Journal of Informetrics 11 (2017) 959–975

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

bibliometrix: An R-tool for comprehensive science mapping analysis

Massimo Aria[a,*], Corrado Cuccurullo[b]

[a] Department of Economics and Statistics, Università degli Studi di Napoli Federico II, Via Cintia, C.sso M.te S.Angelo, 80126 Naples, Italy
[b] Department of Economics and Management, Università della Campania Luigi Vanvitelli, Corso Gran Priorato di Malta, Capua, CE, Italy

ARTICLE INFO

Article history:
Received 14 February 2017
Received in revised form 27 August 2017
Accepted 27 August 2017

Keywords:
Bibliometrics
Science mapping
Workflow
Co-citation
Bibliographic coupling
R package

ABSTRACT

The use of bibliometrics is gradually extending to all disciplines. It is particularly suitable for science mapping at a time when the emphasis on empirical contributions is producing voluminous, fragmented, and controversial research streams. Science mapping is complex and unwieldy because it is multi-step and frequently requires numerous and diverse software tools, which are not all necessarily freeware. Although automated workflows that integrate these software tools into an organized data flow are emerging, in this paper we propose a unique open-source tool, designed by the authors, called *bibliometrix*, for performing comprehensive science mapping analysis. *bibliometrix* supports a recommended workflow to perform bibliometric analyses. As it is programmed in R, the proposed tool is flexible and can be rapidly upgraded and integrated with other statistical R-packages. It is therefore useful in a constantly changing science such as bibliometrics.

1. Introduction

The number of academic publications is increasing at a rapid pace and it is becoming increasingly unfeasible to remain current with everything that is being published. Moreover, the emphasis on empirical contributions has resulted in voluminous and fragmented research streams (Briner & Denyer, 2012). This hampers the ability to accumulate knowledge and actively collect evidence through a set of previous research papers. Therefore, literature reviews are increasingly assuming a crucial role in synthesizing past research findings to effectively use the existing knowledge base, advance a line of research, and provide evidence-based insight into the practice of exercising and sustaining professional judgment and expertise (Rousseau, 2012).

Scholars use different qualitative and quantitative literature reviewing approaches to understand and organize earlier findings. Among these, bibliometrics has the potential to introduce a systematic, transparent, and reproducible review process based on the statistical measurement of science, scientists, or scientific activity (Broadus, 1987; Diodato, 1994; Pritchard, 1969). Unlike other techniques, bibliometrics provides more objective and reliable analyses. The overwhelming volume of new information, conceptual developments, and data are the milieu where bibliometrics becomes useful by providing a structured analysis to a large body of information, to infer trends over time, themes researched, identify shifts in the boundaries of the disciplines, to detect the most prolific scholars and institutions, and to present the "big picture" of extant research (Crane, 1972).

* Corresponding author.
E-mail addresses: aria@unina.it, massimo.aria@unina.it (M. Aria), corrado.cuccurullo@unicampania.it (C. Cuccurullo).

http://dx.doi.org/10.1016/j.joi.2017.08.007
1751-1577/

Although over time, the use of bibliometrics has been extended to all disciplines, bibliometric analysis is complex because it entails several steps that employ numerous and diverse analyses and mapping software tools, which are frequently available only under commercial licenses (Guler, Waaijer, and Palmblad, 2016). These difficulties are compounded by the reality that few researchers and practitioners are trained in how to review literature and to identify evidence-based practices (Briner & Denyer, 2012). The cumbersome nature of the process reduces the possibilities and the potential of bibliometrics, especially for scholars who have no general programming skills.

Recently, automated workflows to assemble specialized software into a comprehensive and organized data flow have begun to emerge for bibliometrics. They are particularly well suited to multi-step analyses using different types of software tools (Guler, Waaijer, Mohammed, & Palmblad, 2016). In this paper, we propose a unique tool, developed in the R language, which follows a classic logical bibliometric workflow that we reconstruct. We have designed and produced an R-tool for comprehensive bibliometric analyses. R is a language and environment for statistical computing and graphics (R Core Team, 2016). It provides a wide variety of statistical and graphical techniques and is highly extensible (Matloff, 2011). In addition to enabling statistical operations, it is an object-oriented and functional programming language; hence, you can automate your analyses and create new functions. It has an open-software nature, which means it is well supported by the user community and new functions are regularly contributed by users, many of whom are prominent statisticians. As it is programmed in R, the proposed tool is flexible, can be rapidly upgraded, and can be integrated with other statistical R-packages. It is therefore useful in a constantly changing field such as bibliometrics.

The aim of this paper is twofold. First, we present the proposed open-source *bibliometrix* R-package for performing comprehensive bibliometric analyses, comparing it to other important software tools. Secondly, we discuss how the proposed tool supports a recommended workflow for performing bibliometric studies. We illustrate the main *bibliometrix* functions in this workflow, using all the articles written in English on bibliometrics in the management, business, and public administration domains over a span of 30 years.

2. Recommended workflow for science mapping

The general science mapping workflow was described by Börner, Chen, and Boyack (2003). Cobo, Lopez-Herrera, Herrera-Viedma, and Herrera, (2011a) compared science mapping software tools using a similar workflow. A standard workflow consists of five stages (Zupic & Čater, 2015):

1. Study design;
2. Data collection;
3. Data analysis;
4. Data visualization;
5. Interpretation.

In study design, scholars define the research question(s) and choose the appropriate bibliometric methods that can answer the question(s). Three general types of research questions can be answered using bibliometrics for science mapping: (i) identifying the knowledge base of a topic or research field and its intellectual structure; (ii) examining the research front (or conceptual structure) of a topic or research field; and (iii) producing a social network structure of a particular scientific community. In study design, one of the most significant choices for scholars is the timespan or decision to divide the timespan into time slices. Bibliometric analysis is performed at a specific point in time to represent a static picture of the field at that moment; it can divide the timespan into multiple time periods to capture the development of the field through time.

In data collection, scholars select the database that contains the bibliometric data, filter the core document set, and export the data from the selected database. This step can involve constructing one's own database (Waltman, 2016).

For data analysis, one or more bibliometric or statistical software tools are employed. Alternatively, scholars can write their own computer code to meet their requirements.

The fourth stage is data visualization. Scholars must decide what visualization method is to be used on the results of the third step and then employ the appropriate mapping software.

The last stage is interpretation, where scholars interpret and describe their findings. Although bibliometric methods will frequently reveal the structure of a field differently to the classification of traditional literature reviews, they are not a substitute for extensive reading in the field. Scholars with in-depth knowledge of the field have a clear distinctive advantage.

The second to fourth stages are typically software-assisted and include different sub-stages.

2.1. Data collection

Data collection is divided in three sub-stages. The first is data retrieval. Many online bibliographic databases, where metadata regarding scientific works are stored, can be sources of bibliographic information, such as Clarivate Analytics Web of Science (WoS at http://www.webofknowledge.com), Scopus (http://www.scopus.com), Google Scholar (http://scholar. google.com), and Science Direct (http://www.sciencedirect.com/) (Cobo et al., 2011a). They do not cover the scientific fields and journals in the same manner and hence the choice is not neutral (Waltman, 2016; Zupic & Čater,

Table 1
Most common bibliometric techniques per unit of analysis (adapted by Cobo et al., 2011).

Bibliometric technique taxonomy	Unit of analysis used	Kind of relation
Bibliographic Coupling	• Author • Document • Journal	• Common references in authors' oeuvres • Common references in documents • Common references in journals' oeuvres
Co-citation	• Author • Reference • Journal	• Co-cited authors • Co-cited documents • Co-cited journals
Co-author	• Author • Country from affiliation • Institution from affiliation	• Co-occurrence of authors in the author list of a document • Co-occurrence of countries in the address list of a document • Co-occurrence of institutions in the address list of a document
Co-word	• Keyword, or term extracted from title, abstract or document's body	• Co-occurrence of terms in a document

2015). Other similar databases exist for specific disciplines (e.g., Medline, Astrophysics Data System), patent data, and digital materials (e.g., arXiv, DBPL, CiteSeerXPatent).

The second sub-stage is data loading and converting, where scholars must convert data into a suitable format for the employed bibliometric tools.

The final sub-stage is data cleaning. The quality of the result depends on the quality of the data. Several preprocessing methods can be applied, for example, to detect duplicate and misspelled elements. Although the majority of bibliometric data are reliable, cited references can contain multiple versions of the same publication and different spellings of an author's name. Moreover, because authors are typically abbreviated by their surname and initials, a problem can arise with common names. Cited journals can also appear in slightly different forms. Books have different editions, which can appear as different citations.

2.2. Data analysis

Data analysis entails descriptive analysis and network extraction. Different approaches have been developed to extract networks using different units of analysis (Table 1). For example, co-word analysis (Callon, Courtial, Turner, & Bauin, 1983) uses the most important words or keywords of documents to study the conceptual structure of a research field. It is the only method that uses the actual content of the documents to construct a similarity measure; the others connect documents indirectly through citations. Co-word analysis produces semantic maps of a field that facilitate the understanding of its cognitive structure. It can be applied to document keywords, abstracts, or full texts. The unit of analysis is usually a concept or keyword, not a document, author, or journal. Another common bibliometric analysis is co-author analysis, which examines the authors and their affiliations to study the social structure and collaboration networks (Glänzel, 2001; Peters & Van Raan,1991). The most common analysis in bibliometrics is citation analysis. It employs citation counts as a measure of similarity between documents, authors, and journals. Citation analysis can be decomposed into bibliographic coupling and co-citation analysis. Examples are author coupling (Zhao & Strotmann, 2008), author co-citation (White & McCain, 1998; White & Griffith, 1981), journal coupling (Gao & Guan, 2009; Small & Koenig, 1977; Yan & Ding, 2012), and journal co-citation (McCain, 1991).

A bibliographic coupling connection is established by the authors of the articles in question, whereas a co-citation connection is established by the authors who are citing the documents analysed. That is, bibliographic coupling (Kessler, 1963) analyses the citing documents, whereas co-citation analysis (Small, 1973) studies the cited documents. Although bibliographic coupling is helpful in detecting the connections of research groups (Yang, Han, Wolfram, & Zhao, 2016), co-citation analysis, when examined over time, is also helpful in detecting a shift in paradigms and schools of thought. The choice of the technique to employ depends on the goals of the analysis. Usually, co-citation analysis is performed for mapping older papers (prospective analysis – it is dynamic and is best performed within different time slices), whereas bibliographic coupling is used to map a current research front (retrospective analysis – it does not change over time). Recently, Klavans and Boyack (2017) suggested that direct citations are more accurate in representing a research front than bibliographic coupling and co-citation.

Once the network has been built, a normalization process can be commonly performed over the relations (edges) between its nodes (vertices) using similarity measures such as Salton's cosine, Jaccard's coefficient, and Pearson's correlation.

Finally, data reduction is helpful in identifying subfields. With the normalized data, different techniques can be used to build the map. Various dimensionality reduction techniques can be applied, such as principal component analysis/factor analysis, multidimensional scaling (MDS), multiple correspondence analysis (MCA), and clustering algorithms.

2.3. Data visualization

Analysis methods allow the extraction of useful knowledge from data and to represent it through intuitive visualizations or maps such as bi-dimensional maps, dendrograms, and social networks. Network analysis allows us to perform a statistical analysis over the maps generated to indicate different measures of the entire network or measures of the relationship or the overlapping of the different clusters detected.

Visualization techniques are used to represent a science map and the result of the different analyses. For example, networks can be represented using heliocentric maps (de Moya-Anegon et al., 2005), geometrical models (Skupin, 2009), thematic networks (Bailón-Moreno, Jurado-Alameda, & Ruiz-Baños, 2006; Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2011b), or maps where the proximity between items represents their similarity (van Eck & Waltman, 2010). Alternatively, temporal analysis aims to indicate the conceptual, intellectual, or social evolution of the research field by discovering patterns, trends, seasonality, and outliers. Burst detection, a temporal analysis, aims to identify features that have high intensity over finite durations of time periods. To demonstrate the evolution in different time periods, cluster strings (Small, 2006; Small & Upham, 2009; Upham & Small, 2010) and thematic areas (Cobo et al., 2011b) can be used. Finally, geospatial analysis aims to discover where an event occurs and its impact on the neighbouring areas.

3. Related bibliometric software tools

3.1. Software tools for science mapping

Numerous software tools support bibliometric analysis; however, many of these do not assist scholars in a complete recommended workflow. The most relevant tools are CitNetExplorer (van Eck & Waltman, 2014), VOSviewer (van Eck & Waltman, 2010), SciMAT (Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2012), BibExcel (Persson, Danell, & Schneider, 2009), Science of Science (Sci2) Tool (Sci2 Team, 2009), CiteSpace (Chen, 2006), and VantagePoint (www.thevantagepoint.com).

CitNetExplorer and VOSviewer are two free Java applications, designed by van Eck and Waltman, for analysing and visualizing citation networks of scientific collections. CitNetExplorer allows the user to (i) analyse the development of a research field over time, (ii) identify the core literature on a research topic, and (iii) explore the publication oeuvre of a researcher and its influence on the publications of other researchers. VOSviewer addresses the graphical representation of bibliometric maps and is especially useful for displaying large bibliometric maps in an easy-to-interpret manner.

SciMAT is an open source software tool developed to perform a science mapping analysis under a longitudinal framework. SciMAT provides three different modules: (i) management of a knowledge base and its entities; (ii) science mapping analysis; and (iii) visualization of the generated results.

BibExcel is designed to assist a scholar in analysing bibliographic data, or any data of a textual nature formatted in a similar manner. It generates data files that can be imported into Excel, or any program that accepts tabbed data records, for further processing. However, BibExcel does not include any module to visualize and map the results.

The Science of Science (Sci2) Tool is free software that supports the temporal, geospatial, topical, and network analysis and visualization of bibliographic collections.

CiteSpace is a free Java application for visualizing and analysing trends and patterns in scientific literature. It focuses on identifying critical points in the development of a field or a domain, especially intellectual turning points and pivotal points.

VantagePoint is commercial software for science mapping analysis. Its major strength is the ability to read virtually any structured text content. It supports more than 190 different import filters. Moreover, VantagePoint includes a tool for visualizing the main bibliometric maps.

3.2. R-packages for bibliometric analysis

In the R environment, other packages have been published recently on the official repository (CRAN, *The Comprehensive R Archive Network*, https://cran.r-project.org/) addressing bibliometrics. Each of them provides for specific analysis functions; however, none addresses the entire workflow. For example, the primary aim of CITAN (Gagolewski, 2011) – CITation ANalysis package for R statistical computing Environment is to support scholars with a tool (i) for preprocessing and cleaning bibliographic data retrieved from Scopus and (ii) for calculating the most popular indices of scientific impact. Moreover, CITAN provides metrics such as h-index, g-index, and L-index. Unlike *bibliometrix*, CITAN (i) can use only data from Scopus and (ii) has no functions for co-citation, bibliographic coupling, scientific collaboration, co-word analysis, or text extraction from titles and abstracts.

ScientoText (Uddin, 2016) is another recent package that is perhaps the most comparable to the *bibliometrix* R-package. Nevertheless, although ScientoText states that it uses data from the WoS and Scopus databases, it currently has no functions for importing and converting data.

H-index Calculator (Alavifard, 2015) uses only data from the Clarivate Analytics WoS for calculating the h-index.

Finally, Scholar (Keirstead, 2015) offers similar functionalities as the well-known software tool Publish or Perish (Harzing, 2007). It enables data to be extracted from Google Scholar for one or more researchers for analysing citations and calculating certain impact metrics. As with CITAN, Scholar does not include any function for co-citation, bibliographic coupling, scientific

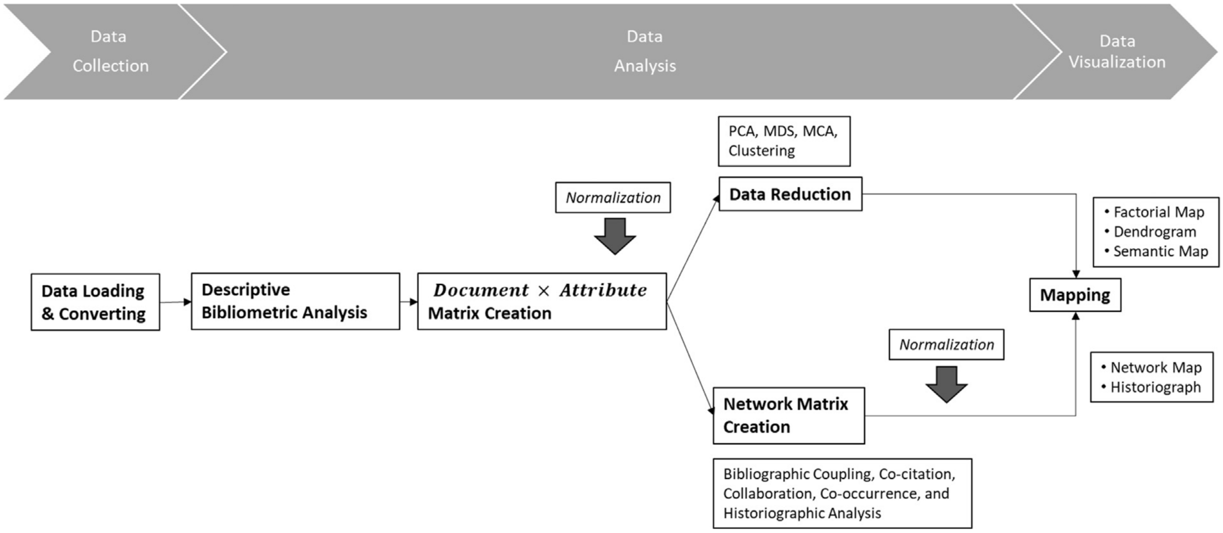

Fig. 1. *bibliometrix* and the recommended science mapping workflow.

collaboration, co-word analysis, or text extraction from titles and abstracts. Moreover, it can only use data from Google Scholar with all the limitations with respect to the WoS and Scopus databases (Bar-Ilan, 2007; Yang & Meho, 2006).

4. *bibliometrix* and the recommended science mapping workflow

The *bibliometrix* R-package (http://www.bibliometrix.org) provides a set of tools for quantitative research in bibliometrics and scientometrics. It is written in the R language, which is an open-source environment and ecosystem. The existence of substantial, effective statistical algorithms, access to high-quality numerical routines, and integrated data visualization tools are perhaps the strongest qualities to prefer R to other languages for scientific computation.

Fig. 1 illustrates the bibliometrix workflow supporting the second through fourth stages of the recommended science mapping workflow presented in Section 2.

1. Data collection. *bibliometrix* supports the following sub-stage:
   a Data loading and conversion to R data frame (Section 4.1).
2. Data Analysis, articulated in three sub-stages:
   a Descriptive analysis of a bibliographic data frame (Section 4.2);
   b Network creation for bibliographic coupling, co-citation, collaboration, and co-occurrence analyses (Section 4.3);
   c Normalization (Section 4.4).
3. Data visualization:
   a Conceptual structure mapping (Section 4.3e);
   b Network mapping (Section 4.5).

To describe the main functions of *bibliometrix* (Table 2), we analysed articles on bibliometrics in the management and business fields between 1985 and 2015. The data is available on the *bibliometrix* website (http://www.bibliometrix.org/datasets/bibliometric_management_business_pa.txt).

4.1. Data loading and converting to R data frame

Data collection is a task composed of different subtasks as follows.

- Data retrieval. *bibliometrix* functions with data extracted from the two main bibliographic databases, namely Clarivate Analytics WoS and Scopus. The *bibliometrix* tutorial (http://www.bibliometrix.org/index.html#header3-16) assists scholars with querying these databases. Moreover, *bibliometrix* connects with the Scopus API to automatically collect metadata regarding the complete scientific production of a list of scholars. In the example that follows, we used Clarivate Analytics – Web of Science core collection – Social Sciences Citation Index (SSCI) and Science Citation Index Expanded (SCI-Expanded) and chose (1) the generic keyword "bibliometric*" as the topic, (2) only articles written in English for the document type, (3) "management", "business", and "public administration" as subject categories, and (4) the timespan 1985–2015.
- Data loading and converting (*hereinafter, square brackets denote the R syntax for commands*). The export files are read by R using the *readFiles* function [D <-

Table 2
Main *bibliometrix* functions.

Software assisted workflow steps	bibliometrix function	Description	Output
Data loading and converting	• readFiles()	• Loads a sequence of Scopus and Clarivate Analytics WoS export files into R	• Bibliographic data frame
	• Convert2df()	• Creates a bibliographic data frame	
	• retrievalByAuthorID()	• Uses Scopus API search to obtain information regarding documents on a set of authors using Scopus ID	
Descriptive bibliometric analysis	• biblioAnalysis()	• Returns an object of class bibliometrix	• Tables of results
	• Summary() and plot()	• Summarize the main results of the bibliometric analysis	
	• citations()	• Identifies the most cited references or authors	
	• localCitations()	• Identifies the most cited local authors	
	• dominance()	• Calculates the authors' dominance ranking	
	• Hindex()	• Measures productivity and citation impact of a scholar	
	• lotka()	• Estimates Lotka's law coefficients for scientific productivity	
	• keywordGrowth()	• Calculates yearly cumulative occurrences of top keywords/terms	
	• keywordAssociation()	• Associates authors' keywords to keywords plus	
Document x Attribute matrix creation	• metaTagExtraction()	• Extracts other field tags, different from the standard WoS/Scopus codify	• Document x Attribute matrix
	• termExtraction()	• Extracts and stems terms from textual fields (abstract, title, author's keywords, and others) of a bibliographic data frame	
	• cocMatrix()	• Computes a Document x Attribute matrix	
Normalization	• normalizeSimilarity()	• Calculates association strength, inclusion index, Jaccard's coefficient, and Salton's similarity coefficient among objects of a bibliographic network	• Similarity matrix
Data Reduction	• conceptualStructure()	• Creates conceptual structure map of a scientific field using MCA and Clustering	• Word occurrence matrix, MCA, and clustering results
Network matrix creation	• biblioNetwork()	• Calculates the most frequently used bibliographic coupling, co-citation, collaboration, and co-occurrence networks	• Network matrix and historical network matrix
	• histNetwork()	• Creates a historical co-citation network from a bibliographic data frame	
Mapping	• networkPlot()	• Plots a bibliographic network using internal R library or VOSviewer software	• Network graph, network Pajek format for VOSviewer, Historiograph, and semantic map
	• histPlot()	• Plots a historical co-citation network	
	• conceptualStructure()	• Plots conceptual structure map of a scientific field using MCA and Clustering	

Table 3
bibliometrix data frame structure.

Field Tag	Class	Description
UT	CHARACTER	Unique Article Identifier
AU	CHARACTER	Authors
TI	CHARACTER	Document Title
SO	CHARACTER	Publication Name (or Source)
JI	CHARACTER	ISO Source Abbreviation
DT	CHARACTER	Document Type
DE	CHARACTER	Authors' Keywords
ID	CHARACTER	Keywords associated by WoS or Scopus database
AB	LARGE CHARACTER	Abstract
C1	CHARACTER	Author Address
RP	CHARACTER	Reprint Address
CR	LARGE CHARACTER	Cited References
TC	NUMERIC	Times Cited
PY	NUMERIC	Year
SC	CHARACTER	Subject Category
DB	CHARACTER	Bibliographic Database

Table 4
Element list of a *bibliometrix* object.

List element	*Description*
Articles	Total number of documents
Authors	Authors' frequency distribution
AuthorsFrac	Authors' frequency distribution (fractionalized)
FirstAuthors	First author of each document
nAUperPaper	Number of authors per document
Appearances	Number of author appearances
nAuthors	Total number of authors
AuMultiAuthoredArt	Number of authors of multi-authored articles
Years	Publication year of each document
FirstAffiliation	Affiliation of the first author for each document
Affiliations	Frequency distribution of affiliations (of all co-authors for each document)
Aff_frac	Fractionalized frequency distribution of affiliations (of all co-authors for each paper)
CO	Affiliation country of first author
Countries	Affiliation countries' frequency distribution
TotalCitation	Number of times each document has been cited
TCperYear	Yearly average number of times each document has been cited
Sources	Frequency distribution of the sources (journals, books, others)
DE	Frequency distribution of the authors' keywords
ID	Frequency distribution of keywords associated to the document by Clarivate Analytics Web of Science and Scopus databases

readFiles(http://www.bibliometrix.org/datasets/bibliometric_management_business_pa.txt)] that creates a large character object called *D*. The function supports plain text (for Clarivate Analytics database) and BibTex (for both Clarivate Analytics and Scopus databases) formats and allows importing simultaneously multiple export files. These can be converted into a data frame using the *convert2df* function [M <- convert2df(D, dbsource = "isi", format = "plaintext")]. *convert2df* creates a bibliographic data frame with cases corresponding to documents and variables to field tags in the original export file. Each document contains several elements such as authors' names, title, keywords and other information. These elements constitute the bibliographic attributes of a document, also called the metadata. We have chosen to use standard column names for the bibliographic data frame adopting the field tags proposed by Clarivate Analytics and for Scopus collections. This facilitates merging different sources and applying R routines. Table 3 contains the structure of the *bibliometrix* data frame considering the main field tags. The column "class" reports the data type of each data frame column.

- Data cleaning. *bibliometrix* does not have specific routines dedicated to data cleaning. It does include in its main functions (e.g., loading and converting, citation analysis) a set of cleaning rules such as: (i) transform full text into uppercase, (ii) remove non-alphanumeric characters, (iii) remove punctuation symbols and extra spaces, and (iv) truncate author's first and middle names to the initials.

4.2. Descriptive analysis of a bibliographic data frame

The descriptive analysis of the bibliographic data frame uses many functions.

- The *biblioAnalysis* function calculates the main bibliometric measures using simple syntax [results <- biblioAnalysis(M, sep = ";")].The *biblioAnalysis* function returns an object of class "*bibliometrix*", which is a list containing the elements reported in Table 4.

Table 5
Descriptive analysis: Main information regarding the collection.

Description	
Articles	304
Period	1985–2015
Annual Percentage Growth Rate	12.40
Average citations per article	26.56
Authors	617
Author Appearances	801
Authors of single authored articles	32
Authors of multi authored articles	585
Articles per Author	0.493
Authors per Article	2.03
Co-Authors per Articles	2.63
Collaboration Index	2.49

Table 6
Descriptive analysis: Top 10–Most productive authors.

Author	*No. of Articles*	*Author*	*No. of Articles Fractionalized*
Kostoff RN	16	*Kostoff RN*	7.77
Kajikawa Y	9	*Vogel R*	3.50
Porter AL	9	*Porter AL*	3.46
Abramo G	5	*Kajikawa Y*	3.00
D'Angelo CA	5	*Shilbury D*	3.00
Moed HF	5	*Talukdar D*	2.33
Bowles CA	4	*Hicks D*	2.08
Hicks D	4	*Eom SB*	2.00
Lee PC	4	*Mcmillan GS*	2.00
Sakata I	4	*Saetren H*	2.00

Table 7
Descriptive analysis: Top 10–Most cited papers.

Paper	Total Citations (TC)	TC per Year
Chen HC, Chiang RHL, Storey VC, (2012), Mis Q.	386	77.20
Daim TU, Rueda G, Martin H, Gerdsri P, (2006), Technol. Forecast.Soc. Chang.	240	21.82
Moed HF, Burger WJM, Frankfort JG, Vanraan AFJ, (1985), Res.Policy	232	7.25
Kostoff RN, Scaller RR, (2001), IEEE Trans. Eng. Manage.	220	13.75
Loh L, Venkatraman N, (1992), Inf. Syst. Res.	210	8.40
Volberda HW, Foss NJ, Lyles MA, (2010), Organ Sci.	202	28.86
Ramos-Rodriguez AR, Ruiz-Navarro J, (2004), Strateg. Manage. J.	190	14.62
Murray F, (2002), Res. Policy	187	12.47
Melin G, (2000), Res. Policy	187	11.00
Gambardella A, (1992), Res. Policy	139	5.56

Table 8
Descriptive analysis: Top 10–Most productive countries (based on first author's affiliation).

Country	*No. of Articles*	*% of Articles*
USA	89	29.5
Netherlands	25	8.3
England	21	6.9
Germany	20	6.6
Italy	20	6.6
Japan	15	5.0
Spain	15	5.0
Sweden	10	3.3
Taiwan	10	3.3
Australia	8	2.7

- The functions *summary* and *plot* summarize the main results of the bibliometric analysis. They display the principal information regarding the bibliographic data frame and six tables. *summary* accepts two additional arguments: *k* is a formatting value that indicates the number of rows for each table; *pause* is a logical value (TRUE or FALSE) used to permit (or not) a pause in screen scrolling. For example, choosing $k = 10$, we expressed the desire to view the first ten authors or first ten sources. The results are displayed in Tables 5–10 and in Fig. 2.
- The *citations* function generates the frequency table of the most cited references or the most cited first authors (of references). For each document, cited references are in a single string stored in the "CR" column of the data frame. For a correct

Table 9
Descriptive analysis: Top 10–Most frequent journals.

Sources	*No. of Articles*	*% of Articles*
Research Policy	58	19.1
Technological Forecasting and Social Change	56	18.4
Technology Analysis & Strategic Management	17	5.6
Technovation	15	4.9
International Journal of Technology Management	6	2.0
R & D Management	6	2.0
Science and Public Policy	6	2.0
African Journal of Business Management	5	1.6
Journal of Technology Transfer	5	1.6
Journal of Business Ethics	4	1.3

Table 10
Descriptive analysis: Top 10–Most frequent keywords.

Author Keywords (DE)	*No. of Articles*	*Keywords-Plus (ID)*	*No. of Articles*
Bibliometrics	87	*Science*	82
Bibliometric Analysis	30	*Innovation*	39
Citation Analysis	26	*Performance*	33
Nanotechnology	17	*Journals*	30
Research	17	*Technology*	30
Innovation	16	*Impact*	28
Analysis	13	*Knowledge*	26
Scientometrics	13	*Management*	26
Text Mining	10	*Bibliometrics*	25
Patent Analysis	9	*Citation Analysis*	25

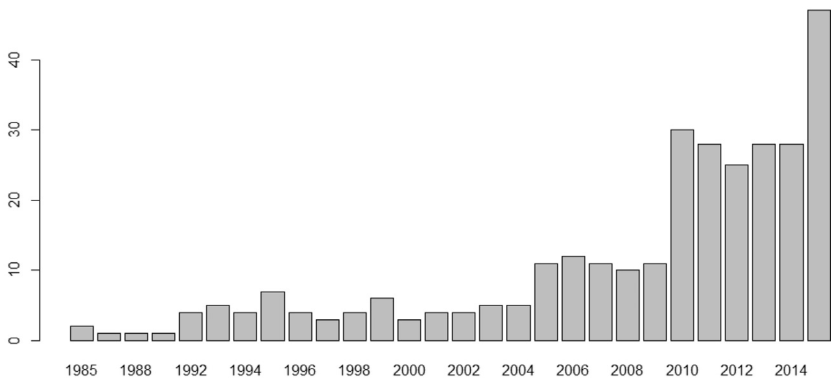

Fig. 2. Publications per year 1985–2015.

extraction, we must identify the separator field among different references used by the selected database. Typically, the WoS default separator is ";". The *bibliometrix* tutorial also describes other separators.

Cited references frequently have numerous inconsistencies in the data format. For example, some databases, such as Scopus, do not have a standardized format. The *citations* function also implements a set of cleaning rules as described in Section 4.1.

Table 11 contains the most frequently cited documents [CR <- citations(M, field = "article", sep = ";")].The *localCitations* function [CR <- localCitations(M, results, sep = ";")] generates the frequency table of the most local cited authors. Local citations measure how many times an author included in this collection has been cited by other authors also in the collection. Table 12 reports the most frequent local authors.

- The authors' h-index is an author-level metric that attempts to measure both the productivity and citation impact of the publications of a scientist or scholar (Hirsch, 2005). The index is based on the set of the scientist's most cited papers and the number of citations that they have received in other publications. The *Hindex* function calculates the authors' h-index and its variants (g-index and m-index) in a bibliographic collection (van Eck & Waltman, 2008). Function arguments are the following: *M* a bibliographic data frame and *authors* a character vector containing the authors' names for which you want to calculate the h-index. For example, to calculate the h-index, in this collection, for author Ronald Kostoff, you would use [Hindex(M, authors = "KOSTOFF R", sep = ";")]

Table 11
Citation analysis: Top 10–Most cited references.

Cited Reference	Citations
Ramos-Rodriguez AR, 2004, Strategic Manage J, V25, P981, Doi 101002/Smj397	32
Small H, 1973, J Am Soc Inform Sci, V24, P265, Doi 101002/Asi4630240406	29
Mccain KW, 1990, J Am Soc Inform Sci, V41, P433	26
Nelson RR, 1982, Evolutionary Theory	22
White HR, 1981, J Am Soc Inform Sci, V32, P163, Doi 101002/Asi4630320302	21
Price DJ, 1963, Little Sci Big Sci	19
*Cohen W*M, 1990, Admin Sci Quart, V35, P128, Doi 102307/2393553	18
*Daim T*U, 2006, Technol Forecast Soc, V73, P981, Doi 101016/Jtechfore200604004	18
*Hoffman D*L, 1993, J Consum Res, V19, P505, Doi 101086/209319	18
Nerur SP, 2008, Strateg Manage J, V29, P319, Doi 101002/SMJ659	18
Small H, 1974, Sci Stud, V4, P17, Doi 101177/030631277400400102	18
White HD, 1998, J Am Soc Inform Sci, V49, P327	18

Table 12
Local citation analysis: Top 10–Most cited authors.

Local Cited Author	*Citations*
Kostoff RN	317
Narin F	103
Porter AL	96
Leydesdorff L	84
Moed HF	71
Pilkington A	53
Hicks D	48
Meyer M	47
Martin B	45
Pavitt K	43

4.3. Network creation for bibliographic coupling, co-citation, collaboration, and co-occurrence analyses

A document's attributes are connected to each other through the document itself (e.g., author(s) to journal, keywords to publication date). These connections of different attributes can be represented through a matrix *Document* × *Attribute*.

cocMatrix is the general function to create a rectangular matrix *Document* × *Attribute* that we call *A*. An attribute is an item of information associated to the document and stored in a field tag within the bibliometric data frame (e.g., authors, publication source, keywords, cited references, affiliations).

In some cases, this matrix can be interpreted as a bipartite or two-mode network (e.g., where attribute is author, keyword, cited reference). For example, to create a *Document* × *Cited reference* matrix, we must use the field tag "CR" [A <- cocMatrix(M, Field = "CR", sep = ";")]. In this case,*A* is a rectangular binary matrix (and also a bipartite network) where each row is a document and each column concerns a cited reference of the collection. The generic element $a_{ij}$is "1" if the document *i* has cited the reference *j*, otherwise it is "0". The *j-th* column sum a_{+j} is the number of documents citing the reference *j*. The *i-th* row sum a_{i+} is the number of references cited by document *i*.

Using the *cocMatrix* function, several matrices can be computed, such as:

- *Document* × *Author* [A <- cocMatrix(M, Field = "AU", sep = ";")];
- *Document* × *Country*. Authors' Countries is not a standard attribute of the bibliographic data frame. We must extract this information from the affiliation attribute using the *metaTagExtraction* function [M <- metaTagExtraction(M,Field = "AU_CO",sep = ";");A <- cocMatrix(M, Field = "AU_CO", sep = ";")]. *metaTagExtraction* allows the following additional field tags to be extracted: Authors' countries(Field = "AU_CO"), First author of each cited reference *(Field = "CR_AU")*, Publication source of each cited reference *(Field = "CR_SO")*, and affiliation for each co-author *(Field ="AU_UN");*
- *Document* × *Authors' keyword* [A <- cocMatrix(M, Field = "DE", sep = ";")] or *Document x Keyword Plus* [A <- cocMatrix(M, Field = "ID", sep = ";")].

a) Bibliographic coupling

Two articles are said to be bibliographically coupled if at least one cited source appears in the bibliographies or reference lists of both articles (Kessler, 1963). A bibliographic coupling network can be obtained using the general formula:

$$B_{coup} = A \times A'$$

where *A* is a *Document* × *Cited reference* matrix. Element b_{ij}indicates how many bibliographic couplings exist between documents *i*and *j*. B_{coup} is a non-negative and symmetrical matrix $B_{coup} = B'_{coup}$.

The strength of the bibliographic coupling of two articles, i and j is defined simply by the number of references that the articles have in common, as given by the element b_{ij} of matrix B_{coup}.

The *biblioNetwork* function calculates, starting from a bibliographic data frame, the most frequently used bibliographic coupling networks such as documents, authors, sources, keywords, and countries. To use *biblioNetwork* it is necessary to set two arguments. First, the type of analysis is set. In this case, the analysis argument is "coupling" (alternatively, "co-citation", "collaboration", and "co-occurrences"). Then, the network unit of analysis, which can be alternatively "authors", "references", "sources", "countries", "keywords", "author_keywords", "titles", or "abstracts" must be set.

The following code calculates a classical document bibliographic coupling network [NetMatrix <- biblioNetwork (M, analysis = "coupling", network = "references", sep = ";")]. Articles with only a small number of references, therefore, would tend to be more weakly bibliographically coupled when bibliographic coupling strength is simply measured according to the number of references that articles have in common.

b) Co-citation analysis

Co-citation of two articles occurs when both are cited in a third article. Thus, co-citation is the counterpart of bibliographic coupling. A co-citation network can be obtained using the general formula:

$$B_{cocit} = A \times A'$$

where A is a *Document* $\times$ *Cited reference* matrix.

Similar to matrix B_{coup}, matrix B_{cocit} is also symmetric. Element b_{ij} indicates how many co-citations exist between documents i and j. The main diagonal of B_{cocit} contains the number of documents where a reference is cited in our data frame. That is, the diagonal element b_{ii} is the number of local citations of the reference i. The *biblioNetwork* function provides a classical reference co-citation network [NetMatrix <- biblioNetwork(M, analysis = "co-citation", network = "references", sep = ";")].

c) Collaboration analysis

A scientific collaboration network is a network where nodes are authors and links are co-authorships. It is one of the most well-documented forms of scientific collaboration (Glänzel & Schubert, 2004). An author collaboration network can be obtained using the general formula:

$$B_{coll} = A \times A'$$

where A is a *Document* $\times$ *Author* matrix. Element b_{ij}indicates how many collaborations exist between authors $i$and j. The diagonal element $b_{ii}$is the number of documents authored or co-authored by researcher i. The *biblioNetwork* function calculates an authors' collaboration network [NetMatrix <- biblioNetwork(M, analysis = "collaboration", network = "authors", sep = ";")] or a country collaboration network [NetMatrix <- biblioNetwork(M, analysis = "collaboration", network = "countries", sep = ";")].

d) Co-word analysis

The aim of the co-word analysis is to draw the conceptual structure of a framework using a word co-occurrence network to map and cluster terms extracted from keywords, titles, or abstracts in a bibliographic collection [NetMatrix <- biblioNetwork(M, analysis = "co-occurrences", network = "keywords", sep = ";")].

A co-word network can be obtained using the general formula:

$$B_{coc} = A \times A'$$

where A is a *Document* $\times$ *Word* matrix, where *Word* is, alternatively, authors' keywords, keywords plus, or terms extracted from titles or abstracts. Element b_{ij} indicates how many co-occurrences exist between words i and j. The diagonal element b_{ii} is the number of documents containing the word i.

The *termExtraction* function extracts terms from a textual field (e.g., abstract, title, author's keywords), deletes stopwords, and applies Porter's stemming algorithm (Porter, 1980). Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form, typically a written word form. Hence, this function normalizes terms before performing the co-occurrence analysis [M <- termExtraction(M, Field = "TI", stemming=TRUE, language="english", verbose=TRUE)].

The *bibliometrix* R-package allows using the *conceptualStructure* function to perform multiple correspondence analysis (MCA) to draw a conceptual structure of the field and K-means clustering to identify clusters of documents that express common concepts.

MCA is an exploratory multivariate technique for the graphical and numerical analysis of multivariate categorical data (Benzécri, 1982; Greenacre & Blasius, 2006; Lebart, Morineau, & Warwick, 1984). MCA performs a homogeneity analysis of an indicator matrix to obtain a low-dimensional Euclidean representation of the original data (Gifi, 1990). In co-word analysis, MCA is applied to a *Document x Word* matrix A. The words are plotted on a two-dimensional map [CS <- conceptualStructure(M, field="ID", minDegree=5, k.max=5, stemming=FALSE), labelsize=5]. The results are interpreted based on the relative positions of the points and their distribution along the dimensions; as words are more similar in distribution, the closer they are represented in the map (Fig. 3) (Cuccurullo, Aria, & Sarto, 2016).

Fig. 3. Conceptual map and keyword clusters.

4.4. Normalization

The *normalizeSimilarity* function allows the user to normalize bibliographic coupling, co-citation, and co-occurrence data calculating a similarity measure. This function computes the following measures (van Eck & Waltman, 2009): the association strength (also called proximity index), the inclusion index (also called Simpson's coefficient), the Jaccard's coefficient, and the Salton's cosine.

Let B denote a bibliographic coupling, co-citation, or a co-occurrence matrix as defined in Section 4.3.

The association strength is the ratio between the observed and expected strength under the assumption of probabilistic independence:

$$S_{A\_ij} = \frac{b_{ij}}{b_{ii}b_{jj}}$$

[S <- normalizeSimilarity(NetMatrix, type = "association")].

The inclusion index is an overlap metric that measures how much a set is included in another:

$$S_{I\_ij} = \frac{b_{ij}}{\min(b_{ii}, b_{jj})}$$

[S <- normalizeSimilarity(NetMatrix, type = "inclusion")].

The Jaccard's index (or Jaccard's similarity coefficient) is a relative measure of the intersection of two sets. It is calculated as the ratio between the intersection and the union of the two objects:

$$S_{J\_ij} = \frac{b_{ij}}{b_{ii}+b_{jj}-b_{ij}}$$

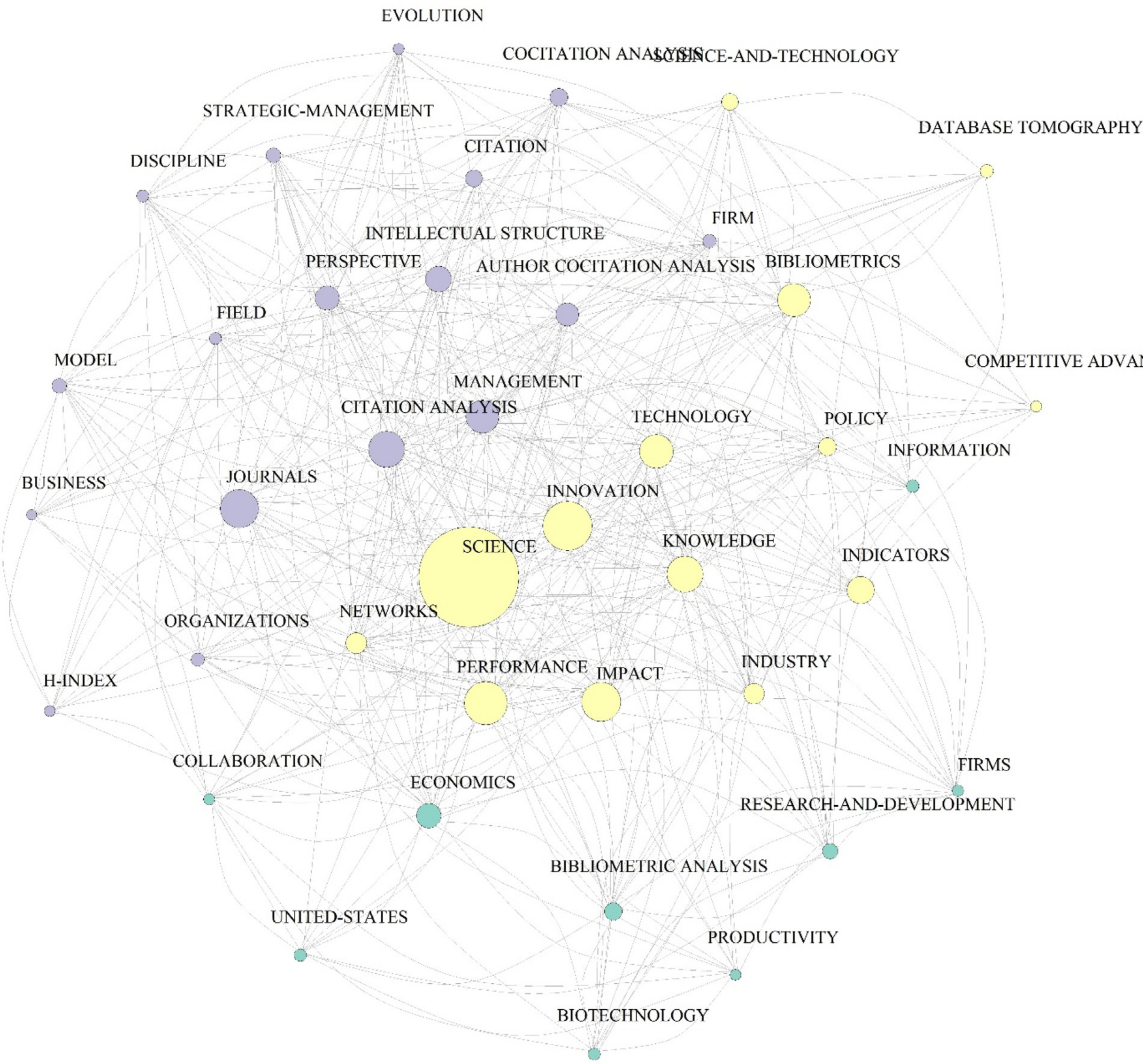

Fig. 4. Keyword Plus co-occurrence network (Kamada & Kawai layout).

[S <- normalizeSimilarity(NetMatrix, type = “jaccard”)].

The Salton’s index relates the intersection of the two objects to the geometric mean of the size of both:

$$S_{S\_ij} = \frac{b_{ij}}{\sqrt{b_{ii} b_{jj}}}$$

[S <- normalizeSimilarity(NetMatrix, type = “salton”)].

The square of Salton’s index is also called the equivalence index [S <- normalizeSimilarity(NetMatrix, type = “equivalence”)].

4.5. Network mapping

All bibliometric networks can be graphically visualized or modelled. The *networkPlot* function plots a network created by *biblioNetwork* using R routines or using *VOSviewer* software by Nees Jan van Eck and Ludo Waltman (Van Eck & Waltman, 2010; van Eck, Waltman, & Noyons, 2010; Waltman, Van Eck, & Noyons, 2010).

Fig. 4 displays a keyword plus co-occurrence network using the kamada-kawai layout (Kamada & Kawai, 1989). The network is drawn selecting the 40 vertices with highest degree [COC <- biblioNetwork(M, analysis = “co-occurrences”, network = “keywords”, sep = “;”), networkPlot(COC, n=40, size=TRUE, remove.multiple = T, Title=“Term co-occurrences”, type=“kamada”, labelsize=0.5)].

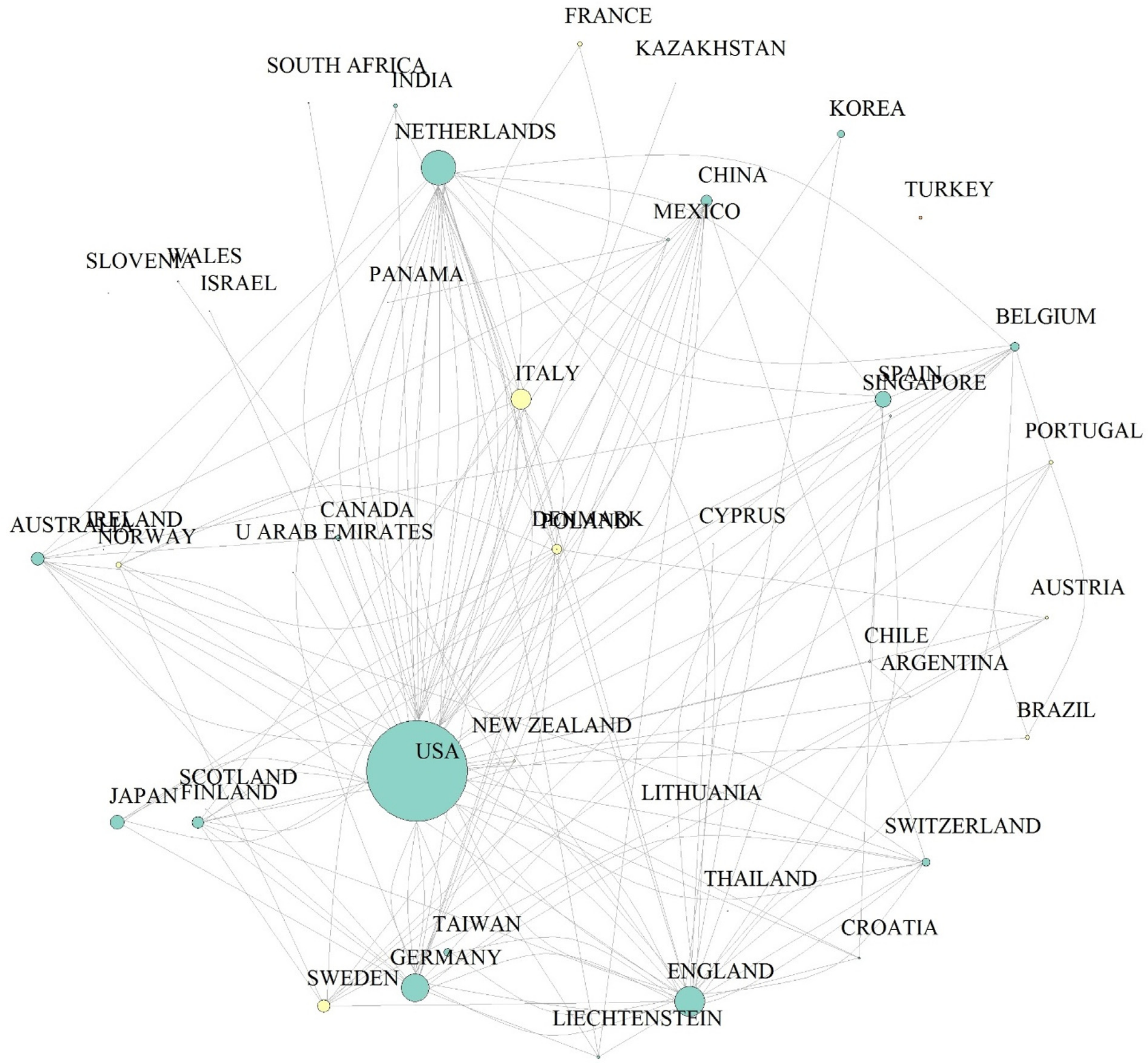

Fig. 5. Country collaboration network (Sphere layout).

Fig. 5 displays another example of a bibliographic network considering collaboration links between countries. In this case, we used sphere layout [M <– metaTagExtraction(M, Field = "AU_CO"); CC <- biblioNetwork(M, analysis = "collaboration", network = "countries", sep = ";"); networkPlot(CC, n=44, size=TRUE, remove.multiple = FALSE, Title="Country Collaboration", type="sphere")].

bibliometrix also performs historiographic analysis, as proposed by Garfield (2004) [histResults <- histNetwork(M, n = 20, sep = ";")]. The *histPlot* function plots a chronological citation network (called a *historiograph*, please see Fig. 6 and Table 13) that represents a chronological map of the most relevant citations resulting from a bibliographic collection [histPlot(histResults, size=FALSE)].

5. Conclusions

Science mapping is becoming an essential activity for scholars of all scientific disciplines. As the number of publications continues to expand at increasing rates and publications develop fragmentarily, the task of accumulating knowledge becomes more complicated. The determination of intellectual structure and the research-front of scientific domains are important not only for the research but also for the policy-making and practice.

Specialized software tools commonly perform only certain steps of science mapping analysis. Only a small number of these allow scholars to follow the complete workflow. *bibliometrix* is an open-source tool for executing a comprehensive

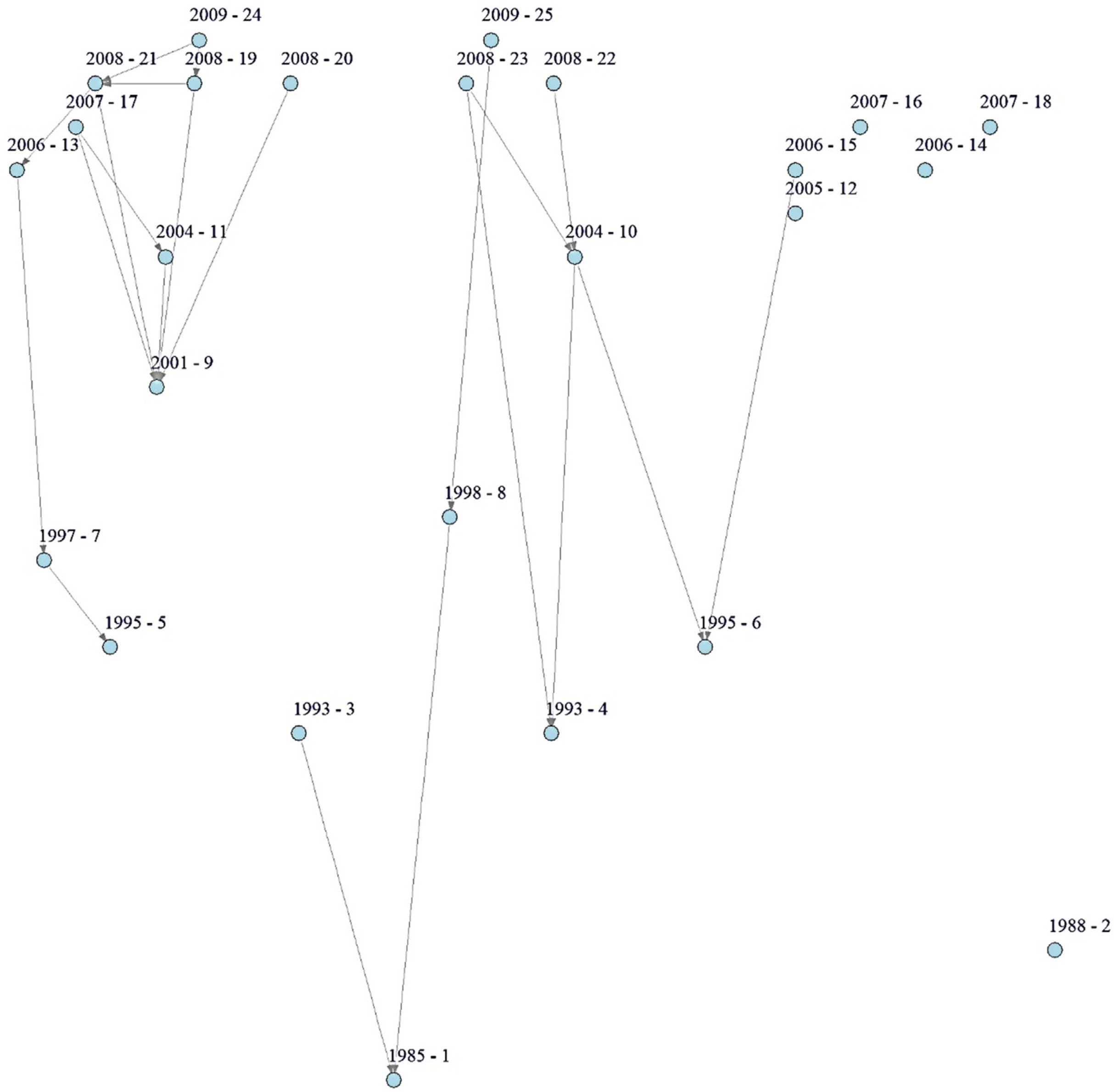

Fig. 6. Historiograph.

science mapping analysis of scientific literature. It was programmed in R to be flexible and facilitate integration with other statistical and graphical packages. Indeed, bibliometrics is a constantly changing science and *bibliometrix* has the flexibility to be quickly upgraded and integrated. Its development can address a large and active community of developers formed by prominent researchers. The advantages are direct. In fact, sources are published on GitHub permitting the creation of a shared development. Other advantages are indirect. In an environment composed of thousands of packages, *bibliometrix* can be a step in a larger workflow, exploiting other R solutions.

We are already working on new developments. They concern (i) the extension of compatibility with other bibliographic databases such as PubMed, (ii) the improvement of reference disambiguation by string metric-based algorithms, (iii) the introduction of direct citation (Klavans & Boyack, 2017) and tri-citation analysis (Marion, 2002; McCain, 2009), and (iv) the use of hybrid methods that combine bibliometric and semantic approaches (Glänzel & Thijs, 2012; Thijs, Schiebel, & Glänzel, 2013). The last-mentioned development includes term-burst detection through expectile smoothing (Schnabel & Eilers, 2009), thematic mapping and evolution (Cobo et al., 2011b), and latent semantic analysis (Dumais, 2004).

Table 13
Historiograph legend.

ID	*Reference*	*DOI*	*Local citations*	*Total citations*
1985–1	MOED HF, 1985, RES POLICY	10.1016/0048–7333(85)90012-5	14	232
1988–2	NARIN F, 1988, RES POLICY	10.1016/0048-7333(88)90039-X	9	44
1993–3	NEDERHOF AJ, 1993, RES POLICY	10.1016/0048–7333(93)90005-3	6	61
1993–4	HOFFMAN DL, 1993, J CONSUM RES	10.1086/209319	18	67
1995–5	PORTER AL, 1995, TECHNOL FORECAST SOC	10.1016/0040–1625(95)00022-3	8	89
1995–6	USDIKEN B, 1995, ORGAN STUD	10.1177/017084069501600306	11	80
1997–7	WATTS RJ, 1997, TECHNOL FORECAST SOC	10.1016/S0040-1625(97)00050-4	14	112
1998–8	RINIA EJ, 1998, RES POLICY	10.1016/S0048-7333(98)00026-2	10	113
2001–9	KOSTOFF RN, 2001, IEEE T ENG MANAGE	10.1109/17.922473	13	220
2004–10	RAMOS-RODRIGUEZ AR, 2004, STRATEGIC MANAGE J	10.1002/SMJ.397	32	190
2004–11	KOSTOFF RN, 2004, TECHNOL FORECAST SOC	10.1016/S0040-1625(03)00048-9	6	100
2005–12	KOSTOFF RN, 2005, TECHNOL FORECAST SOC	10.1016/J.TECHFORE.2005.02.001	6	14
2006–13	DAIM TU, 2006, TECHNOL FORECAST SOC	10.1016/J.TECHFORE.2006.04.004	18	240
2006–14	SCHILDT HA, 2006, ENTREP THEORY PRACT	10.1111/J.1540-6520.2006.00126.X	8	59
2006–15	PILKINGTON A, 2006, TECHNOVATION	10.1016/J.TECHNOVATION.2005.01.009	6	38
2007–16	KOSTOFF RN, 2007, TECHNOL FORECAST SOC	10.1016/J.TECHFORE.2007.02.007	6	26
2007–17	KOSTOFF RN, 2007, TECHNOL FORECAST SOC	10.1016/J.TECHFORE.2007.04.004	7	47
2007–18	BIEMANS W, 2007, J PROD INNOVAT MANAG	10.1111/J.1540-5885.2007.00245.X	6	20
2008–19	KAJIKAWA Y, 2008, TECHNOL FORECAST SOC	10.1016/J.TECHFORE.2008.04.007	7	44
2008–20	SHIBATA N, 2008, TECHNOVATION	10.1016/J.TECHNOVATION.2008.03.009	7	83
2008–21	KAJIKAWA Y, 2008, TECHNOL FORECAST SOC	10.1016/J.TECHFORE.2007.05.005	14	76
2008–22	NERUR SP, 2008, STRATEG MANAGE J	10.1002/SMJ.659	18	96
2008–23	CHARVET FF, 2008, J BUS LOGIST	<NA>	9	34
2009–24	KAJIKAWA Y, 2009, TECHNOL FORECAST SOC	10.1016/J.TECHFORE.2009.04.004	7	34
2009–25	ABRAMO G, 2009, RES POLICY	10.1016/J.RESPOL.2008.11.001	6	50

Author contributions

Massimo Aria, Corrado Cuccurullo: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

Acknowledgements

The authors would like to thank the editor and referees for their helpful comments. These have allowed us to significantly improve the quality of this paper.

References

Alavifard, S. (2015). *hindexcalculator: H-index calculator using data from a web of science (WoS) citation report. R package version 1.0.0*. https://CRAN.R-project.org/package=hindexcalculator

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179–255.

Bailón-Moreno, R., Jurado-Alameda, E., & Ruiz-Baños, R. (2006). The scientific network of surfactants: Structural analysis. *Journal of the American Society for Information Science and Technology*, *57*(7), 949–960.

Bar-Ilan, J. (2007). Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, *74*(2), 257–271.

Benzécri, J. P. (1982). *L'Analyse des Donnés. II. L'analyse des correspondances*. Paris: Dunod.

Briner, R. B., & Denyer, D. (2012). Systematic review and evidence synthesis as a practice and scholarship tool. In *Handbook of evidence-based management: Companies, classrooms and research*. pp. 112–129.

Broadus, R. (1987). Toward a definition of bibliometrics. *Scientometrics*, *12*(5–6), 373–379.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, *22*(2), 191–235. http://dx.doi.org/10.1177/053901883022002003

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the Association for Information Science and Technology*, *57*(3), 359–377.

Cobo, M. J., Lopez-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science Mapping Software Tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: a practical application to the fuzzy sets theory field. *Journal of Informetrics*, *5*(1), 146–166.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, *63*(8), 1609–1630.

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press.

Cuccurullo, C., Aria, M., & Sarto, F. (2016). Foundations and trends in performance management. A twenty-five years bibliometric analysis in business and public administration domains. *Scientometrics*, *108*(2), 595–611.

Diodato, V. (1994). *Dictionary of bibliometrics*. Binghamton, NY: Haworth Press.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, *38*, 189–230.

Gagolewski, M. (2011). Bibliometric impact assessment with R and the CITAN package. *Journal of Informetrics*, *5*(4), 678–692.

Gao, X., & Guan, J. (2009). Networks of scientific journals: An exploration of Chinese patent data. *Scientometrics*, *80*(1), 283–302.

Garfield, E. (2004). Historiographic mapping of knowledge domains literature. *Journal of Information Science*, *30*(2), 119–145.

Gifi, A. (1990). *Nonlinear multivariate analysis*. John Wiley & Sons Incorporated.

Glänzel, W., & Schubert, A. (2004). *Analysing scientific networks through co-authorship*. pp. 257–279. *Handbook of quantitative science and technology research (11)*.

Glänzel, W., & Thijs, B. (2012). Using core documents for detecting and labelling new emerging topics. *Scientometrics*, *91*(2), 399–416. http://dx.doi.org/10.1007/s11192-011-0591-7

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, *51*(1), 69–115.

Greenacre, M., & Blasius, J. (Eds.). (2006). *Multiple correspondence analysis and related methods*. CRC Press.

Guler, A. T., Waaijer, C. J., Mohammed, Y., & Palmblad, M. (2016). Automating bibliometric analyses using Taverna scientific workflows: A tutorial on integrating Web Services. *Journal of Informetrics*, *10*(3), 830–841.

Guler, A. T., Waaijer, C. J., & Palmblad, M. (2016). Scientific workflows for bibliometrics. *Scientometrics*, *107*(2), 385–398.

Harzing, A. W. (2007). *Publish or Perish*. [available from]. http://www.harzing.com/pop.htm

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 16569–16572.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, *31*(1), 7–15 [Elsevier].

Keirstead, J. (2015). *scholar: analyse citation data from Google Scholar. R package*.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *Journal of the Association for Information Science and Technology*, *14*(1), 10–25.

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984–998.

Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate descriptive statistical analysis (correspondence analysis and related techniques for large matrices)*. Chichester: Wiley.

Marion, L. (2002). A tri-citation analysis exploring the citation image of Kurt Lewin. *Proceedings of the American Society for Information Science and Technology*, *39*(1), 3–13.

Matloff, N. (2011). *The art of R programming: A tour of statistical software design*. No Starch Press.

McCain, K. W. (1991). Mapping economics through the journal literature: An experiment in journal cocitation analysis. *Journal of the American Society for Information Science*, *42*(4), 290.

McCain, K. W. (2009). Using tricitation to dissect the citation image: Conrad Hal Waddington and the rise of evolutionary developmental biology. *Journal of the American Society for Information Science and Technology*, *60*(7), 1301–1319. http://dx.doi.org/10.1002/asi

Persson, O., Danell, R., & Schneider, J. W. (2009). . pp. 9–24. *How to use Bibexcel for various types of bibliometric analysis. Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th Birthday* (5).

Peters, H., & Van Raan, A. (1991). Structuring scientific activities by co-author analysis: An expercise on a university faculty level. *Scientometrics*, *20*(1), 235–255.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation*, *25*, 348.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org

Rousseau, D. M. (Ed.). (2012). *The Oxford handbook of evidence-based management*. Oxford University Press.

Schnabel, S. K., & Eilers, P. H. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, *53*(12), 4168–4177.

Sci2 Team. (2009). *Science of Science (Sci2) Tool*. Indiana University and SciTech Strategies. https://sci2.cns.iu.edu

Skupin, A. (2009). Discrete and continuous conceptualizations of science: Implications for knowledge domain visualization. *Journal of Informetrics*, *3*(3), 233–245.

Small, H. G., & Koenig, M. E. (1977). Journal clustering using a bibliographic coupling method. *Information Processing & Management*, *13*(5), 277–288.

Small, H., & Upham, P. (2009). Citation structure of an emerging research area on the verge of application. *Scientometrics*, *79*(2), 365–375.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology*, *24*(4), 265–269.

Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics*, *68*(3), 595–610.

Thijs, B., Schiebel, E., & Glänzel, W. (2013). Do second-order similarities provide added-value in a hybrid approach? *Scientometrics*, *96*(3), 667–677. http://dx.doi.org/10.1007/s11192-012-0896-1

Uddin, A. (2016). *scientoText: Text & Scientometric Analytics. R package version 0.1*. [https://CRAN.R-project.org/package=scientoText version 0.1.4, http://github.com/jkeirstead/scholar].

Upham, S. P., & Small, H. (2010). Emerging research fronts in science and technology: patterns of new knowledge development. *Scientometrics*, *83*(1), 15–38.

Waltman, L., Van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629–635.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391.

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. http://dx.doi.org/10.1002/asi.4630320302

White, D., & McCain, K. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science*, *49*(4), 327–355.

Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326.

Yang, K., & Meho, L. I. (2006). Citation analysis: a comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for information science and technology*, *43*(1), 1–15.

Yang, S., Han, R., Wolfram, D., & Zhao, Y. (2016). Visualizing the intellectual structure of information science (2006–2015): Introducing author keyword coupling analysis. *Journal of Informetrics*, *10*(1), 132–150.

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science*, *59*(1998), 2070–2086. http://dx.doi.org/10.1002/asi

Zupic, I., & Čater, T. (2015). Bibliometric methods in management and organization. *Organizational Research Methods*, *18*(3), 429–472.

de Moya-Anegon, F., Vargas-Quesada, B., Chinchilla-Rodriguez, Z., Corera-Alvarez, E., Herrero-Solana, V., & Munoz-Fernández, F. J. (2005). Domain analysis and information retrieval through the construction of heliocentric maps based on ISI-JCR category cocitation. *Information Processing & Management*, *41*(6), 1520–1533.

van Eck, N. J., & Waltman, L. (2008). Generalizing the h-and g-indices. *Journal of Informetrics*, *2*(4), 263–271.

van Eck, N. J., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the Association for Information Science and Technology*, *60*(8), 1635–1651.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

van Eck, N. J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, *8*(4), 802–823.

van Eck, N. J., Waltman, L., & Noyons, C. M. (2010). A unified approach to mapping and clustering of bibliometric networks14. *Eleventh International Conference on Science and Technology Indicators* [p. 284].