & Councill, 2004, p.17601).

Fig. 11 Top 30 acknowledged entities which fall into the UNI (universities) category. **A** represents entities from oceanography, **B** from economics, **C** from social science, and **D** from computer science

Nevertheless, it was observed that some entities do not follow the power law mentioned above. This applies to the entities of UNI category from economics domain (Fig. 11B) and IND category from computer science domain (Fig. 15D).

Relationships between analysed variables

A Chi-square test of independence was conducted to examine the relationship between an entity, entity type and scientific domain. As Table 4 demonstrates, the relation between these variables was significant. The P-value of all pairs of examined variables tend toward zero.

However, a weak association in a large sample size may also result in a very low *P* value[18]. To assess the strengths of association between an acknowledged entity, entity type and a scientific domain the Cramer's V test was conducted. The results of the analysis are represented in Fig. 13. Analysis was performed on the disambiguated data (see Section 3). Table 5 demonstrates the degrees of freedom for the analyzed variable pairs.

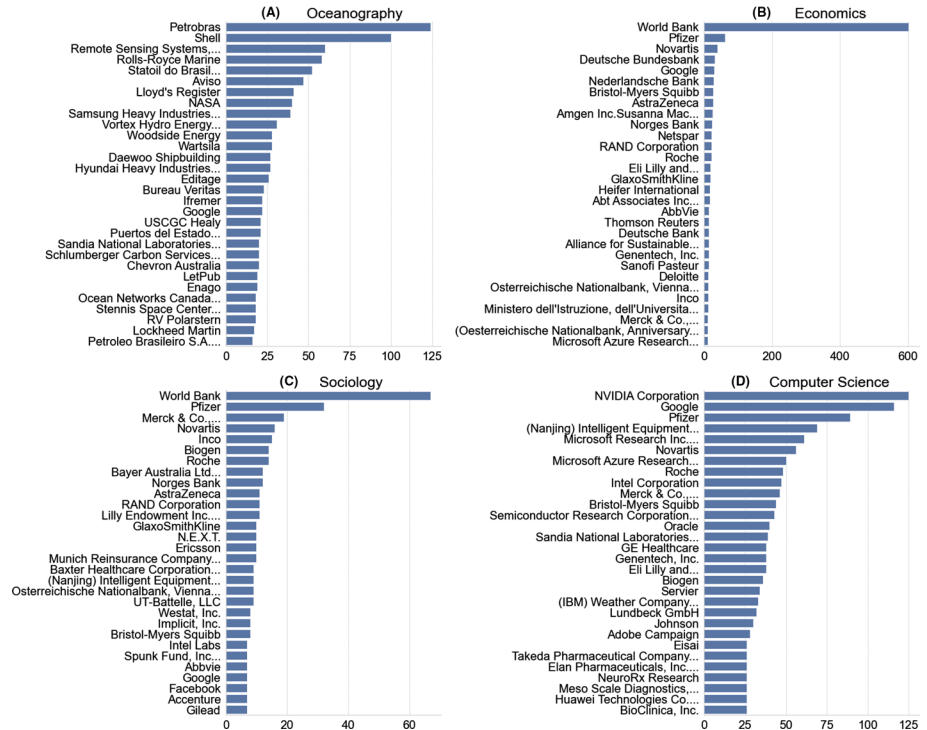[18] https://www.spss-tutorials.com/cramers-v-what-and-why/.

Fig. 12 Top 30 acknowledged entities which fall into the COR (corporation) category. **A** represents entities from oceanography, **B** from economics, **C** from social science, and **D** from computer science

Table 4 The results of the Chi-square test for multiple variables

| Variables	Chi-square	*P* value	Degrees of freedom
Entity type—scientific field	143606.086	0.0	15
Entity—scientific field	3264463.167	0.0	1343037
Entity—entity type	7107380.717	0.0	2238395

The column "Variables" represents variable pairs that were examined

A large association indicates that by knowing the value of one category, the value of a second related category can be predicted. No association, on the other hand, indicates that the categories are not related. As expected, there is a large association between the acknowledged entity and the entity type (the Cramer's *V* is equal to 0.98). The variable pair entity type and scientific domain showed a small association of 0.18. Furthermore, a large association between the entity and scientific domain was discovered (0.85). Thus, a precise acknowledged entity would, with high probability, have a precise label and would belong to the precise scientific domain.
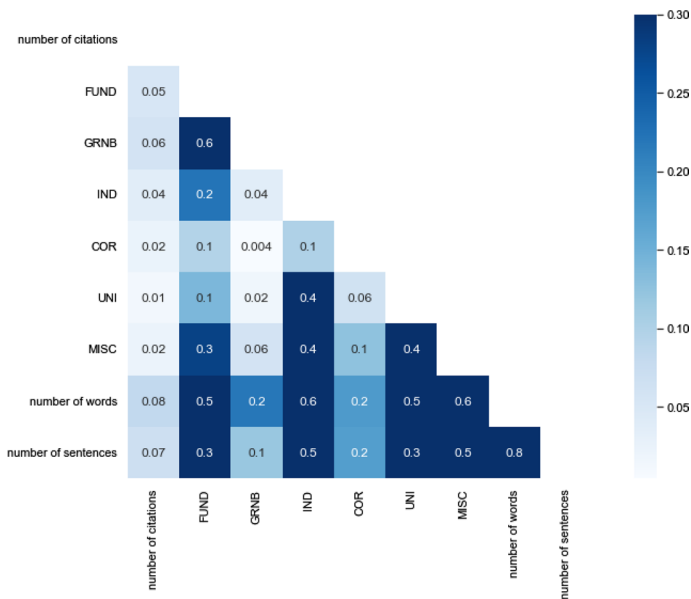
Relationships between different numerical variables (number of citations, number of acknowledged entities, number of words and number of sentences in the acknowledgement

Fig. 13 Distribution of the Cramer's *V* values between scientific domains, entity types and acknowledged entities
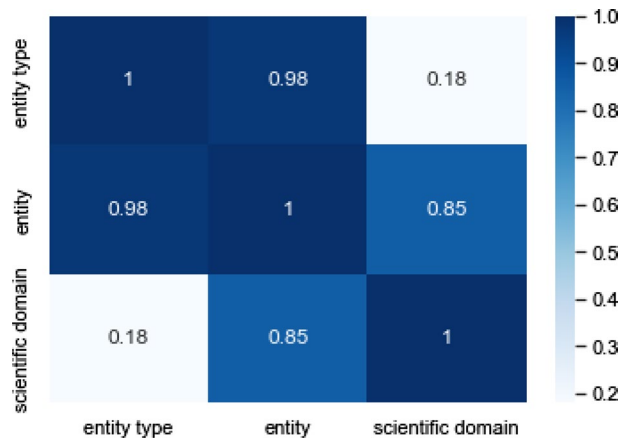
Table 5 Degrees of freedom for the variable pairs

Variable	Entity type	Scientific domain
Entity type		3
Entity	5	3

Fig. 14 Correlation between number of citations of the article, number of acknowledged entities of different types, number of sentences and number of words in the acknowledgement text

texts) were examined with the use of the Pearson correlation coefficient (Pearson's R)[19]. As Fig. 14 demonstrates, no correlation between number of citations and number of acknowledged entities was found: values of the Pearson's R for all variable pairs is under 0.1. Furthermore no correlation was found between the number of citations and number of words and sentences in the acknowledgement text.

High correlation was observed between number of words and number of acknowledged funding organizations (0.5), individuals (0.6), universities (0.5) and miscellaneous entities (0.6). The number of sentences showed a high correlation with the number of acknowledged individuals (0.5) and miscellaneous entities (0.5) and a moderate correlation with the number of acknowledged funding organizations (0.3) and universities (0.3). Clearly, the number of sentences in the acknowledgement text is highly correlated with the number of words (0.8).

Additionally, a high correlation (0.6) was discovered between the number of grant numbers and number of acknowledges funding organizations. A moderate correlation of 0.4 was observed between the number of acknowledged universities and the number of individuals and miscellaneous entities along with between the number of acknowledged individuals and miscellaneous entities.

The ANOVA test was performed on the analysed variables. The *P* value is very low (0.000), which means that the results are statistically significant.

Discussion

As we expected, the relations between entity types and scientific domains were found. A strong association was found between acknowledged entity and scientific domain, and acknowledged entity and entity type, as we anticipated. Nevertheless, an association between scientific domain and entity type was minor. Furthermore, social sciences was expected to have the highest number of acknowledged individuals [as in Paul-Hus et al. (2017)] and economics [as in Diaz-Faes and Bordons (2017)], the results partly coincided with our expectation: our analysis showed that individuals were most acknowledged in economics. Social sciences and computer science in general showed a smaller number of acknowledged entities than economics and oceanography. This can possibly be explained by the fact that the acknowledgement corpus generally contains fewer acknowledgements from these two scientific domains than from economics and oceanography (Table 1).

The prevalence of the FUND category is not surprising as WoS stores only acknowledgements containing funding information. Unexpectedly, the IND category was the second most frequent category: GRNB was expected to be one of the most frequent categories. Interesting trends were observed in the COR category. The most research in social sciences and economic domains is funded by pharmaceutical corporations and banks, while the top corporations acknowledged in the oceanography domain come from the petroleum industry. The top corporations in computer science are in the software and computing field.

Generally, the number of words in the acknowledgement texts positively correlates with the number of acknowledged funding organizations, universities, individuals and miscellaneous entities. At the same time, acknowledgement texts with the larger number of sentences have more acknowledged individuals and miscellaneous categories. Correlations

---

[19] Analyses was performed on the normalized data.

between frequencies of entities of different types were found. Thus, acknowledgement texts with more acknowledged funding organizations would posses more grant numbers. Acknowledgements with more acknowledged individuals will have more acknowledged universities and miscellaneous entities and vice versa.

Our analysis revealed that oceanography in general have the longest acknowledgement texts, while computer sciences domain the shortest ones. Generally number of words in an acknowledgement text decreased from 2014 to 2015 in all scientific domains.

Giles and Councill's (2004) analysed the correlations between the number of acknowledgements received by an individual and that individual's h-index. H-index is an index used to assess an individual's scientific research output (Hirsch, 2005). Giles & Councill (2004) used computer science research papers from the CiteSeer digital library as a data source. We tried to replicate the Giles and Councill's (2004) study, but faced difficulties in referencing an acknowledged individual name and a real person[20]. To solve this problem additional disambiguation techniques should be developed. Many acknowledgement texts contain name of a person together with the person's institutional affiliation (usually names of universities or corporations), therefore it is possible to identify the acknowledged person through the name of person's affiliation. The Flair framework additionally provides the position of the acknowledged entity in the text (i.e. index range of the characters). Thus, an acknowledged university or corporation, which is in the close index range with the acknowledged individual would highly likely be the the person's institutional affiliation.

Additionally, we would propose to include an ORCID of the acknowledged individual while writing an acknowledgement text, to make the identification procedure easier.

Conclusion

The analysis of the automatically extracted entities revealed differences and distinct patterns in the distribution of acknowledged entities of different types between different scientific domains. The extracted funding information poses various possibilities and challenges for further research. Thus, grant numbers, funding organizations and corporations might give an insight on the impact of funding sponsorship on scientific research. Names of individuals might assess one's scholarly performance, and the miscellaneous category might reveal other non-sponsorship variables that might influence academic research.

One of the future research aims might be the analysis of correlations between the acknowledged entities and such characteristics as journal impact, other scientific domains, researcher's productivity and impact. Furthermore, some funding organizations might find the extracted acknowledgements information useful in order to track if funding recipients acknowledge funding in their publications. The disambiguation of acknowledged entities poses further interesting challenges, since in the present paper only a superficial disambiguation of the data was conducted which was necessary for the analysis and visualization of the results.

The main limitation of our study was the peculiarities of the WoS funding information indexing. The WoS includes only acknowledgements containing funding information; therefore, not every database entry has an acknowledgement. That can explain the prevalence of the funding agencies over other types of entities.

---

[20] For example, it is yet impossible to detect individuals, which were acknowledged with abbreviated names. Some identical names can belong to different people.

# Appendix 1: List of WoS disciplines used for the present paper

Classification	Scientific field
Engineering, Ocean	Oceanography
Marine & Freshwater Biology	Oceanography
Oceanography	Oceanography
Engineering, Marine	Oceanography
Economics	Economics
Agricultural Economics & Policy	Economics
Business & Economics	Economics
Biomedical Social Sciences	Social sciences
Social Work	Social sciences
Social Sciences, Biomedical	Social sciences
Social Sciences, Interdisciplinary	Social sciences
Social Sciences - Other Topics	Social sciences
Social Issues	Social sciences
Psychology, Social	Social sciences
Social Sciences, Mathematical Methods	Social sciences
Sociology	Social sciences
History Of Social Sciences	Social sciences
Social Sciences	Social sciences
Mathematical Methods In Social Sciences	Social sciences
Computer Science, Software Engineering	Computer science
Computer Science, Hardware & Architecture	Computer science
Computer Science, Artificial Intelligence	Computer science
Computer Science	Computer science
Medical Informatics	Computer science
Computer Science, Information Systems	Computer science
Computer Science, Interdisciplinary Applications	Computer science
Computer Science, Cybernetics	Computer science
Computer Science, Theory & Methods	Computer science

Appendix 2: Top 30 acknowledged individuals and miscellaneous entities

See Figs. .
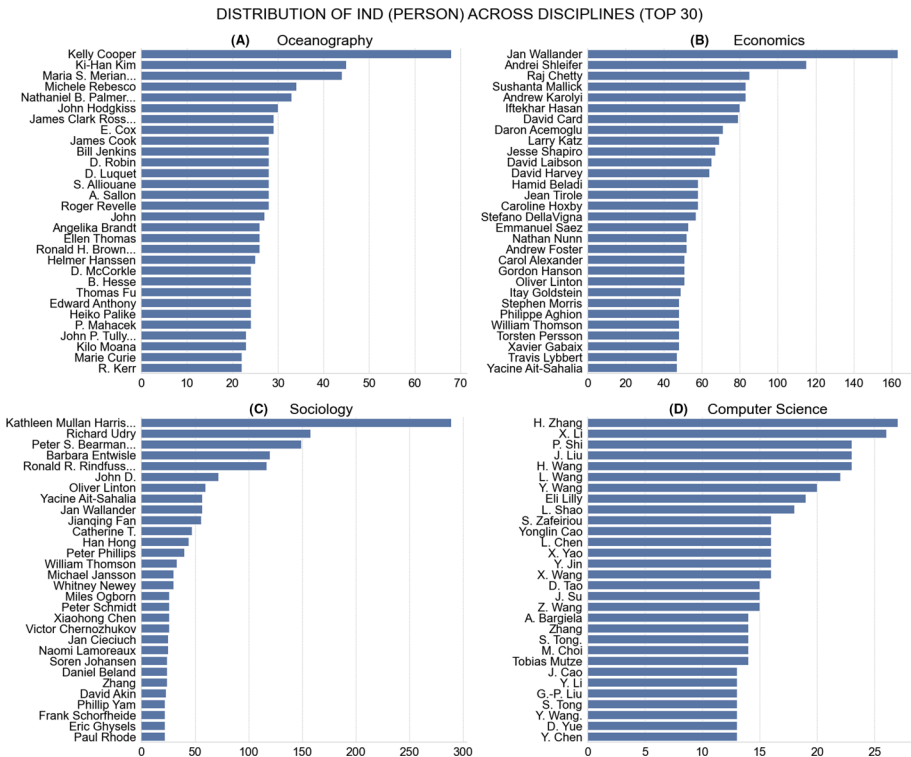
Fig. 15 Top 30 acknowledged entities which fall into the IND (person) category. **A** represents entities from oceanography, **B** from economics, **C** from social science, and **D** from computer science
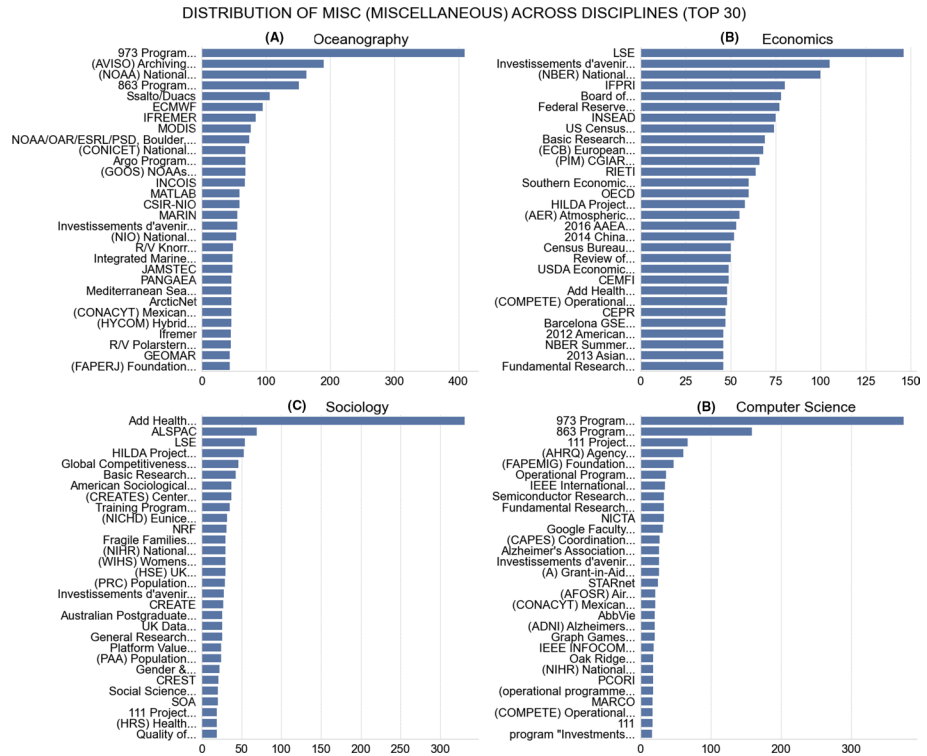
Fig. 16 Top 30 acknowledged entities which fall into the MISC (miscellaneous) category. **A** represents entities from oceanography, **B** from economics, **C** from social science, and **D** from computer science

Appendix 3: Mean number of acknowledged entities per paper in different disciplines including standard deviation

Scientific field	Label	Mean	Standard deviation
Computer science	COR	2.601320	4.790851
Computer science	FUND	3.810634	3.411334
Computer science	GRNB	4.213542	4.129411
Computer science	IND	3.360900	4.488084
Computer science	MISC	2.302349	1.951288
Computer science	UNI	2.275596	2.118370
Economics	COR	2.103382	1.808882
Economics	FUND	3.125622	2.751300
Economics	GRNB	2.778183	2.649553
Economics	IND	9.017434	11.094021
Economics	MISC	3.773150	3.924555
Economics	UNI	4.002978	4.856648
Oceanography	COR	2.633848	1.977578
Oceanography	FUND	5.871298	5.543036

Scientific field	Label	Mean	Standard deviation
Oceanography	GRNB	4.906725	4.525458
Oceanography	IND	9.569943	11.648552
Oceanography	MISC	4.067774	3.938683
Oceanography	UNI	3.547739	3.105968
Social sciences	COR	2.332203	2.235140
Social sciences	FUND	3.807435	3.771419
Social sciences	GRNB	3.118261	4.026767
Social sciences	IND	7.541158	9.967386
Social sciences	MISC	3.324291	3.095832
Social sciences	UNI	3.506444	3.447353

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). *FLAIR: an easy-to-use framework for state-of-the-art NLP* (pp. 54–59). Association for Computational Linguistics.

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. 2018, 27th International Conference on Computational Linguistics, pp. 1638–1649.

Alexandera, D., & Vries, A. P. (2021). This research is funded by...: Named entity recognition of financial information in research papers. BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval at ECIR, pp. 102–110.

Alvarez-Bornstein, B., & Montesi, M. (2021). Funding acknowledgements in scientific publications: A literature review. *Research Evaluation, 29*(4), 469–488. https://doi.org/10.1093/reseval/rvaa038

Baccini, A., & Petrovich, E. (2021). Normative versus strategic accounts of acknowledgment data: The case of the top-five journals of economics. *Scientometrics*. https://doi.org/10.1007/s11192-021-04185-6.

Cronin, B. (1995). *The Scholar's courtesy: The role of acknowledgement in the primary communication process.* London: Taylor Graham.

Cronin, B., & Weaver, S. (1995). The praxis of acknowledgement: From bibliometrics to influmetrics. *Revista Española de Documentación Científica, 18*(2), 172.

Diaz-Faes, A. A., & Bordons, M. (2017). Making visible the invisible through the analysis of acknowledgements in the humanities. *Aslib Journal of Information Management, 69*(5), 576–590. https://doi.org/10.1108/AJIM-01-2017-0008.

Giles, C. L., & Councill, I. G. (2004). Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Proceedings of the National Academy of Sciences USA, 101*(51), 17599–17604. https://doi.org/10.1073/pnas.0407743101.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences USA, 102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102.

Kassirer, J. P., & Angell, M. (1991). On authorship and acknowledgments. *The New England Journal of Medicine, 325*(21), 1510–1512. https://doi.org/10.1056/NEJM199111213252112.

Kayal, S., Afzal, Z., Tsatsaronis, G., Katrenko, S., Coupet, P., Doornenbal, M. & Gregory, M. (2017). Tagging funding agencies and grants in scientific articles using sequential learning models. In: BioNLP 2017, Vancouver, Canada, pp. 216–221. Association for Computational Linguistics.

Lewison, G. (1994). Publications from the European community's biotechnology action programme (BAP): Multinationality, acknowledgement of support, and citations. *Scientometrics, 31*(2), 125–142. https://doi.org/10.1007/BF02018556.

Mackintosh, K. (1972). Acknowledgements patterns in sociology. Ph. D. thesis, University of Oregon.

Mccain, K. (2017). 12. Beyond garfield's citation index: An assessment of some issues in building a personal name acknowledgments index. *Scientometrics*. https://doi.org/10.1007/s11192-017-2598-1.

McCain, K. W. (1991). Communication, competition, and secrecy: The production and dissemination of research-related information in genetics. *Science, Technology, & Human Values, 16*(4), 491–516. https://doi.org/10.1177/016224399101600404.

McCain, K. W. (2018). Beyond garfield's citation index: An assessment of some issues in building a personal name acknowledgments index. *Scientometrics*. https://doi.org/10.1007/s11192-017-2598-1.

Mejia, C., & Kajikawa, Y. (2018). Using acknowledgement data to characterize funding organizations by the types of research sponsored: The case of robotics research. *Scientometrics, 114*(3), 883–904. https://doi.org/10.1007/s11192-017-2617-2.

Paul-Hus, A., & Desrochers, N. (2019). Acknowledgements are not just thank you notes: A qualitative analysis of acknowledgements content in scientific articles and reviews published in 2015. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0226727.

Paul-Hus, A., Desrochers, N., & Costas, R. (2016). Characterization, description, and considerations for the use of funding acknowledgement data in web of science. *Scientometrics, 108,* 167–182. https://doi.org/10.1007/s11192-016-1953-y.

Paul-Hus, A., Díaz-Faes, A., Sainte-Marie, M., Desrochers, N., Costas, R., & Larivière, V. (2017). Beyond funding: Acknowledgement patterns in biomedical, natural and social sciences. *PLOS ONE, 12,* e0185578. https://doi.org/10.1371/journal.pone.0185578.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Petrovich, E. (2022). Acknowledgments-based networks for mapping the social structure of research fields. A case study on recent analytic philosophy. *Synthese*. https://doi.org/10.1007/s11229-022-03515-2.

Rigby, J., & Julian, K. (2014). On the horns of a dilemma: Does more funding for research lead to more research or a waste of resources that calls for optimization of researcher portfolios? An analysis using funding acknowledgement data. *Scientometrics, 101*(2), 1067–1075. https://doi.org/10.1007/s11192-014-1259-x.

Rose, M., & Georg, C. P. (2021). What 5,000 acknowledgements tell us about informal collaboration in financial economics. *Research Policy, 50,* 104236. https://doi.org/10.1016/j.respol.2021.104236.

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics, 126*(6), 5113–5142. https://doi.org/10.1007/s11192-021-03948-5.

Smirnova, N., & Mayr, P. (2022). Evaluation of embedding models for automatic extraction and classification of acknowledged entities in scientific documents. In C. Zhang, P. Mayr, W. Lu, and Y. Zhang (Eds.), Proceedings of the 3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2022), pp. 48–55. CEUR, Aachen.

Song, M., Kang, K. Y., Timakum, T., & Zhang, X. (2020). Examining influential factors for acknowledgements classification using supervised learning. *PLOS ONE, 15*(2), e0228928.

Thomer, A. K., & Weber, N. M., (2014). Using named entity recognition as a classification heuristic. In: iConference 2014 Proceedings, pp. 1133–1138. iSchools.

Tian, S., Xu, X., & Li, P. (2021). Acknowledgement network and citation count: The moderating role of collaboration network. *Scientometrics, 126*(9), 7837–7857. https://doi.org/10.1007/s11192-021-04090-y.

Tollison, R., & Laband, D. (2003). Good colleagues. *Journal of Economic Behavior & Organization, 52,* 505–512. https://doi.org/10.1016/S0167-2681(03)00070-2.

Wang, J., & Shapira, P. (2011). Funding acknowledgement analysis: An enhanced tool to investigate research sponsorship impacts: The case of nanotechnology. *Scientometrics, 87*(3), 563–586. https://doi.org/10.1007/s11192-011-0362-5.