Scientometrics (2022) 127:5613–5631
https://doi.org/10.1007/s11192-022-04475-7

Combining Web of Science and Scopus datasets in citation-based literature study

Miika Kumpulainen[1] · Marko Seppänen[1]

Received: 2 April 2021 / Accepted: 29 July 2022 / Published online: 25 August 2022

Abstract

Scientific research builds on previous studies and scientifically proven knowledge. Researchers must master the recent developments in the field when designing research to answer new questions. Today, the accessibility of research literature is abundant due to digitized publications, extensive coverage of citation indexes, and several literature databases. The means for conducting systematic literature reviews have greatly improved. In recent years, a lot of research with methods, such as systematic literature reviews, literature mapping, and visualizing studies, has been published. Despite the tools having been perfected, they still have limitations. Particularly, database selection for a literature search sets a bias, because the citation indexes differ in coverage according to the scientific domain. In this paper, bibliometric data from two citation indexes (Web of Science and Scopus) are combined for the purposes of bibliometric analyses in the context of inter-firm relationships. The combination process requires a lot of unifying and repairing of data in practice that is referred as data wrangling. This study describes the process steps and the lessons learned and considers the amount of effort required.

Keywords Citation analysis · Literature review · Data combining · Web of Science · Scopus

Introduction

The essential part of any scientific study is a review of the existing literature. By this means, the researcher can grasp the existing scientific knowledge of the phenomenon in question. Literature searches have become rather easy today, thanks to technology improvements. Many databases and citation indexes contain a vast amount of literature; however, the coverage of citation indexes varies depending on the scientific domain. When conducting a bibliometric study and applying visualization techniques to demonstrate the results, often the data

✉ Miika Kumpulainen
miika.kumpulainen@tuni.fi

Marko Seppänen
marko.seppanen@tuni.fi

[1] Information and Knowledge Management, Faculty of Management and Business, Tampere University, Tampere, Finland

stems from one citation index. This leads to the fact that the illustrations of the bibliometric results obtained are dependent on the citation index selected (Mongeon & Paul-Hus, 2016).

This paper examines the differences between Clarivate Analytics' Web of Science (WoS) and Elsevier's Scopus (Scopus) citation data and assesses the working procedure needed to combine this data into one uniform dataset. The study is conducted in the context of inter-firm relationships, which is a rich research area in terms of published scientific literature. The differences in document co-citation analysis are sought between the citation indexes, and the challenging process of combining the two sets of citation index data into a uniform dataset is described in detail. In addition, the amount of manual work in relation to the obtained results is assessed.

Bibliometrics is a quantitative method which uses scientific literature as data for its analyses (Forsman, 2016). As for any analyses, the quality of the data is essential for flawless results and conclusions. The data unifying process description clarifies the citation data details and all the executed corrective actions to purify, unify, and repair the data. The corrective operations performed resemble "data wrangling", a term that is defined as data cleaning and unifying in the context of Big Data. The document co-citation analysis results are presented, and insights offered for other scholars to consider carefully whether it is worthwhile engaging in time- and effort-consuming data wrangling.

This study combines two data sources to improve the quality of bibliometric study. Two databases, Clarivate Analytics' Web of Science (WoS) and Elsevier's Scopus (Scopus), were used. These two typically represent the major databases and citation indexes for general-purpose scientific literature including journal articles, conference proceedings, and books. WoS and Scopus also cover a vast number of scientific domains albeit with only partially overlapping domains: WoS covers natural sciences and engineering widely, whereas the Scopus coverage of the social sciences is relatively higher than that of WoS (Mongeon & Paul-Hus, 2016). Thus, it is assumed that combining these data sources may result in richer results in terms of literature fields.

This bibliometric study was conducted on inter-firm or inter-organizational relationships by focusing on document co-citation analysis. To assess whether it is worth unifying this data, the process of so-called data wrangling is presented in detail. More specifically, this process involves unifying the citation data of the scientific articles that have been included in the study. Unifying includes automized processes but also a rather great deal of manual work. If the results of combined datasets appear to yield improved bibliometric analyses, one can approximate the pay-off for the time and effort invested.

Method

Bibliometric research

The study is based on bibliometric analyses. For these purposes, BibExcel and VOSviewer were used. BibExcel is a program dedicated to performing various types of bibliometric analyses and it is compatible with many different software suites (Persson et al., 2009) (https://homepage.univie.ac.at/juan.gorraiz/bibexcel/). The VOSviewer is a versatile program for creating, analyzing, and visualizing network data, especially bibliometric data (van Eck & Waltman, 2019) that can be provided for instance, by BibExcel. Version 1.6.13 of the program was used.

A systematic approach was employed to obtain the literature data. This process consisted of several steps that are similar to those used in a systematic literature review (Denyer & Tranfield, 2009; Tranfield et al., 2003). The systematic literature review method emphasizes the transparency of the data selection and attempts to minimize the biased views of the researchers conducting the review. The literature data serves as the source of the bibliographies, which are the material for the bibliometric analyses.

The aim was to describe the unifying process of converting Scopus citation data into a form compatible with WoS citation data. The target was to create a unified dataset in terms of the information contained in the citation and how this information was expressed in written form. Achieving this required the use of computer assisted data processing, which is called "data wrangling" in this paper. Common software programs (Microsoft Office's Excel and Notepad++) were used.

Document co-citation analysis

Document co-citation is based on the notion that two documents are cited together by another document, and that the strength of the co-citation of the two documents increases, the greater the number of documents that cite these documents together (Small, 1978). Co-citation analysis is used to identify groups of documents that share a similarity of subject, hence they could be interpreted as forming a semantic relation to each other (Leung et al., 2017; Small, 1980). Bibliometrics also includes other quantitative analyses that are based on the identification of co-occurrences in the analyzed data: e.g., co-word and co-author analysis (Cuc, 2019; He, 1999; White & Griffith, 1981). Bibliographic coupling analysis has been used to identify documents that share a research focus (Chang et al., 2015; Jarneving, 2007). Further, bibliographic coupling exists between two documents if they share a reference (Kessler, 1963).

Document co-citation analysis emphasizes the significance of the document as a unit of analysis. Small (1978) introduced the idea that cited documents could be used as symbols of concepts. Document co-citation analysis forms a network of document clusters. Interpreting the contents of the document clusters and the relationships between the clusters forms an interesting area for bibliometric research (Chen et al., 2010).

Literature search

Context in inter-firm relationships

The inter-firm relationship concept and the related literature form the context of this study. On the other hand, the phenomenon of two firms or organizations having a relationship with one another is easy to identify but at the same time challenging to define. Scholars have studied inter-firm relationships from many angles, for example why firms engage in a relationship (Dyer & Singh, 1998), and what kinds of relationships exist and how they develop over time (Cannon & Perreault, 1999; Dwyer et al., 1987; Wilson, 1995).

The concept of inter-firm relationships is often studied in the contexts of transaction cost economics, resource dependency, network theory, and social exchange theory (Golicic & Mentzer, 2005). These widely accepted theories have their own focus areas or viewpoints, which has guided the research on inter-firm relationships.

The research on inter-firm relationships is often conducted in the context of transactional exchange, i.e., the relationship between a seller and buyer company. Inter-firm relationships can be viewed in terms of the closeness or distance between the two parties. The extremes are a long-distance relationship (arms-length) and a close collaborative relationship (Anderson & Narus, 1991). Also, more detailed taxonomies of the different kinds of relationships have been introduced in the scientific literature (Cannon & Perreault, 1999; Wong et al., 2010). The relationships have been studied as a development process (Anderson, 1995; Dwyer et al., 1987; Ford, 1980; Wilson, 1995). These studies examine the business relationships as a sequence of stages starting from the search and selection of the other party and ending in a state of maintaining or dissolving of the relationship. During the process, the firms build up commitment to and trust in each other. Commitment and trust as well as satisfaction are the main dimensions of the concept of relationship quality (Athanasopoulou, 2009). Morgan and Hunt's (1994) theory of trust and commitment as important variables in relationship success has had a big impact on the inter-firm relationship research area in terms of the number of citations their article has received (27,443 citations, Google Scholar on January 23, 2020).

Search string generation

A detailed description of a systematic literature review method has been presented earlier (2003). At the beginning of the literature searching process, the aim is to identify the relevant terms and keywords in order to create a search string that fits the study scope most appropriately. In order to identify the relevant terms, first a set of seminal papers on inter-firm relationships was chosen and studied in detail with the aim of observing the terms used. The papers selected were (Anderson, 1995; Cannon & Perreault, 1999; Dwyer et al., 1987; Ford, 1980; Frazier, 1983; Jap & Anderson, 2007; Ring & van de Ven, 1994; Wilson, 1995; Wong et al., 2010). To make this process more transparent and systematic, computer assistance was employed for term identification by transferring the article text in.txt format to the Notepad++ program and running a search function around the term "relations" (including the wildcard functionality in the search e.g., "relationship"). The target was to find the terminology used around the "relations" word. The search concerned and included the word preceding and following "relations". The search also took into account possible capital letters. The different variations were listed; the results included the following terms, which were selected for the search string:

buyer-seller relations *, buyer-supplier relations *, customer-supplier relations *,
dyadic relations *, exchange relations *, inter-company relations *, intercompany relations *,
inter-firm relations *, interfirm relations *, inter-organizational relations *,
interorganizational relations *, inter-organisational relations *, interorganisational relations *

(*Denotes a wildcard, making e.g., "relationship" an inclusive term).

Literature data search

The literature data search was done on May 25, 2018 in the Web of Science Core Collection and in Scopus. An inclusive filter was set for the articles and the English language. There were no criteria for the time period of the publication. The subject areas in the WoS searches were Management, Business, Operations Research, Management Science, Engineering Industrial, Economics, Applied Psychology. For Scopus, the following subject areas were chosen: Business, Management and Accounting, Social Sciences, Economics, Econometrics and Finance, Psychology, Decision Sciences, and Engineering.

The search results were 4100 articles for WoS and 4997 articles for Scopus making a sum total of 9097 articles. All the articles were listed in Excel and coded based on whether the article was found in WoS (W + running number) or Scopus (S + running number). From this dataset, duplicates (2197 articles) were identified by sorting the articles by the author names and then manually comparing whether two articles with similar details existed and excluded from the initial results (n = 6900 articles). It is worthwhile mentioning that in the case of duplicates, the WoS article was chosen for the literature data, because the final aim was to transform the Scopus citations into WoS citation format.

Data selection was done in sequential steps or rounds. The first round of inclusion/exclusion was conducted based on the journal name, i.e., the field it represents. The journal names that clearly did not refer to the study's scope were marked. These consisted of journals within medical science (e.g., psychiatry), strong inter-personal focus in the context of family or social relationships (i.e., lacking the organizational perspective), a purely technical focus, chemistry, anthropology, or a cultural history focus. Also, journals with a geography focus were marked. After this, the article title in question was checked and evaluated whether this specific article in the marked journals corresponded to our study focus. This process excluded 522 articles leaving the final set of articles (n = 6378).

The second round of inclusion/exclusion was the most time- and effort-consuming in the process of literature data gathering. The evaluation of the articles' inclusion/exclusion was based on the article title. The inclusion criteria applied were:

Does the title directly refer to an inter-firm or organizational setting (does the title contain these or similar words e.g., buyer–seller relationship, relationship between companies, relations)?

In borderline cases, the article subject and content were examined in more detail. In practice, these cases were marked separately (in the Excel list with a question mark) and then all checked at the same time afterwards. The checking consisted of reading the article abstract and possibly examining the article's full text. The borderline articles included for example a focus on the intra-organizational relationships between business units, joint ventures, investor relations, franchising settings, and inter-individual relationships. The inclusion/exclusion decision-making was not always easy, but as a guideline, an included article had to contain a discussion of inter-firm or inter-organizational relationships. In the inclusion/exclusion round there were 2957 article inclusions and 2996 exclusions, and the number of borderline cases was 425. Out of these borderline cases, 194 articles ended up being included. Altogether, the number of included articles was 3151. WoS accounted for 2154 articles (68%) and Scopus 997 articles (32%).

Citation data acquisition and methods

Citation data acquisition consisted of the bibliographical data retrieval from the literature data articles. The citation data is essential for bibliographic coupling analysis and document co-citation analysis. The process included using a variety of computer programs and big amount of manual work. In the following sections, this process is described in detail and emerging challenges are highlighted.

Step one: exporting the citation data

The starting point for citation data acquisition was the 2154 articles in WoS (68%) and 997 articles in Scopus (32%). The citation data was extracted from WoS and Scopus in the formats that suited BibExcel. The format was "plain text". The Scopus citation export data was in RIS format. After this, some data preparation processes were conducted separately in BibExcel for the exported citation files from WoS and Scopus. The process consisted of converting the original citation files into a format that is readable for BibExcel, then creating a file including the extracted citation data (bibliographies) [to see more detailed information on data preparation in BibExcel, see (Persson et al., 2009)]. Further, the citation data was appended from the WoS and Scopus records, which resulted in 205,372 citation lines. For the appended citation data, the trimming functionality in BibExcel was employed. DOI (digital object identifier) code was removed since the Scopus data mostly did not have this included.

To summarize the outcome, in these 3151 articles (WoS 2 154 + Scopus 997), there were altogether 205,372 references. The WoS data accounted for 152,700 citation lines and Scopus data 52,672 lines. This txt format file was exported to Excel for the next step: data unifying, which is the first and major part of data wrangling.

Before continuing on the process of unifying citation data, we will now report a few detailed observations of the acquired citation data. In order to run bibliometric analysis with designated software, the quality of the data becomes crucial. The bibliometric analysis is based on statistics and bibliometric software simply runs the analysis with the available data so the quality of analysis corresponds with the quality of the data. Document co-citation analysis uses bibliographies of the selected literature. Unfortunately, the data provided by the citation indexes WoS and Scopus are often "broken" (e.g., spelling errors), and needs fixing before analysis. There are also significant differences in terms of the details of the citation data between these two big citation indexes. In this study, citation data from WoS and Scopus for the analyses was combined as much as possible. In Fig. 1, there is an example of typical journal article citation data from WoS and Scopus in terms of the information that is included in

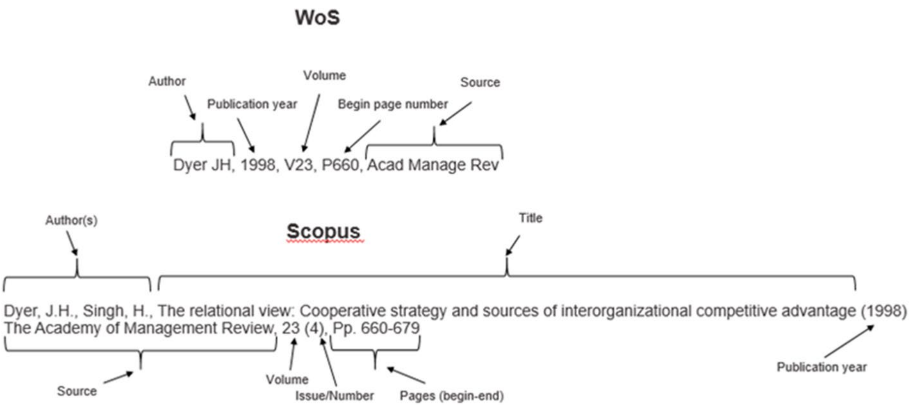
WoS
Author
Publication year
Volume
Begin page number
Source
Dyer JH, 1998, V23, P660, Acad Manage Rev
Author(s)
Scopus
Title
Dyer, J.H., Singh, H., The relational view: Cooperative strategy and sources of interorganizational competitive advantage (1998)
The Academy of Management Review, 23 (4), Pp. 660-679
Source
Volume
Issue/Number
Pages (begin-end)
Publication year

Fig. 1 Example of article citation. WoS version above and Scopus version below.

the citation. The example is Dyer and Sing's article "The Relational View: Cooperative Strategy and Sources of Interorganizational Competitive Advantage", published in the Academy of Management Review in 1998. The WoS citation version of this article is in the upper part of Fig. 1 and the Scopus citation version below.

The article was published in volume 23, issue/number 4 and the article begins on page 660 and ends on page 679 in this journal. The example's cited article is the same, yet the WoS and Scopus citation versions differ significantly. The WoS version is very brief compared to the Scopus version. The WoS citation gives only the first author of the reference, whereas Scopus lists all the authors (two in this case: Dyer J. H. & Singh H.). In addition, WoS does not mention the title of the reference. The WoS version includes the first author, publication year, volume number, beginning page number, and the source. The Scopus version includes all the authors, the title of the reference, publication year, source, volume, issue/number, and pages (beginning-end). It is also worth underlining that in the WoS version the source information is in abbreviated form (Acad Manage Rev in this case), whereas the Scopus version provides the whole name. The Scopus citation in this example includes all the information elements that are present in the WoS citation version, i.e., first author, publication year, volume, start page, and source. However, the source information poses a challenge, because there are no general rules to apply when turning a journal's full name into an abbreviated form.

There were several problems in the Scopus citation data that needed to be fixed. First, the special characters, e.g., Scandinavian letters (ä, å, ö) and letters of other alphabets (ñ, é, ß to name but a few) had turned into blurred characters in Excel since the character coding standards were not supported by the Microsoft Windows standards. All the special (blurred) characters were then sought in the Scopus citation data and corrected. The Scandinavian and other special characters were replaced with a non-special corresponding character, for example ä → a, ö → o, ñ → n, etc. In addition, the observed blurred characters that resulted from some punctuation marks, e.g., quotation marks, apostrophes, etc. were changed to characters understandable for Excel. Altogether over 6100 characters were replaced.

During the process, it was also realized that there were lines in the citation data that included more than one citation. The aim was to have each citation on a separate line in the citation data file. These citations were separated into individual lines, which resulted in 1350 new lines/citations.

Step two: extracting information from the Scopus citation string

The citation unifying (Scopus citation version to WoS citation version) process began with the analysis of the required information in the citation data. The citation information was in a single Excel cell. The target was to extract the necessary information from the Scopus citation string. The original Scopus citation string was divided into "pieces" and each of these "pieces" or extractions of the original string were in separate Excel cells. This demanded Excel coding or the executing of a series of Excel functions. Figure 2 illustrates the Scopus citation string with the target information for extraction circled.

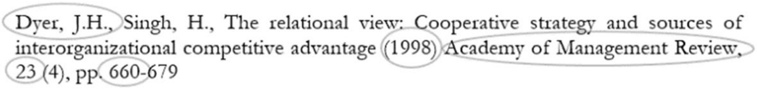
Dyer, J.H., Singh, H., The relational view: Cooperative strategy and sources of interorganizational competitive advantage (1998) Academy of Management Review, 23 (4), pp. 660-679

Fig. 2 The Scopus citation string, required information circled

First author: Scopus lists all the authors in its citation information (if applicable), starting with the last name of the first author and a comma plus the first letters of the author's first names and a period mark between the first letters, ending with a comma. The target was to extract the first author name, in this example, Dyer J. H. In Excel, the LEFT function was utilized and sought for the first combination of period mark and comma ".," which resulted in a cell with the text string Dyer, J. H. Next, the punctuation marks from the text string were trimmed with the SUBSTITUTE function in Excel.

Publication year: The publication year in the Scopus citation version is marked in brackets. The necessary information was extracted by a combination of MID and SEARCH functions in Excel, by searching for a text string having four characters in between the brackets ("(????)"). The presumption was of course that the citation reference was published earliest in the year of 1000!

Volume: The extraction of the volume number was done by quite complicated MID and FIND functions in Excel. A string fulfilling the condition that the volume number was provided in the composition with the issue/number (in brackets) was sought and the volume information before the issue/number was extracted.

Beginning page: Identifying the beginning page number was done in the first stage by the MID and FIND functions in Excel. The string "pp." was sought and the subsequent character string extracted. In the second stage, the number in front of the "–" character was separated by the LEFT and FIND functions.

Source: Extracting the source information was the most challenging task in the citation data unifying process. As mentioned earlier, the WoS citation consisted of the source (e.g., journal name) information in abbreviated form and unfortunately this abbreviation cannot be "calculated" purely with Excel functions, i.e., the logic behind the WoS abbreviations remained unsolved. Instead, a list from Caltech's webpage (https://www.library.caltech.edu/journal-title-abbreviations) that had a link to the full names of ISI journals and the corresponding abbreviations was used. The list also includes a set of conference proceeding names and corresponding abbreviations. In this list there were 84 599 full name–abbreviation pairs. This information was inserted in the citation data Excel file in another sheet, giving the full name and abbreviated form next to each other (this list is called "full name–abbreviated name"). The next phase was to extract the source information from the full citation string.

For the source string location, first the string after the publication year was extracted with the MID, FIND, LEN functions in Excel. Then, an output of the string before the ","-character was sought with the LEFT and FIND functions. With these procedures, the source information of the Scopus data was extracted into a separated cell. Next, the source information was matched with the "full name–abbreviated name" list using the VLOOKUP function in Excel.

The actions described above were effective for extracting the required information from the Scopus article citation string to the WoS article citation version. The extracted strings of the first author, publication year, volume, beginning page, and source were joined in one cell using the CONCATENATE function.

Finally, the Excel procedures resulted in 16,088 successful citation transformations into WoS format (out of 54,021), which is roughly 30% of all Scopus data citations. In order to achieve a successful transformation, the Excel functions had to return a valid value for the citation information fields: publication year, volume, page start, and source. The unsuccessful citation transformations (in terms of number of invalid values for the citation information fields) were publication year (1019), volume (28,263), page start (14,595), and source (26,755). No attention was paid to the author information field at this point, because resolving the author field defects was rather straightforward.

There were several reasons for the low score of successful transformations. The first and most obvious reason was that not all the citations of Scopus data were journal articles references (which the Excel procedures were aimed at). In principle, it could not be expected to reach a 100% transformation coverage. The data also included other types of citations: e.g., books, conference proceedings, web pages, and reports. BibExcel includes a reporting tool for roughly categorizing Scopus citations into three categories: "Journal", "Book," and "Don't know". According to this report, the division of Scopus citations between these categories was "Journals" approx. 75% (39,525 citations), "Books" approx. 22% (11,348 citations), and "Don't know" approx. 3% (1799 citations). (It is worth noting that this report could be applied for the original Scopus citation data before the manual splitting that resulted in 1350 new citation lines). During the unifying process it transpired that this division was not completely reliable, e.g., some citations marked "journal" were book citations and vice versa. However, a cautious conclusion may be made that the largest proportion of Scopus citation data was journal article citations and book citations accounted for the second largest group. The earlier described procedures in Excel were targeted only at Scopus journal article citations, leaving the book and other citations without treatment. The 16,088 transformed Scopus citations covered roughly 40% of the Scopus journal citations (39,525).

Although full coverage of citation data transformation was not expected, the proportion achieved was disappointingly low. The biggest problems in the extraction of the necessary information was related to volume information and source information. The beginning page number challenge was also significant, whereas the publication year represented a minor problem (which probably reflects scholarly habits in academic writing, too).

From the journal article citation perspective, Scopus data unfortunately included other kinds of citation methods than the example case described in the beginning this section (see Fig. 1, Scopus version). For example, there was a large number of journal article citations where the issue/number information was not included in the string, which makes the Excel function unable to extract the volume number information (see the "Volume" subsection above explaining that the Excel function extracted the string before the issue/number that was in brackets). This fact made the unifying work more challenging and created the need to execute other types of Excel operations to increase the coverage of successful article citation transformations.

The large quantity of unidentified sources was a big reason for unsuccessful transformations. This occurred when the source name was unidentified by the listing of journals' full names and their abbreviations (see above "Source" subsection explaining how the source name was transformed from the full name to abbreviated form). This in turn was due to two reasons. Either the "full name–abbreviated name" list did not include the specific source in question or the source spelling in the citation did not match the list. The latter reason underlines unfortunately the lack of standardization in expressing a source name.

To summarize the result so far, approximately 30% of the Scopus citations were successfully transformed into WoS format. The reasons for this low proportion were the fact that not all of the citations were journal articles (for which the unifying Excel functions were meant), unstandardized citation styles, and the incomplete list of full names of journals and their corresponding abbreviations. These conclusions marked the beginning of a big campaign for data supplementation. The focus was placed on journal article citations, because this category made up the majority of the whole.

Step three: Scopus citation source and volume information

The problem of source information defects in the data was due to two reasons: the incomplete “full name–abbreviated name” list and the various ways of spelling the source information. The unidentified source information was examined in detail, by viewing this information in a separate Excel file and sorting the source names. The target was to find the highest frequencies in the unidentified source names in the Scopus citation data.

Spelling problems

At this point it became clear that there were different ways to express the names of the sources. For example, there were many different types of errors in the Scopus data for the source (journal) “Administrative Science Quarterly”. Table 1 illustrates these errors.

There was a large number of different “possibilities” for spelling any journal name. The variations shown in Table 1 cover 110 citations in the data. The next stage in this case was to change these source spellings to the one that was in the “full name–abbreviated name” list. Although the number of variations was high in this specific case, the matching/correct name (Administrative Science Quarterly) was used in 980 citations. This example shows a glimpse of the data and the challenges in terms of transforming the citations into analogous form. It was noticed that in the Scopus citations abbreviations for the journal names were also used—or at least some kinds of abbreviations. In the literature data, there were 300 Scopus articles in which the citations had an abbreviated version of the journal name. However, this number contains uncertainty, because drawing a conclusion on whether the source is abbreviated or just incompletely informed is unclear.

A typical defect regarding the spelling source information was related to the use of the definite article “the” or indefinite article “a”/“an”, or leaving them unused. For example, the citation data included “The Accounting Review” or “The Journal of Supply Chain

Table 1 Errors in the “Administrative Science Quarterly” journal name (source) in the Scopus citation data

Source name in citation	Number of cases
Adm. Sci. Q	44
Adm Sci Q	14
Admin. Sci. Quart	14
Admin. Sci. Q	11
Admin. Sci. Q	7
Admin Sci Q	4
Administration Science Quarterly	3
Administrative Science Quarterlv	3
Administrative Science Quaterly	2
Adminstrative Science Quarterly	2
Adminis. Sci. Quart	1
Administative Sciences Quaterly Organization	1
Administratioe Science Quarterly	1
Administrative Science Ouarterly	1
Administrative Sciences Quarterly	1
AdministrativeScience Quarterly	1

Management", which were unidentified because of the "the" (i.e., in the "full name–abbreviated name" list these journals were spelled without the definite article). Also, the coordination conjunction "and" was a source of defects when the spelling in the "full name–abbreviated" list required an "&" mark. For example, "International Journal of Operations and Production Management" was spelled in the "full name–abbreviated name" list as "International Journal of Operations & Production Management". A significant amount of time was spent locating small spelling differences in vast amounts of citation data. The differences needed to be put into a form that resulted in the matching source information. Over 4 500 corrections were made to the spelling.

Incomplete "full name–abbreviated name" list

Another problem with the source information were the cases when the source (journal) was not included in the "full name–abbreviated name" list. For example, in the Scopus citation data there were over one hundred citations from the Journal of Marketing Management, which was not on the "full name–abbreviated name" list. In such a case, first a proportion of the references was manually validated by checking on the internet (e.g., Google Scholar) that the articles had in fact been published in the journal that was stated in the citation. Then, similar articles were searched in the WoS citation data. If the same articles could be identified in the WoS citation data, matching with the source's full name and abbreviation could be made. In the Journal of Marketing Management case, the WoS citation data included articles with the abbreviation "J Marketing Manageme". This pair was added to the "full name–abbreviated name" list.

The correcting of the source data defects was finally a task which was not tackled completely, due to the limited time and resources available. However, a substantial amount of corrections was made. In total 105 journal names were added to the "full name–abbreviated name" list. Thus over 2 300 citations were transformed into WoS citation form.

Volumes without issue or number information

The WoS citation format included the volume number in a journal article reference (see Fig. 1). The Excel function designed to extract the volume number from the original citation string was based on the assumption that the string was written in a form including the issue or number, e.g., "23 (4)". In this example "23" is the journal volume information and "4" is the issue or number in this specific volume. However, there was a significant amount of journal article citations that lacked the issue information. For these cases a new Excel function was designed that would suit journal article citations without issue or number information. This function was able to extract a section including the volume number from the original citation string. By applying an additional extraction, the bare volume information was separated. This was a breakthrough, because this treatment was successful for 13,121 Scopus citations.

Step four: Scopus book citation data, special cases of journal article citations, and author information

First, the most frequently occurring book citations were identified. Then, it was checked whether these book citations appeared in the WoS citation data. The form in which they appeared in the WoS citation data was naturally different compared to that of journal article citations. For example, Williamson's book "The Economic Institutions of Capitalism" from 1985 was cited in the WoS citation data as "Williamson O. E, 1985, EC I CAPITALISM". Hence, the WoS citation format contained the author, publication year, and the book title in an abbreviated form. The Scopus citation data versions of the same book reference varied. A typical version was: "Williamson, O.E., (1985) The Economic Institutions of Capitalism, The Free Press, New York, NY" again including much more information compared to WoS: author, year of publication in parentheses, full title of the book, publisher, and city of publication. The WoS version of the book title (EC I CAPITALISM) was added in the "full name–abbreviated name" list in order to have Excel perform the correct transformation. Let us take another example of a book citation in the Scopus data, Pfeffer and Salancik's book "*The external control of organizations*" published in 1978. In the WoS data, this book was cited as: "*Pfeffer J., 1978, EXTERNAL CONTROL ORG*". The WoS version contained only the first author, year of publication and the book title in a short format. The Scopus citation data contained variations of the reference, with a typical example being: "Pfeffer, J., Salanick, G.R., (1978) The external control of organizations: A resource dependence perspective, Harper and Row, New York". This citation includes both (all) authors, year of publication in parentheses, book's full title, publisher, and city of publication.

Citation transformation was done for 45 books altogether. For these book citations there were 248 different variations in the data. From the Scopus citation data, a total of 1 834 book citations were transformed into WoS citation format. The data also included book chapter citations. For these citations the page number information was added, which was the procedure in such cases according to the WoS citation data. The final Scopus data included 217 book chapter citations.

The data unifying process also revealed journal article citations without volume or page number information. The number of journal articles without volume information was 199 and the number without page numbering 27. These citations were simply left without the missing information. This was in line with the WoS citation data for similar articles.

There was after all a large proportion of author field problems in the citation data. In many cases the initial letter(s) of the author's first name(s) was missing or the author's name was lacking completely. In these cases, the missing information was searched and added manually. When comparing the starting point and the end of process, a total of 1 146 author name fields were corrected.

Summary of data wrangling results

Table 2 summarizes the Scopus data transformation work in terms of citation numbers. In total 36 482 Scopus citations were successfully transformed to the WoS citation format, which is over double compared to the starting point. During the process two citations were found to be located in one line. These were separated into two lines, which explains why there is a difference of one citation in the total Scopus citations between the starting point and final data. The successful transformations represent 67.5% of all the Scopus citations.

Our study showed that this process results in about two-thirds of successful Scopus citation transformations. The proportion of unsuccessful citation transformations includes various references: internet pages, reports, conference proceedings, and also journal articles and books. The transformation process was based on covering the most frequently occurring sources (journals and books), leaving the smaller proportions untreated.

Journal article citations represent most of the transformed Scopus citations: 94.4% when the journal article citations without volume or page number information are also included.

Table 2 Scopus citation data transformation success and the division of final data in terms of citation types

	Starting point	Final data
Scopus citations	54,021	54,022
Successful transformations	16,088	36,482
Unsuccessful transformations	37,933	17,540
Scopus citations in the final data		
Journal articles		34,285
Without volume number information		119
Without page number information		27
Book citations		1834
Book chapters		217
Total		36,482

Transformed book citations including book chapter citations account for the rest: 5.6%. It is difficult to estimate the proportion of successful journal article transformations of the whole Scopus citation data. Referring to the BibExcel report based on the original Scopus citation data of the division between "Journals", "Books," and "Don't know" (see chapter V. "Wrapping-up the citation string and identified troubles"), the coverage of successful journal article transformations was approx. 87% (34,431/39,525) with approx. 18% for books (2051/11,348). However, as discussed in section V, the reliability of the BibExcel report's figures was questionable.

Visualization of document co-citation analysis

In this section, the final steps of the process are described, and the visualizations of the selected analyses are presented. The process includes the data preparation to make it compatible with BibExcel and the visualization software.

All of the data, WoS data included, was uploaded into a file that was exported to BibExcel. The combined file contained 189 182 citation lines (Scopus 36 482 citation lines + WoS 152 700 citation lines). In order to run the data in BibExcel, the file had to be in.txt file. The citation data in Excel format was exported to Notepad++ and saved in the required file extension format. In BibExcel the data was trimmed by taking away the second initial letter of the author information from all citations. The intention was to minimize any possible confusion where two citations referring to the same document would be interpreted as two different documents because of a difference between the initial letters (e.g., Williamson OE vs. Williamson O). In addition, words written using only capital letters was trimmed into normal type (e.g., WILLIAMSON → Williamson).

The document co-citation analyses were run with BibExcel for the bare WoS citation data and the combined Scopus and WoS data. The process included creating a file with citation frequencies and performing a co-occurrence analysis against the cited reference strings. The computing capacity limited the carrying out of analyses with all the citation data. The document co-citation analyses were run on the references that had gained three or more citations. For the WoS citation data this meant 9387 citations and for the unified citation data 11,388 citations. Network files of the analyses were generated with BibExcel. The network files were uploaded to VOSviewer and the visualizations were run. Figure 3

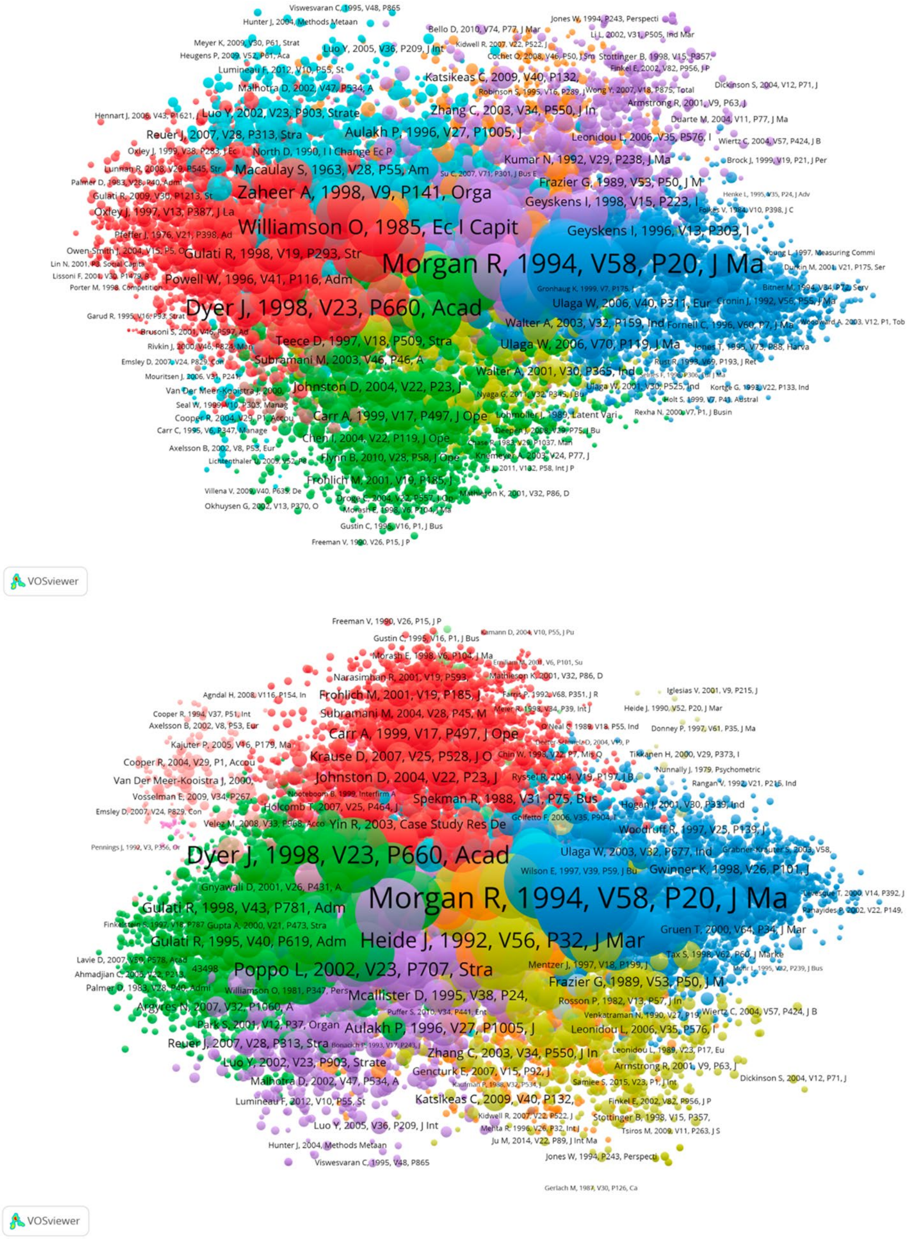

Fig. 3 Document co-citation visualization of the WoS citation data (above) and the combined citation data (below)

presents the visualization of the document co-citation analysis for the WoS citation data and the combined citation data.

The colors in the document co-citation network visualizations represent the clusters and the circles represent the documents. The document co-citation network visualizations consist of 9 386 items for WoS data and 11,337 items for the combined data. In the middle of both networks there is Morgan and Hunt's article "The Commitment-Trust Theory of Relationship Marketing". Both networks are shaped in quite a dense form, which could indicate a fairly consistent integrity of the inter-firm relationship research area. However, the WoS data document co-citation analysis has formed 10 document clusters and the combined data 14 document clusters. The difference in the numbers of clusters is an interesting result, although the greater number of documents may partly explain this difference. The document cluster sizes in terms of the number of documents is shown in Table 3.

Discussion and limitations

The objective of the study was to demonstrate the procedure of unifying the WoS and Scopus citation data into one combined dataset, and then to assess the meaningfulness engaging researchers in such a unifying process. These results offer also insights related to data wrangling in bibliometrics. The document co-citation analysis results between combined data and plain WoS data show minor differences in terms of the domains of inter-firm relationship literature. This result was mostly in-line with expectations, since WoS data dominates in the final dataset (about 81% of the combined citation data). However, the preliminary citation quantity analysis between the WoS data and combined data indicates that the highly cited references in the WoS data have received relatively more additional citations in the combined data than the more rarely cited references. Thus, the combined data has strengthened already "powerful" references in the WoS data. The WoS data was used as the reference data against which the Scopus data were compared. The selected methodology of transforming Scopus citation data into WoS form emphasizes the WoS data, because the automatic source name matching with the "full name–abbreviated name" list prioritized the journals that were included in WoS. This can be justified because WoS is often considered as the main source in academic research and Scopus is more often seen as a challenger (in citation indexes). However, more analysis is welcome, because in the WoS citation

Table 3 Document co-citation cluster sizes (numbers of documents per cluster)

Document cluster	WoS data	Combined data
1	2082	2714
2	1979	2533
3	1737	2148
4	1041	994
5	930	756
6	771	753
7	443	506
8	348	354
9	47	251
10	8	223
11		46
12		35
13		23
14		1

data there were such source journals that were not included in the “full name–abbreviated name” list.

A comparison of the document co-citation visualization between the WoS data and the combined data provides some interesting insights. For instance, the document co-citation analysis reveals more document clusters for the combined data, which could be interpreted as finding more fine-grained streams within the literature. In both clusterings, roughly 80% of the documents were divided between the five biggest document clusters (following the classic Pareto distribution). The marketing literature domain is rather well represented among the documents with the strongest links in the networks. For both datasets the top 10 documents are from the Journal of Marketing (5), Journal of Marketing Research (3), Academy of Management review (1), and Oliver E. Williamson’s book “The Economic Institutions of Capitalism” (1). This supports the a priori assumption (see Athanasopoulou, 2009) that the inter-firm relationship concept is strongly related to the marketing stream of scientific literature. Another a priori assumption was related to the richer data sources in the combined citation data compared to the WoS data. According to tentative analysis, the high end of the document co-citation link strengths shows no big differences in data. However, the possible differences are in the weaker document link strengths. This indicates an interesting research topic, i.e., to investigate smaller streams within the literature of inter-firm relationship research.

To the question of whether data wrangling makes sense in bibliometrics in light of the workload required, this study offers two complementing perspectives. First, this study demonstrates that the citation data itself includes many errors, for example, in terms of spelling and information completeness.

Conducting corrective actions on the erroneous data is justified as such. Second, before investing time and effort in corrective actions, the amount of data must be considered carefully. In case of smaller datasets, purification of data is do-able. However, in large datasets, based on the experiences of this study, it is suggested to either be aware of the data incompleteness and take this limitation into account when analyzing the results or create more effective data wrangling methods than have been presented in this study. Finally, this study shows that unifying data from two citation indexes and using them in analyses in bibliometrics software is possible. However, it is advised to consider carefully the amount of data before starting a similar unifying project.

This study identified four limitations that are worth mentioning. First, one of the objectives was to study the differences between WoS and Scopus citation indexes. The context was set as the concept of interfirm relationships since this context offers many publications in both databases. However, the literature area chosen does not show results in terms of highlighting differences between WoS and Scopus on a general level. Instead, this study merely gives insights into the differences between them, regarding the inter-firm relationship context. Further, analysis of the bare Scopus citation data was not done.

Second, the literature data search and the process of article inclusions and exclusions set some limitations on the results. One limitation concerns the article search. One search string (“inter-organizational relations”) was missing in the WoS search (detected in the analysis phase). The effects of this shortcoming are most probably minor, but rectifying has not yet taken place. It is also questionable whether this study covers all the data necessary for encompassing the inter-firm relationship literature. The search strings were selected in order to find all the relevant material, but it may be that some relevant aspects of the phenomenon are still missing. The evaluation and selection of search strings originated mainly from the marketing literature, which could have placed a bias on the study. The article inclusion and exclusion process does have its limitations. In the process, there were situations when deciding whether to include or exclude an article was difficult. The choices made naturally have potential minor effects on the results.

Third, the study comprises a lot of data processing. Although computers provide powerful tools, the risk of human error always exists. During the process some faults were identified in the unified data, for example "long" space characters in Excel had turned into erroneous characters in BibExcel. In addition, a systematic error was found in the unified citation data regarding the comma character. These errors were fixed but the possibilities for additional errors naturally remain.

Fourth, the utilized tools and their compatibility in terms of detailed properties have their limitations. The datasets processed with Excel, BibExcel, Notepad++, and VOSviewer were large. In many cases, the process was slowed because of a shortage of computing capacity. The computers that were used in the process were up to date, but still not completely capable of performing extremely large datasets. In addition, the programs used were perhaps not the most suitable for the sizes of the data. For example, the document co-citation calculation in BibExcel was an extremely long operation. The document co-citation analysis with the limitation on references that had gained three or more citations took several hours to perform.

Conclusions

As the contribution of this study, two main issues are emphasized regarding citations in scholarly work. First, the citing standards should be consolidated among research communities even more. More unified standards could have a positive effect on enhancing the quality of certain areas within bibliometric studies. For scholars and publishers, one piece of advice would be to pay more attention to the accuracy of reference expressions. This means, for example, having all the correct issues, volumes, and page numberings in place. By doing this, the journals and the articles would be better recognized by quantitative analyses, which would have an indirect positive effect on the journals' impacts and finally citations that any scholar would happily welcome. Finally, it is highly recommended to increase the coverage of DOI (digital object identifier) utilization. This information should be attached in citation data that is exported from citation indexes. An unambiguous code may eliminate most of the problems of the type of unifying endeavors described here.

Another contribution of this study is in describing the data unifying process and its challenges. The literature of inter-firm relationships represented the context of this study. There are four important topics for future research that would extend this study's results in terms of methodology and deepening knowledge of the inter-firm relationship literature. First, comparison of bibliographic coupling analysis of WoS and combined data would probably highlight certain differences in terms of connections between the journal articles, because Scopus literature data complements the total dataset with its share. Second, this study lacks a co-citation analysis of plain Scopus data. The comparison between the results of WoS, Scopus, and their combined data would enrich the results and clarify the differences between these citation indexes in the selected scope of inter-firm relationships. Third, the inter-firm relationship concept intersects various scientific domain areas. Other citation indexes (e.g., EbscoHost or IEEE), with their domain focus characteristics, could bring more specialized points of view to the literature study. Regarding data wrangling, there is room for improvement and further development. The Excel functions used delivered the results needed in terms of information field extractions. Finally, we suggest that the scientific consequences of the successful transformations would deserve more scholarly attention, since this study the efforts focused on the citation data transformation process and its challenges. Further research would focus on the question of what information do the transformed citation data carries.

Funding We would like to thank Business Finland and Future Electrified Mobile Machines (FEMMa) project for partial funding of this manuscript preparation. In addition, we would like to thank PhD Nees Jan van Eck in helping with VOSviewer's properties.

Data availability The data that support the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Anderson, J. C. (1995). Relationships in business markets: Exchange episodes, value creation, and their empirical assessment. *Journal of the Academy of Marketing Science, 23*(4), 346–350.

Anderson, J. C., & Narus, J. A. (1991). Partnering as a focused market strategy. *California Management Review, 33*, 95–113. https://doi.org/10.2307/41166663

Athanasopoulou, P. (2009). Relationship quality: A critical literature review and research agenda. *European Journal of Marketing, 43*(5/6), 583–610. https://doi.org/10.1108/03090560910946045

Cannon, J. P. J., & Perreault, W. D. (1999). Buyer–seller relationships in business markets. *Journal of Marketing Research, 36*(4), 439–460. https://doi.org/10.2307/3151999

Chang, Y. W., Huang, M. H., & Lin, C. W. (2015). Evolution of research subjects in library and information science based on keyword, bibliographical coupling, and co-citation analyses. *Scientometrics, 105*(3), 2071–2087. https://doi.org/10.1007/s11192-015-1762-8

Chen, C., Sanjuan, F. I., & Hou, J. (2010). The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. *Journal of the American Society for Information Science and Technology, 61*(7), 1386–1409.

Cuc, J. E. (2019). Trends in business model research: A bibliometric analysis. *Journal of Business Models, 7*(5), 1–24.

Denyer, D., & Tranfield, D. (2009). Producing a systematic review. In D. A. Buchanan & A. Bryman (Eds.), *The SAGE handbook of organizational research methods* (pp. 671–689). Sage Publications Ltd.

Dwyer, F. R., Schurr, P. H., & Oh, S. (1987). Developing buyer–seller relationships. *Journal of Marketing, 51*(2), 11–27. https://doi.org/10.2307/1251126

Dyer, J. H., & Singh, H. (1998). The relational view: Cooperative strategy and sources of interorganizational competitive advantage. *The Academy of Management Review, 23*(4), 660–679. https://doi.org/10.5465/AMR.1998.1255632

Ford, D. (1980). The development of buyer–seller relationships in industrial markets. *European Journal of Marketing, 14*(5/6), 339–353. https://doi.org/10.1108/EUM0000000004910

Forsman, M. (2016). *Julkaisut ja tieteen mittaaminen: Bibliometriikan käännekohtia*. Enostone.

Frazier, G. L. (1983). Interorganizational exchange behavior in marketing channels: A broadened perspective. *Journal of Marketing, 47*(4), 68–78. https://doi.org/10.2307/1251400

Golicic, S. L., & Mentzer, J. T. (2005). Exploring the drivers of interorganizational relationship magnitude. *Journal of Business Logistics, 26*(1), 47–72.

He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends, 48*(1), 133–159.

Jap, S. D., & Anderson, E. (2007). Testing a life-cycle theory of cooperative interorganizational relationships: Movement across stages and performance. *Management Science, 53*(2), 260–275. https://doi.org/10.1287/mnsc.1060.0610

Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics, 1*(4), 287–307. https://doi.org/10.1016/j.joi.2007.07.004

Kessler, M. M. (1963). Bibliographic coupling between scientific articles. *American Documentation, 24*(January), 123–131.

Leung, X. Y., Sun, J., & Bai, B. (2017). Bibliometrics of social media research: A co-citation and co-word analysis. *International Journal of Hospitality Management, 66*, 35–45. https://doi.org/10.1016/j.ijhm.2017.06.012

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics, 106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Morgan, R. M., & Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *Journal of Marketing, 58*(3), 20–38. https://doi.org/10.1177/1356766710391135

Persson, O., Danell, R., & Schneider, J. W. (2009). How to use Bibexcel for various types of bibliometric analysis. In F. Åström, R. Danell, B. Larson, & J. W. Schneider (Eds.), *Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th birthday (Issue Special volume of the e-zine of the ISSI)* (Vol. 5, pp. 9–24). International Society for Scientometrics and Informetrics.

Ring, P. S., & van de Ven, A. H. (1994). Developmental processes of cooperative interorganizational relationships. *The Academy of Management Review, 19*(1), 90–118.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*(3), 327–340.

Small, H. (1980). Co-citation context analysis and the structure of paradigms. *Journal of Documentation, 36*(3), 183–196.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14*, 207–222. https://doi.org/10.1111/1467-8551.00375

van Eck, N. J., & Waltman, L. (2019). VOSviewer manual. In *Leiden: Universiteit Leiden* (p. 52). http://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.1.pdf

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*(3), 163–171.

Wilson, D. T. (1995). An integrated model of buyer–seller relationship. *Journal of the Academy of Marketing Science, 23*(4), 335–345.

Wong, C., Wilkinson, I. F., & Young, L. (2010). Towards an empirically based taxonomy of buyer–seller relations in business markets. *Journal of the Academy of Marketing Science, 38*(6), 720–737. https://doi.org/10.1007/s11747-010-0191-8