TITLE PAGE

Title: **Survival analysis of author keywords: An application to the Library and Information Sciences area.**

Author (s): **Peset F[1], Garzón-Farinós F[2], González LM[3], García-Massó X[4], Ferrer-Sapena A[1], Toca-Herrera JL[5], Sánchez-Pérez EA[1*]**

1 Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.

2 Universidad Católica de Valencia San Vicente Mártir, Quevedo 2, 46001 Valencia, Spain.

3 Departamento de Educación Física y Deporte, Universitat de València. Gascó Oliag, 3. 46010 Valencia, Spain.

4 Departamento de Didáctica de la Expresión Musical, Plástica y Corporal, Universitat de València, Avd. de los Naranjos 4, 46022, Valencia, Spain.

5 Institute for Biophysics, Department of Nanobiotechnology, University of Natural Resources and Life Sciences, Vienna, Austria.

*Corresponding Author: easancpe@mat.upv.es, +34963877000, ext. 76637.

Declarations of interest: none.

ABSTRACT

Our purpose is to adapt a statistical method for the analysis of discrete numerical series to the keywords appearing in scientific papers of a given area. As an example, we apply our methodological approach to the study of the keywords in the Library and Information Sciences (LIS) area. Our objective is to detect the new author keywords that appear in a fixed knowledge area in the period of one year in order to quantify the probabilities of survival for 10 years as a function of the impact of the journals where they appeared. Many of the new keywords appearing in the LIS field are ephemeral. Actually, more than half are never used again. In general, the terms most commonly used in the LIS area come from other areas. The average survival time of these keywords is approximately 3 years, being slightly higher in the case of words that were published in journals classified in the second quartile of the area. We believe that measuring the appearance and disappearance of terms will allow to understand some relevant aspects of the evolution of a discipline, providing in this way a new bibliometric approach.

KEYWORDS

Keyword survival analysis, author keywords, keyword frequency, library and information science research

INTRODUCTION

Researchers recurrently review the literature that exists regarding their field of study as a part of their scientific methodology. This work is becoming increasingly complex, since the number of articles produced in a year exceeds the assimilation capacity of any research group. All the actors involved in scientific production have made an effort for adapting themselves to the distribution of digital information, changing their habits and their ways of working (Athukorala, Hoggan, Lehtiö, Ruotsalo, & Jacucci, 2013; Lancaster, 2003; Niu & Hemminger, 2012).

Despite the undoubted advances in the field of information sciences, we can affirm that the selection of scientific literature has become another problem for working groups. Simplifying the problems in the collection of information, we find two key points: on the one hand, the way in which information is stored and handled, and on the other hand, what methods exist to analyse it.

We will focus on the second point of the assimilation of information by the researcher who consumes it. Even assuming that the information can be retrieved correctly, it is still open to determine which is the best way to synthesize it. Knowing the trends in certain topics is a complex problem. To face it, researchers use different methodological approaches. It is common to find reviews, systematic reviews or meta-analyses in the scientific literature whose objective is to establish the limits of what is already known regarding a specific topic. On the other hand, bibliometric methods allow us to determine other parameters, such as the growth of articles and citations per year, the rankings of the most prolific authors or institutions, the geographical distribution of the authors, the collaboration patterns, or the frequency of subject descriptors, among many others (Naseer & Mahmood, 2009).

When an analyst studies a part or parts of the publications about a certain topic, he always prioritises one type of information over another, consciously or not. For example, when an expert analyses co-authorship, he assumes that the author of the article is the 'engine' behind scientific discoveries. In contrast, if the words of the article are analysed (e.g., co-words), the centre of the action is placed in the contents, as if they were independent of their creators. We are aware that one approach does not exclude the other; however, scientific advances can be understood without their creators, but not the other way around. With all of this in mind, we must focus the analysis on the contents to establish the trends in a certain field of knowledge.

As we have already mentioned, reading and reviewing the best articles may be a viable but not efficient solution, due to the massive amount of information. Moreover, in the knowledge fields with intense scientific production it may become a titanic task. Thanks to new approaches with text mining, a part of the process that previously fell to the researcher is automated (O'Mara-Eves, Thomas, McNaught, Miwa, & Ananiadou, 2015). With these text-mining techniques, hidden knowledge is discovered in large masses of text. Many analyses can be executed on the content of an article. Among them, those that apply to the author keywords (AK) that accompany a paper have been shown to be meaningful to understanding a discipline (Li, 2018; Onyancha, 2018; Xu et al., 2018; Zhang et al., 2016).

The AK telegraph what happens in the text they present, describing it. In 1970, Jones and Jackson offered a generic definition of the term: 'Keywords are a list of words or phrases that are provided by the author and signify the meaning or main ideas presented in the paper' (Jones & Jackson, 1970). Therefore, the AK are authentic indicators of the topics that appear in the articles. The authors choose the AKs voluntarily, without subjecting themselves to a controlled vocabulary in most disciplines. Therefore, we believe that through this concretising exercise, authors reflect their personal visions of their subjects (Kevork & Vrechopoulos, 2009). In this choice, we detect at least two types of AK. On the one hand, authors choose terms already used by others in their area because a work is fed by previous articles. Thus, the appearance of terms already known indicates the influence that previous discoveries have in their research. On the other hand, when researchers contribute a new topic, the authors adopt words from other areas or simply coin new terms.

The investigations of AK are numerous and present highly varied approaches. However, most mainly count terms and use co-occurrence networks (Aizawa & Kageura, 2003; Dotsika & Watkins, 2017; Névéol, Doğan, & Lu, 2010; Radhakrishnan, Erbis, Isaacs, & Kamarthi, 2017; Su & Lee, 2010; Uddin & Khan, 2016; Yang, Han, Wolfram, & Zhao, 2016). Only a few works focus on the dynamics of keywords over time (Gil-Leiva & Alonso-Arroyo, 2007; Han, Gui, & Xu, 2014; Lee, 2016; Liu, Tian, Kong, Lee, & Xia, 2019; Mela, Roos, & Deng, 2013). Our work is focused in the latter direction. We believe that measuring the appearance and disappearance of terms will allow us to understand the evolution of a discipline through the literature accumulated on a given topic.

Our proposal seeks to take a step forward and apply survival analysis techniques to the AK of the Library and Information Science (LIS) knowledge field as an example of application of our methodology. As far as we know, there are no previous studies that have proposed this type of analysis applied to AK. Our purpose is to adapt a statistical method usually focused on discrete numerical series to the keywords of our area of knowledge. Consequently, we have set as our main objective to detect the new AKs that appear in the Library and Information Sciences knowledge area for the period of one year to quantify their probabilities of survival over a period of 10 years. Also, we have analysed the survival probabilities of AK depending on the impact of the journals in which they appeared. Secondarily, the origin of the new keywords that appeared in the LIS area will be studied.

METHODS

As we said in the Introduction, we will show how to apply our technique in a concrete scientific area. LIS is a professional an academic domain drawing on many kinds of knowledge. LIS studies the problems related to scientific information from different points of view (Hjørland, 2000; Milojević, Sugimoto, Yan, & Ding, 2011). We have chosen the LIS area for several reasons–for example, it is our usual field of study and we are familiar with its characteristics–, but mainly because it is a very active research area in which new technical words appear every day. The coherence in the use of the AK by LIS researchers is another reason. Finally, the interdisciplinary nature of the field and the influence of other areas is also reflected in the broader vocabulary used in LIS. So, our method can be applied to any other scientific field.

Selection of the author keywords to analyse

All the journals appearing in the Scimago Journal Rank (SJR*)* in the LIS category were selected (‘Scimago Journal & Country Rank’, n.d.). SJR classifies the journals appearing in the Scopus® database by subject area and subject category. It also includes information on various quality indicators and establishes a ranking according to the impact of each of the indexed journals.

All the AKs published in LIS subject-category journals composed the universe of our study. As an inclusion criterion, it was established that the keywords must appear in any of the journals indexed in the SJR in 2014 in this category. It was also necessary that the journal had appeared in the SJR since at least 2004. If the journal only had coverage since 2005, then only the keywords that were published in the coverage period were considered.

Search strategy and keyword extraction procedure

To establish the search strategy, the ‘Journal rankings’ tool of the Scimago Journal Rank website was used. From the categories subject filter, LIS and 2014 were selected to download the Excel document with the relevant journals, using the ‘Download data’ option. The same procedure was carried out for the year 2004. From the different fields declared in the files, the fields ‘ISSN’ and ‘SJR Quartile’ were extracted for the search and subsequent analysis. ‘SJR Quartile’ indicates the quartile to which a given journal belongs according to its impact. To do this, SCImago Journal & Country Rank scores all journals based on their citations and assigns them a value. Those journals with the highest scores are included in quartile 1 and those with the lowest impact in quartile 4. Because many of the journals have had more than one ISSN throughout their history, the extracted information was completed by separately performing individual searches for each of the journals. Finally, 295 ISSNs were combined in a single search strategy [see supplementary material S1 (Peset et al., 2018a)].

The search strategy was introduced in the Scopus® database, establishing as a time limit that the document had been published before 2015. The aforementioned search yielded a result of 120,489 documents. The records were downloaded in batches of 2000 documents in RIS format, selecting the following fields: (i) Author keywords, (ii) Year, (iii) Source title and (iv) ISSN. See supplementary material S2 (Peset et al., 2018b).

For the extraction and storage of the keywords in a single document, the software BibExcel (version 2011-02-03, Olle Persson, Umeå University, Umeå, SWE) was used. Because many journals do not declare the AK field or have just recently begun declaring it, the resulting file contained fewer references than the original search, namely, 110,731 documents (46,905 articles contained AK), with a total of 248,693 AK from 161 journals.

Search of new keywords appearing in 2004 and frequency of later appearance

As the first step of the analysis, 2004 was chosen as the year of study. We are aware that selection of the year is a key point that conditions the results. In the discussion section, the different criteria that can be used are indicated. A priori, because there are no prior works using this type of methodology, an intermediate year of the historical series was chosen, taking into account that there would be a large number of previous years and a sufficient number of subsequent years to carry out a valid analysis.

To locate the new keywords that emerged during 2004 and to track their frequency of emergence during the subsequent years, custom-written software routines (MATLAB R2013a, MathWorks Inc., Natick, MA, USA) were used. In this analysis, uppercase was not differentiated from lowercase.

Prior to the analysis, the records were divided into three periods based on the year in which the AK was published: historic period [1889-2003], emergence period [2004] and survival period [2005-2014]. Each AK of the year 2004 was searched in the historic period. The keywords that did not appear prior to 2004 formed our category of 'new keywords' in LIS.

Subsequently, each new word in the survival period was searched every year. The result for each word is a vector of 10 columns, one per year, in which 1 indicated the appearance of the term and 0 the non-appearance. In addition, the number of times the word appeared every year was counted (Figure 1). Frequency tables were calculated for the entire file generated.

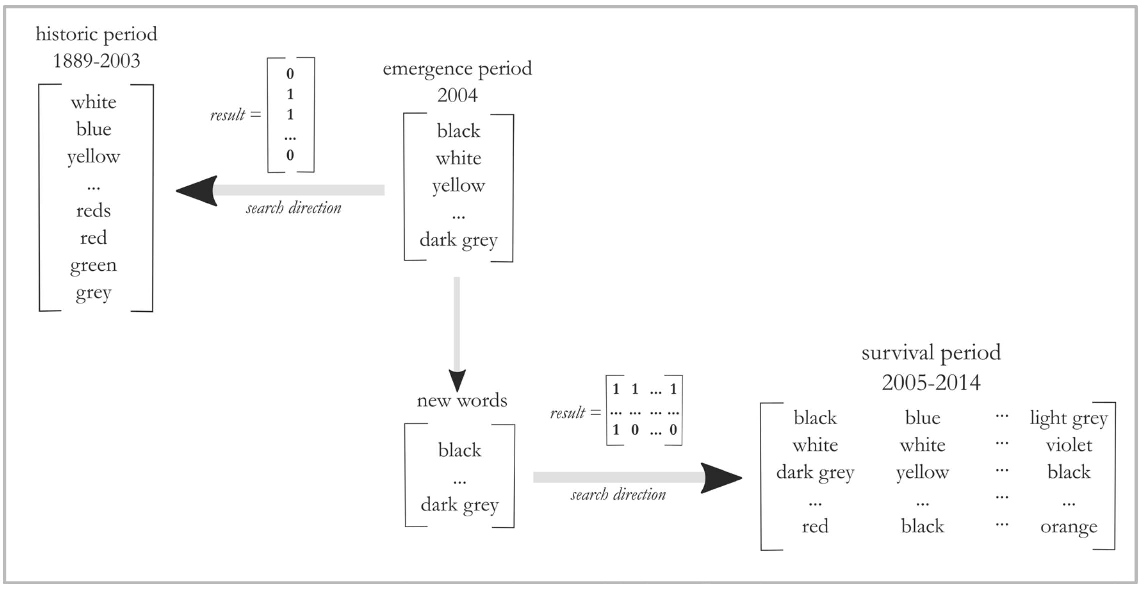

Figure 1. Map of new keywords and calculation of survival time. *The words chosen for the example do not correspond to real records; they only show the chosen methodology. In the centre of the figure appears the emergence period, with all the words that were published during 2004. With these words (4 in the example), a retrospective search is performed in the historic period (left section of the panel). As a result of this search, those words that do not appear in the historic period are selected (in the example, 2 words, 'black' and 'dark grey'). Finally, there is a follow-up during the survival period (right section of the panel) noting in each year if the word appears or does not appear (in the example, the word 'black' appears every year, while the word 'dark grey' only appears the first year). Additionally, the number of times per year the word appeared in the survival period was recorded (in the example, the word 'black' appears with a frequency of 1 in the three years and the word 'dark grey' with a frequency of 1 in the first year).*

Finally, in order to be able to compare the performance of 2004 with previous years, the same process was carried out for the years 1990 to 2003. Thus, survival was computed starting from 14 different years over the following 10 years.

Analysis of the new keywords from the emergence period through the survival period

The present study aims to determine the evolution of the new AK that appear in a given period. These 'new' keywords can be the result of migration from other areas or may be new words in any area covered by the database. Likewise, new keywords are those that present some variation, except in upper and lower case. Singular and plural forms of a term are considered as distinct words; this methodological choice will be justified later. Only AK in English are taken into account.

Furthermore, those that have been written with some orthographic or typographical incorrectness will be considered to be in the 'new' category. Regarding this last aspect, it should be noted that orthographic or typographic anomalies are not frequent but are possible. Typically, the incorrect word disappears over the years and becomes an anomaly without impact. If it survives, we understand that the scientific community has accepted a new use. This would mean a change in the way researchers write. Therefore, our analysis has a collateral utility for language researchers, who could establish new writing patterns.

With the 'new' keywords detected, the first analysis was performed, which consisted in calculating the probability that each of the new keywords had of surviving or disappearing. To do this, an analysis of survival through Kaplan-Meier curves was proposed. This type of analysis estimates the time that passes until a certain event occurs. The survival analysis can be applied to all those events that occur over time and that have been previously defined. In fact, this type of analysis has been applied to numerous areas of knowledge as diverse as medicine, economics, production engineering or social sciences (Singer & Willett, 1993).

The proposed procedure is an adaptation of the standard survival analysis to the case of the AKs. The Kaplan-Meier method is a statistical technique for modeling a survival function. Given a set of elements that change their state from 1 to 0 after a certain time t, the Kaplan-Meier formula is used to estimate the fraction of elements that retains the state equal to 1 after t.

Consider a sample of n individuals. At a time t_i, n_i individuals of the original population survives (i.e., there are n_i elements that maintain a state value equal to 1), and d_i changes of state are observed. The risk estimator at that time t_i is given by the fraction $\frac{d_i}{n_i}$ so the survival probability is $1-\frac{d_i}{n_i}$. This allows defining the known Kaplan-Meier estimator of the survival function used in this work, which provides the probability that an individual will survive a longer time than t. For a time t, the estimator is given by the formula

$$S(t) = \prod_{i:\, t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where t_i represents each of the instants that are considered in the model, d_i the number of state changes that occurred at the time t_i and n_i the elements whose state remains equal to their original value 1. This type of function is used for survival analysis. The standard application example is to model the survival rate of a population affected by a given disease. In our study, the population is the total set of AKs, and the periods of time that are considered are given by the years of publication. In this case, the problem is to define the event that produces the change of state. Since words can "be reborn", we have to define when we consider that a word has passed to state 0 using some arbitrary criterion, for example, that does not appear in a given time t, and does not appear again in all the analyzed period.

Before we can conduct this analysis, it is necessary to establish three fundamental aspects: the time of observation, the moment in which the event of interest occurs and when a subject is censored. In our analysis, the research subjects are the AKs, and they will be followed up over the survival period. In other words, we will test whether, after 10 years, these keywords appear in at least one article published in the LIS category.

Regarding the definition of the event, it is necessary to consider that a word can appear and disappear discontinuously over a period of 10 years. It could be argued that the data are interval-censored (Box-Steffensmeier et al., 2015). Interval-censored failure time data arise when the time of occurrence of the event of interest is known only to lie in an interval (Sun, 1997). However, we believe that if there are subsequent documents in which the word appears, the AK is alive, as researchers use it while preparing their articles. As a general criterion, it was established that a word 'died' in the last year in which there were records of it during the ten year period after 2004. As the reader can see, this criterion is somewhat arbitrary, as it may happen that the word reappears the year after completion of the study. This is a requirement of the Kaplan-Meier technique, and we have to assume that the final event must be set somehow. Our decision is consistent if we consider the following "time-dependent" definition of the final event (the death of a word): in 2014 −the last date of our analysis− a word is dead if it has not been used in the previous year. Measuring the number of years that the word has not appeared makes it possible to calculate the necessary additional "survival time". Thus, when the event was detected, it could not be repeated, and the analysis was terminated for this word (Singer & Willett, 1993). The last necessary step is to establish which keywords were censored: in our work, only those keywords that appeared throughout all the observation period were considered censored (right-censoring).

Once the data matrix was prepared, the analysis was performed through the SPSS 20 software (IBM, Armonk, New York, USA). The Kaplan-Meier curves were calculated for the total set of new keywords and for four subgroups, created according to the quartile of 2004 in which the journal published the AK. In cases where the keyword was published in two or more journals, the median of the quartile it belonged to was calculated, rounding the decimal places down. To determine if there were differences between the survival data of the 4 groups, a Wilcoxon (Gehan-Breslow) test of the equality of survivor functions was used. Finally, frequency and percentage tables were created to describe the new keywords discovered.

RESULTS

General descriptions of the retrieved records

Our database contains a total of 248,693 keywords included by authors in 46,905 articles over a period of 69 years. However, the use of keywords became widespread only in 1980. The ratio of keywords per article was 5.3.

If we discard duplicate keywords, the total is reduced to 102,349 keywords. As expected, the total of AK has grown exponentially (Figure 2) over the years. This is due to the increase in scientific production and the popularization of the use of the AK field as a form of indexing.

Seven keywords appeared more than 1,000 times throughout all the years tested: 1. 'LIBRARIES' (2,413); 2. 'ACADEMIC LIBRARIES' (2,232); 3. 'INTERNET' (2,029), 4. 'KNOWLEDGE MANAGEMENT' (1,449); 5. 'DIGITAL LIBRARIES' (1,341); 6. 'INFORMATION RETRIEVAL' (1,303); and 7. 'INFORMATION LITERACY' (1,249).

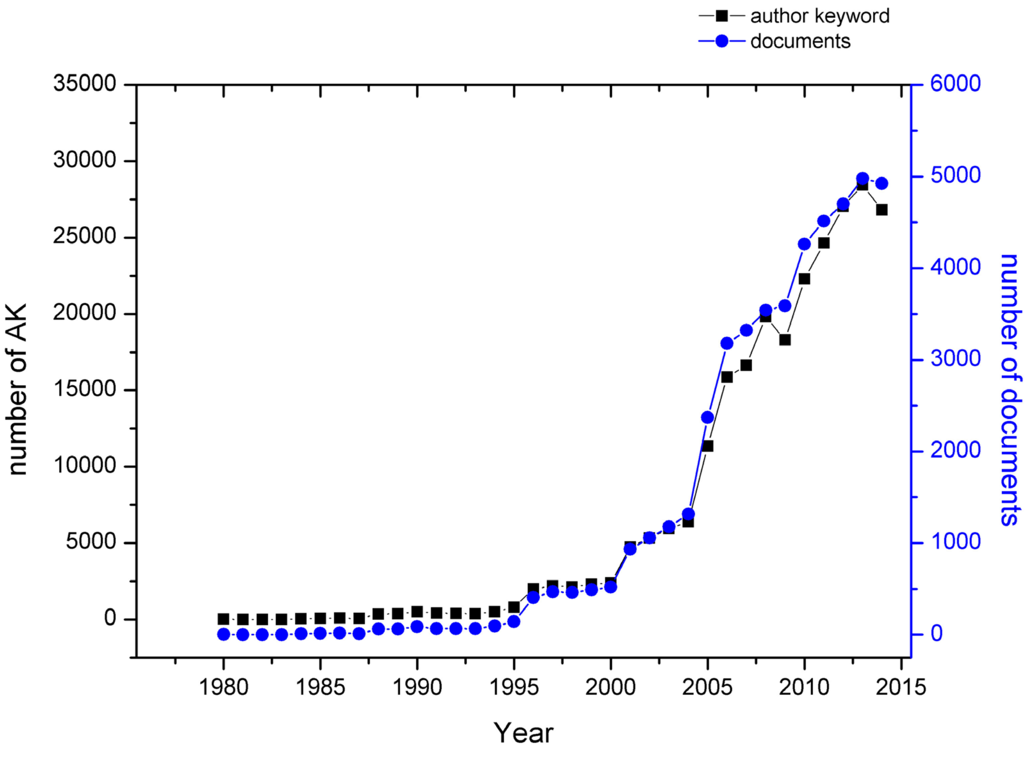

Figure 2. Frequency of appearance of keywords and documents in the LIS area between 1980 and 2014.

New keywords appearing in 2004 and their behaviour in subsequent years

During 2004, 6,372 keywords—without duplicates, 4,343—were published. Among them, we detected 2,810 new keywords. After eliminating the duplicates, 2,655 keywords were finally used in the subsequent analyses [see supplementary material S3 (Peset et al., 2018c)]. The three journals that published the greatest number of new keywords were as follows, in this order: *IEEE Transactions on Information Theory* (436), *Journal of Information and Computational Science* (339) and *Journal of Information Science and Engineering* (237). A total of 43 journals in the LIS area published at least a new keyword.

The word that was repeated the most in the survival period was 'GOOGLE' (266). The importance of these AK during the survival period varies, of which 'SENSOR NETWORK' and 'ARCHIVES MANAGEMENT' have lost the most influence during recent years (Figure 3).

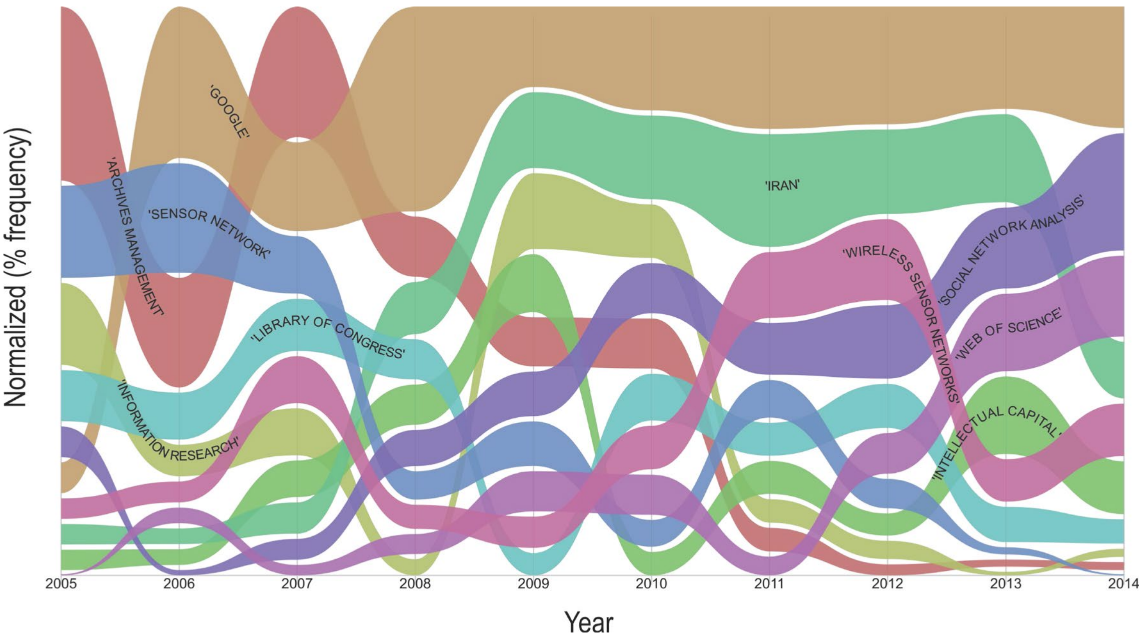

Figure 3. Bump chart of keyword ranking throughout the survival period. *The figure has been created considering only the 10 most frequent words that appeared as new keywords in 2004 and indicates the changes in their ranking throughout the period.*

Survival analysis

Figure 4 shows the survival curves of AKs over time after their appearance. The analysis was carried out with all the keywords involved in the process (panel a; green line) and revealed that during the first year, 65.3% of the keywords disappeared and were not used again to the end of the period analysed. From that moment on, a gentle decrease is observed that is accentuated slightly towards the end. Only 11.5% of the keywords reached the end of the period without any event being observed. In other words, only 306 keywords considered new, out of a total of 2,655, were used during the 10 years after their appearance. The average time (95% CI) of survival for the series was 3.33 (3.20 to 3.46) years. This average survival time is similar to that of the previous years tested 1990-2003 (panel b; red line). Survival during these previous years was 3.74 years (3.63 to 3.86). No statistically significant differences were found between the two periods.

The analysis was then carried out by classifying the file according to the quartile to which the journal publishing the keyword belonged, revealing that the keywords that debuted in the second quartile show a longer average survival time than those from other quartiles (Panel b). The worst prognosis is observed in the fourth quartile. Table 1 shows the estimated average survival times for each of the quartiles. Long-rank test and Wilcoxon (Gehan-Breslow) test for the equality of survivor functions both show P values below 0.001, confirming that there are significant differences between the curves analysed.

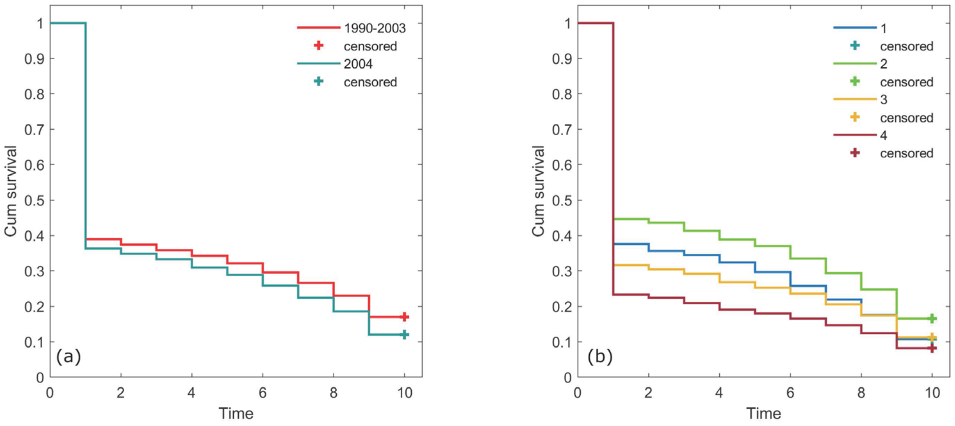

Figure 4. Kaplan-Meier curves for new AK appeared in 2004 and before. *The panel on the left (a) shows the survival curves of the data set. The red line shows the 10-year survival curves of the author keywords published in the period 1990-2003. The green line corresponds to the survival curves of the new words that appeared in 2004 over the following 10 years. The panel on the right (b) shows survival curves based on the quartile of the journal in which the keyword was published (1=first quartile; 2=second quartile; 3=third quartile and 4=fourth quartile).*

Table 1. Average survival time for AKs based on the quartile in which they were published

SJR quartile	Mean (Estimate)	SE	Lower Bound	Upper Bound
1	3.37	0.11	3.14	3.60
2	4.20	0.17	3.86	4.53
3	3.17	0.12	2.94	3.41
4	2.55	0.14	2.26	2.84

SE= standard error of the mean

DISCUSSION

About the methodology used

In designing our study, one of the first important decisions was the choice of the area to be analysed. The choice of the LIS area as a test ground for our methodology responds to double criteria: utilitarian and methodological. On the one hand, the LIS area is our usual field of study, and consequently, our 'know how' facilitates the immediate interpretation of the results. On the other hand, the authors of articles published in the LIS area are usually aware of the importance of the AK included in their works, and thus, the keywords are usually better conceived and have a greater homogeneity in their use. In addition, our article provides both the raw data collected and the data processed, so it can be reused by colleagues in our area who want to test other hypotheses. Of course, it is necessary to bear in mind that our methodology is applicable to most areas. In fact, more than one field of knowledge could even be combined and compared at the same time. In this way, we can affirm that the novelty of our work is linked to the type of statistical analysis used.

A second aspect to take into account is the choice of the year under study (i.e., 2004). In our case, the year selected does not respond to any qualitative criteria (e.g., a notorious event). We are aware that a qualitative or quantitative criterion (Dehdarirad, Villarroya, & Barrios, 2014) could add richness to the analysis. In our case, we have avoided choosing a specific date to avoid possible contamination of the results. Along these lines, the year 2004 was chosen because it was the most current in the historical series analysed, which allowed us to have a survival period of at least 10 years. Our first intention at the time of undertaking the analysis was to establish a methodology that can be used in future work. It would be interesting if other researchers reproduced our methodology by choosing another year or years as the period of emergence based on their interests or hypotheses. For those readers interested in knowing how different time windows would affect our analyses, other options have been added in the supplementary material S4 (Peset et al., 2018d) that cover different time periods (i.e. 5, 10, 15 and 20 years).

A third noteworthy aspect of our methodology is the use of survival analysis in a series of data that appear and disappear at intervals. As already indicated, a word not appearing in a given year did not necessarily indicate its death. Only when it did not appear in all subsequent years was it marked as missing. This approach is not completely new in the field of bibliometrics. Santos and Irizo (2002, 2005) used a criterion similar to ours, although in their case, the analysis was performed on the citations received by an article. Like us, these two authors assumed in their work that an article could be cited at intervals. Obviously, this criterion may be to the detriment of the spirit with which this type of analysis was created. However, we believe that the results obtained are valid and, more importantly, that they are useful for establishing behavioural models over time. In fact, our model has been tested with data from the year after the end of the analysis (year 2015) and we have been able to verify that only 3% of the words that were supposed 'to have died' appear again. However, future studies may use our supplementary material to search for words that have been considered dead and have subsequently become highly relevant (i.e. "Sleeping Beauty").

About retrieved records

Regarding the general descriptive results of our study, the number of manuscripts retrieved in our search is quite high. We obtained approximately 110,000 articles, a figure slightly higher than that presented by Figuerola et al. (2017). These authors retrieved fewer records (≈92,000) in a search similar to ours but was restricted to the years 1978-2014. This discrepancy is largely explained by the database used (i.e., LISA vs. SCOPUS) and by the smaller number of years included in that study because our study tracked all the indexed production since 1889.

As regards the number of AKs obtained, our word ratio based on the number of manuscripts is slightly higher than that obtained by other authors (Hu, Hu, Deng, & Liu, 2013). We have observed that, on average, each manuscript is accompanied by 5 keywords. It is necessary to keep in mind that our raw data include very old manuscripts indexed in which AK were not declared systematically. Obviously, over the years this trend has reversed, and although the number of manuscripts has increased, the use of keywords has increased even faster.

The general calculation of keywords highlights the use of five keywords over the years. The term 'LIBRARIES' is the most frequently used keyword, which could scarcely be otherwise. This is logical because the term encompasses both the professional phenomenon under study in the LIS area and the academic discipline that gives the area its name. The rest of the words described have already been observed in AK studies of other areas (Aharony, 2012; Hu et al., 2013), with most of them being area-specific terms. Only the appearance of the term 'INTERNET' is an exception to this pattern (Onyancha, 2018). Obviously, we are talking about a word that has had a universal impact and that has changed the behaviour of people in practically all areas (Colley & Maltby, 2008), which is more than sufficient reason to explain its importance in the LIS field.

New words appearing in 2004 and their behaviour in subsequent years

In 2004, we found 4,343 words in the published articles. Of these, 2,655 were new. Although the main objective of our work is to keep track of these words throughout the subsequent years, we want to dwell on some interesting issues that have arisen. The word 'GOOGLE' bursts in with force in the area. Being the most popular search engine among Internet users (Jan Brophy & David Bawden, 2005), its adoption by the LIS area does not require any additional explanation.

However, the second most used word, 'IRAN', does not have such an obvious explanation. Before venturing an explanation, we emphasise that considering all the years a word is recorded constitutes a residual, minority term. Two institutions in that country have the most works published with this word: Ferdowsi University of Mashhad and Islamic Azad University. According to Walters and Wilder (2016), Iran, as a country, is located at the modest thirteenth position in the LIS area ranking but has 5 active research groups that are among the 50 best such groups in the world. In conclusion, the most plausible explanation is that one or several of these groups are very active and use the word 'IRAN' as a way to familiarise the world with the local reality of their country.

As for the rest of the most common keywords that appeared in 2004, six of them come from other areas of knowledge. In other words, part of the novelty detected comes from journals dedicated to other areas of study. Our methodology does not allow us to affirm whether the same thing happens with all the new words detected. However, the influence of other areas on the LIS area is in line with the data collected by Tan in (2004). In his work based on an analysis of citations, he showed that the discipline that contributed most citations to the LIS area was 'Computer sciences', followed by 'Communication', 'Education' and 'Management sciences'. In our case, of the 10 most productive keywords, we have been able to confirm that 'Computer sciences' is the area with the most influence (Walters & Wilder, 2016). This is to be expected because the predominance of the 'digital phenomenon' has reached all fields of knowledge, including the LIS area.

Nevertheless, from a mathematical point of view there are other options to establish how keywords migrate from one area to another. In the work of Losee (1995) an equation is proposed that assumes that term growth may be modelled by the diffusion of the concept through a population (in our case keywords). This mathematical model can quantify how the importance of terms grows or decreases in different disciplines and can provide additional information to the approach we have taken in this article.

The AKs 'GOOGLE', 'SOCIAL NETWORK ANALYSIS' and 'WEB OF SCIENCE' have improved their position over the 10 years of the survival period. All three have been used throughout the period, and their influence continues to increase. In addition to marking an upward trend among the group of words that appeared in 2004, they are also among the most frequently used in all years (Hu et al., 2013). According to the classification of Zins (2007), two of these keywords belong to the category 'Data Organization and Retrieval' and the other to 'Information/Learning Society'. At the opposite pole are 'SENSOR NETWORK' and 'ARCHIVES MANAGEMENT', which could be included in the category of 'Information Technology' and 'Data Organization and Retrieval', respectively. These data agree with the main idea expressed by Aharony (2012). In view of the above, it seems that the trends do not relate to the large thematic categories because they are proportionally represented over the years. However, according to the dynamics of the words of our study, it seems that the large areas of 'World Wide Web', 'Libraries' and 'Education' research occupy a predominant position in the LIS area (González-Alcaide, Castelló-Cogollos, Navarro-Molina, Aleixandre-Benavent, & Valderrama-Zurián, 2008). Of course this variety of terms coming from different areas confirms the interdisciplinary nature of the LIS area.

Survival of author keywords

Half of the AK fail to survive beyond the first year. According to our survival analysis, the average time it takes for a new keyword to disappear is approximately 3 years. Only 11% of the keywords were used throughout the 'survival' period. As far as we know, no previous studies have applied a survival analysis to AKs. The works of Santos and Irizo (2002, 2005) employ a model of analysis closer to ours, using the citations received by the articles. Obviously, the behaviour of citations and keywords does not have to be similar; however, we have found some similarities. Although the results section has simplified the analyses carried out to improve reading fluency, like Santos and Irizo, we have tested our empirical model with different theoretical models. As with their findings, the distribution that best fits here is the Weibull distribution (k = 0.8, SE = 0.01, where k is the shape parameter), which indicates that the failure rate is constant over time. In other words, despite the sharp decline in the first year, our data indicates that keywords age steadily. As we said before, the notion of "death" of a word must be adapted to the context: a word is considered dead if it does not appear in any document in the last year of the analysis. This methodological option does not fit the original requirements of the Kaplan-Meier curves, but it allows a conceptual adaptation: the attention is limited to the period being analysed, since it is assumed that a word that disappears in it dies. If it reappears a new study, it should be considered as a new word in the context of a new analysis.

Finally, we have worked with the hypothesis that the keywords that appear in certain journals with varying impact may have different probabilities of survival. Testing this hypothesis would reveal our methodology's true potential. Once the T = survival time values and the δ = (0, 1) variable for failure/censorship have been obtained, all the desired parameters can be added to the survival analysis (e.g., words that come from two or more different areas, from different journals, simple vs. compound words, etc.). As an example, in our work, we have divided the keywords into four subgroups according to the quartile in which the publishing journal was ranked. Surprisingly, we have found differences, but contrary to what was expected, it was not the words included in Q1 that obtained the best results. Apparently, the terms that appeared in the second quartile had a longer survival time than those in the other quartiles ($p < 0.001$). These results cannot be compared with similar studies, nor does the scientific literature give us clues about the reasons behind this discovery. Therefore, we will try here to offer an argument that can be useful for future works.

It is common for researchers, when seeking information systematically, to place more importance on journals with greater impact. This tendency is due to the pressure usually put on researchers by the authorities and agencies in charge of evaluating their merits (Michael Hall, 2011; Nosek, Spies, & Motyl, 2012). In this way, there is a vicious cycle, with no easy solution, in which the agencies only evaluate high-impact publications positively, so the researchers try to publish their work in these journals and, consequently, unconsciously grant them more value. Finally, closing the circle, the agencies, seeing the response, see their criteria reinforced by the attitudes of the researchers. The journals are also part of this 'game' and are not exempt from pressure. It is possible that these Q1 journals tend not to accept breakthrough ideas, instead seeking authors and very secure works within existing lines of work that help to preserve their status. According to Wang et al. (2014), the most innovative ideas tend to come from exploratory studies, but this type of methodological design is rare in the best-ranked journals. Perhaps the journals of the second quartile are freer to accept approaches outside the established circles. However, this explanation can only be tentative because our experimental data do not allow us to conclude this. We believe that future work should address this problem by looking for other factors or co-variables that explain the phenomenon.

Before presenting the limitations of our study, it is necessary to highlight the novel aspects of our analysis compared to other similar approaches published to date. As already commented in the introduction, there are many works that have taken the AK as a subject of analysis. However, in the last years there have appeared different studies that try to establish the patterns of the AK over time. These studies use methods mainly based on the frequency of the occurrence of terms over the years, Before presenting the limitations of our study, it is necessary to highlight the novel aspects of our analysis compared to other similar approaches published to date. As already commented in the introduction, there are many works that have taken the AK as a subject of employing in some cases sophisticated methods of normalization, bibliometric indicators and linear regression models (Chang, Huang, & Lin, 2015; Chen & Xiao, 2016; Faust, 2018; M. Wang & Chai, 2018; Xu et al., 2018). Like our work, some authors (Cheng, Huang, Yu, & Wu, 2018; Khan & Wood, 2015, 2015) have focused their analyses on the time of appearance and disappearance of AK. Our article offers a similar analysis, but providing the calculation of the survival time of the AK. This fact, from our point of view, improves the ability to quantify and predict the behavior over the years of the AK and therefore, it shows the character and the evolution of LIS as a field of knowledge. To show a complete view of the field is possible through quantitative or qualitative studies (Tuomaala, Järvelin, & Vakkari, 2014). Our method allows us to offer a general view of the field and show findings that match with qualitative studies about LIS (Borgman, 2007; Hjørland, 2017; McClure & Bishop, 1989; Rayward, 2005; Vakkari, 1994). Like us, other authors studied the ephemeral, the constant evolution and the interdisciplinarity of the LIS topics with similar results using other methodologies (Halevi & Moed, 2013; Hjørland & Albrechtsen, 1995; Rayward, 1985). As an example, Buckland (2012), reviewing the obsolescence in assigned subject descriptors, affirmed that "linguistic expressions [as AKs] are necessarily culturally grounded, and, for that reason, in conflict with the need to have stable. Being pragmatic, our machine-based quantitative approach provides with few resources clues that experts could develop with human-based qualitative approaches.

Limitations

The main limitations of our study are linked to the methodology used for the selection of new words. Our results are restricted to AKs. In this way, when we say that an AK is totally new in the LIS area, it is not considered that the word may have appeared in other documents located within fields such as title, summary, full text, etc. Although this is the main limitation of our study, we believe that when authors introduce an AK in an article, they are marking the moment when the word acquires value for them because they are deciding that the term synthesizes a large part of their work.

Another criterion that may have limited our results is the similarity between words. In our analysis, the terms that have similar writing have been considered as different terms. For example, in our study, the AK 'SENSOR NETWORKS' and 'SENSOR NETWORK' (Levenshtein distance 0.93) have been considered as two different words. This procedure differs from most of the studies consulted (Ding, Chowdhury, & Foo, 2001; Leung, Sun, & Bai, 2017; Z.-Y. Wang, Li, Li, & Li, 2012). Any observer can intuit that, conceptually, both terms are the same, but in our study, we consciously decided not to do so. From our point of view, the authors are autonomous when choosing how to write their keywords. Moreover, when they introduce a term, for example, plural or singular, in fact they are subconsciously following patterns that they have already seen or read in other places (He, 1999). Consequently, these slight variations may be due to the sources they consult or the cultural environment (Buckland, 2012), and from our point of view, they should not be corrected by the authors. Despite the methodological option that we have decided on in this paper, future studies should test other options that unify the singular and plural terms using or adapting existing algorithms.

For us, the subjective criterion of homogenization of terms, which is applied profusely in other works, can cause the observers (i.e., the authors of the study) to influence their uses and customs in the studied phenomenon. We also do not alter the terms with a defective presentation (e.g., spelling). We believe that these errors are rectified naturally by the passing of the years because usually, the defective word tends to disappear. In the event that it survives, the students of a certain topic should analyse the reasons for and consequences of this variation.

In future work, those experts who do not agree with these arguments can homogenize and correct this limitation. To do so, we recommend that part of the process be automated (e.g., perform

Levenshtein distance calculations (Runkler & Bezdek, 2000) and then standardised subjective criteria be applied to transform the terms with high values. For those specialists who want to reinterpret our results by unifying terms, we have included the calculation of Levenshtein's distances from our study in the supplementary material S5 (Peset et al., 2018e). Another interesting choice is to simplify the words using the Porter stemming algorithm. This process focuses on removing morphological and inflectional endings from words in English. The results obtained in the present research when this method is applied, can be consulted in our supplementary material S6 (Peset et al., 2018a). However, as can be seen, the data differ only slightly from those provided in the results section of this article.

Conclusions

In our work we have presented a new aspect of the bibliometric analysis of a scientific field that concerns the dynamics of author keywords in research papers. As an example, our procedure has been applied for developing a survival analysis of these keywords in the case of the LIS area. We have shown that in a standard year, approximately 60% of new keywords appear in the LIS knowledge area. Many are ephemeral because more than half are not used again later. In general, the terms most used in the LIS area come from other areas. The average survival time of these keywords is approximately 3 years, being slightly higher in the case of keywords that were published in journals classified in the second quartile of the LIS area. Despite these results, it is necessary to bear in mind that this survival period could be affected by the end date and by the observation time used.

REFERENCES

Aharony, N. (2012). Library and Information Science research areas: A content analysis of articles from the top 10 journals 2007–8. *Journal of Librarianship and Information Science*, *44*(1), 27–35. https://doi.org/10.1177/0961000611424819

Aizawa, A., & Kageura, K. (2003). Calculating association between technical terms based on co-occurrences in keyword lists of academic papers. *Systems and Computers in Japan*, *34*(3), 85–95. https://doi.org/10.1002/scj.1197

Athukorala, K., Hoggan, E., Lehtiö, A., Ruotsalo, T., & Jacucci, G. (2013). Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–11. https://doi.org/10.1002/meet.14505001041

Borgman, C. L. (2007). The View from Here. In *Scholarship in the digital age: information, infrastructure and the internet* (p. 360). MIT Press, MA.

Box-Steffensmeier, J. M., Cunha, R. C., Varbanov, R. A., Hoh, Y. S., Knisley, M. L., & Holmes, M. A. (2015). Survival Analysis of Faculty Retention and Promotion in the Social Sciences by Gender. *PLOS ONE*, *10*(11), e0143093. https://doi.org/10.1371/journal.pone.0143093

Buckland, M. K. (2012). Obsolescence in subject description. *Journal of Documentation*, *68*(2), 154–161. https://doi.org/10.1108/00220411211209168

Colley, A., & Maltby, J. (2008). Impact of the Internet on our lives: Male and female personal perspectives. *Computers in Human Behavior*, *24*(5), 2005–2013.

Dehdarirad, T., Villarroya, A., & Barrios, M. (2014). Research trends in gender differences in higher education and science: A co-word analysis. *Scientometrics*, *101*(1), 273–290.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management*, *37*(6), 817–842. https://doi.org/10.1016/S0306-4573(00)00051-0

Dotsika, F., & Watkins, A. (2017). Identifying potentially disruptive trends by means of keyword network analysis. *Technological Forecasting and Social Change*, *119*, 114–127. https://doi.org/10.1016/j.techfore.2017.03.020

Figuerola, C. G., Marco, F. J. G., & Pinto, M. (2017). Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. *Scientometrics*, *112*(3), 1507–1535. https://doi.org/10.1007/s11192-017-2432-9

Gan, C., & Wang, W. (2014). A Bibliometric Analysis of Social Media Research from the Perspective of Library and Information Science. *I3E*, 23–32. Retrieved from http://link.springer.com/content/pdf/10.1007/978-3-662-45526-5.pdf#page=37

Gil-Leiva, I., & Alonso-Arroyo, A. (2007). Keywords given by authors of scientific articles in database descriptors. *Journal of the American Society for Information Science and Technology*, *58*(8), 1175–1187. https://doi.org/10.1002/asi.20595

González-Alcaide, G., Castelló-Cogollos, L., Navarro-Molina, C., Aleixandre-Benavent, R., & Valderrama-Zurián, J. C. (2008). Library and information science research areas:

Analysis of journal articles in LISA. *Journal of the Association for Information Science and Technology*, *59*(1), 150–154.

Halevi, G., & Moed, H. F. (2013). The thematic and conceptual flow of disciplinary research: A citation context analysis of the journal of informetrics, 2007. *Journal of the American Society for Information Science and Technology*, *64*(9), 1903–1913. https://doi.org/10.1002/asi.22897

Han, H., Gui, J., & Xu, S. (2014). *Revealing research themes and their evolutionary trends using bibliometric data based on strategic diagrams*. 653–659. https://doi.org/10.1109/ISCC-C.2013.121

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

Hjørland, B. (2000). Library and information science: practice, theory, and philosophical basis. *Information Processing & Management*, *36*(3), 501–531. https://doi.org/10.1016/S0306-4573(99)00038-2

Hjørland, B. (2017). Library and information science (LIS). In *Encyclopedia of Knowledge Organization*. Retrieved from http://www.isko.org/cyclo/lis

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science*, *46*(6), 400–425. https://doi.org/10.1002/(SICI)1097-4571(199507)46:6<400::AID-ASI2>3.0.CO;2-Y

Hu, C.-P., Hu, J.-M., Deng, S.-L., & Liu, Y. (2013). A co-word analysis of library and information science in China. *Scientometrics*, *97*(2), 369–382. https://doi.org/10.1007/s11192-013-1076-7

Jan Brophy, & David Bawden. (2005). Is Google enough? Comparison of an internet search engine with academic library resources. *Aslib Proceedings*, *57*(6), 498–512. https://doi.org/10.1108/00012530510634235

Jones, K. S., & Jackson, D. M. (1970). The use of automatically-obtained keyword classifications for information retrieval. *Information Storage and Retrieval*, *5*(4), 175–201.

Kevork, E. K., & Vrechopoulos, A. P. (2009). CRM literature: conceptual and functional insights by keyword analysis. *Marketing Intelligence & Planning*, *27*(1), 48–85. https://doi.org/10.1108/02634500910928362

Lancaster, F. W. (2003). *Indexing & Abstracting in Theory & Practice* (Edición: 3). Champaign, Ill.: Univ of Illinois Graduate School of.

Lee, S. (2016). A Study on Research Trends in Public Library Research in Korea Using Keyword Networks. *Libri*, *66*(4), 263–274. https://doi.org/10.1515/libri-2016-0052

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.

Li, M. (2018). Classifying and ranking topic terms based on a novel approach: role differentiation of author keywords. *Scientometrics*, *116*(1), 77–100. https://doi.org/10.1007/s11192-018-2741-7

Liu, J., Tian, J., Kong, X., Lee, I., & Xia, F. (2019). Two decades of information systems: a bibliometric review. *Scientometrics*, *118*(2), 617–643. https://doi.org/10.1007/s11192-018-2974-5

Losee, R. M. (1995). The Development and Migration of Concepts from Donor to Borrower Disciplines: Sublanguage Term Use in Hard & Soft Sciences. *ArXiv:Cmp-Lg/9509004*. Retrieved from http://arxiv.org/abs/cmp-lg/9509004

McClure, C. R., & Bishop, A. (1989). The Status of Research in Library/Information Science: Guarded Optimism. *College & Research Libraries*, *50*(2), 127–143. https://doi.org/10.5860/crl_50_02_127

Mela, C., Roos, J., & Deng, Y. (2013). A keyword history of Marketing Science. *Marketing Science: The Marketing Journal of INFORMS*, *32*(1), 8–18.

Michael Hall, C. (2011). Publish and perish? Bibliometric analysis, journal ranking and the assessment of research quality in tourism. *Tourism Management*, *32*(1), 16–27. https://doi.org/10.1016/j.tourman.2010.07.001

Milojević, S., Sugimoto, C. R., Yan, E., & Ding, Y. (2011). The cognitive structure of Library and Information Science: Analysis of article title words. *Journal of the American Society for Information Science and Technology*, *62*(10), 1933–1953. https://doi.org/10.1002/asi.21602

Naseer, M. M., & Mahmood, K. (2009). Use of Bibliometrics in LIS Research. *LIBRES: Library and Information Science Research Electronic Journal; Perth*, *19*(2), 1–11.

Névéol, A., Doğan, R. I., & Lu, Z. (2010). Author Keywords in Biomedical Journal Articles. *AMIA Annual Symposium Proceedings*, *2010*, 537–541.

Niu, X., & Hemminger, B. M. (2012). A study of factors that affect the information-seeking behavior of academic scientists. *Journal of the American Society for Information Science and Technology*, *63*(2), 336–353. https://doi.org/10.1002/asi.21669

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*, *4*, 5. https://doi.org/10.1186/2046-4053-4-5

Onyancha, O. B. (2018). Forty-Five Years of LIS Research Evolution, 1971–2015: An Informetrics Study of the Author-Supplied Keywords. *Publishing Research Quarterly*, *34*(3), 456–470. https://doi.org/10.1007/s12109-018-9590-3

Peset, F., Garzón-Farinos, F., González, L., García-Massó, X., Ferrer-Sapena, A., Toca-Herrera, J., & Sánchez-Perez, E. (2018a). Supplementary material S1. Search strategy. *Figshare*. Retrieved from https://figshare.com/s/60c1a3aa37b9bd1db596

Peset, F., Garzón-Farinos, F., González, L., García-Massó, X., Ferrer-Sapena, A., Toca-Herrera, J., & Sánchez-Perez, E. (2018b). Supplementary material S2. Database. *Figshare*. Retrieved from https://figshare.com/s/5313334d203f800e869c

Peset, F., Garzón-Farinos, F., González, L., García-Massó, X., Ferrer-Sapena, A., Toca-Herrera, J., & Sánchez-Perez, E. (2018c). Supplementary material S3. New Keywords (2004). *Figshare*. Retrieved from https://figshare.com/s/cf78765ab8320de9eea0

Peset, F., Garzón-Farinos, F., González, L., García-Massó, X., Ferrer-Sapena, A., Toca-Herrera, J., & Sánchez-Perez, E. (2018d). Supplementary material S4. Survival analysis 5, 10, 15, 20 years. *Figshare*. Retrieved from https://figshare.com/s/4b687afad323182e4ecd

Peset, F., Garzón-Farinos, F., González, L., García-Massó, X., Ferrer-Sapena, A., Toca-Herrera, J., & Sánchez-Perez, E. (2018e). Supplementary material S5. Levenshtein distance. *Figshare*. Retrieved from https://figshare.com/s/c82e03d769aedbe24efa

Radhakrishnan, S., Erbis, S., Isaacs, J. A., & Kamarthi, S. (2017). Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature. *PLOS ONE*, *12*(3), e0172778. https://doi.org/10.1371/journal.pone.0172778

Rayward, W. B. (1985). Library and information science: An historical perspective. *The Journal of Library History*, *20*(2), 120–136.

Rayward, W. B. (2005). The historical development of information infrastructures and the dissemination of knowledge: A personal reflection. *Bulletin of the American Society for Information Science and Technology*, *31*(4), 19–22. https://doi.org/10.1002/bult.1720310407

Runkler, T. A., & Bezdek, J. C. (2000). Automatic keyword extraction with relational clustering and Levenshtein distances. *Ninth IEEE International Conference on Fuzzy Systems. FUZZ- IEEE 2000 (Cat. No.00CH37063)*, *2*, 636–640 vol.2. https://doi.org/10.1109/FUZZY.2000.839067

Santos, J. B., & Irizo, F. J. O. (2002). Modelización de la antigüedad de las citas en la literatura científica con datos censurados a la derecha. *Revista Española de Documentación Científica*, *25*(2), 141–150.

Santos, J. B., & Irizo, F. J. O. (2005). Modelling citation age data with right censoring. *Scientometrics*, *62*(3), 329–342.

Scimago Journal & Country Rank. (n.d.). Retrieved 29 August 2017, from http://www.scimagojr.com/

Singer, J. D., & Willett, J. B. (1993). It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics*, *18*(2), 155–195. https://doi.org/10.3102/10769986018002155

Su, H.-N., & Lee, P.-C. (2010). Mapping knowledge structure by keyword co-occurrence: A first look at journal papers in Technology Foresight. *Scientometrics*, *85*(1), 65–79. https://doi.org/10.1007/s11192-010-0259-8

Sun, J. (1997). Regression Analysis of Interval-Censored Failure Time Data. *Statistics in Medicine*, *16*(5), 497–504. https://doi.org/10.1002/(SICI)1097-0258(19970315)16:5<497::AID-SIM435>3.0.CO;2-J

Tang, R. (2004). Evolution of the interdisciplinary characteristics of information and library science. *Proceedings of the American Society for Information Science and Technology*, *41*(1), 54–63. https://doi.org/10.1002/meet.1450410107

Tuomaala, O., Järvelin, K., & Vakkari, P. (2014). Evolution of library and information science, 1965–2005: Content analysis of journal articles. *Journal of the Association for Information Science and Technology*, *65*(7), 1446–1462. https://doi.org/10.1002/asi.23034

Uddin, S., & Khan, A. (2016). The impact of author-selected keywords on citation counts. *Journal of Informetrics*, *10*(4), 1166–1177. https://doi.org/10.1016/j.joi.2016.10.004

Vakkari, P. (1994). Library and Information Science: Its Content and Scope. In *Advances in Librarianship*: *Vol. 18*. *Advances in Librarianship* (Vol. 18, pp. 1–55). https://doi.org/10.1108/S0065-2830(1994)0000018003

Walters, W. H., & Wilder, E. I. (2016). Disciplinary, national, and departmental contributions to the literature of library and information science, 2007–2012. *Journal of the Association for Information Science and Technology*, *67*(6), 1487–1506. https://doi.org/10.1002/asi.23448

Wang, L., Lin, J., & Cui, W. (2014). *Where do breakthrough ideas come from? Characteristics of scientists' research behaviors*. 61–65. https://doi.org/10.1109/ICMIT.2014.6942401

Wang, Z.-Y., Li, G., Li, C.-Y., & Li, A. (2012). Research on the semantic-based co-word analysis. *Scientometrics*, *90*(3), 855–875. https://doi.org/10.1007/s11192-011-0563-y

Xu, J., Bu, Y., Ding, Y., Yang, S., Zhang, H., Yu, C., & Sun, L. (2018). Understanding the formation of interdisciplinary research from the perspective of keyword evolution: a case study on joint attention. *Scientometrics*, *117*(2), 973–995. https://doi.org/10.1007/s11192-018-2897-1

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. https://doi.org/10.1016/j.joi.2015.12.003

Zhang, J., Yu, Q., Zheng, F., Long, C., Lu, Z., & Duan, Z. (2016). Comparing keywords plus of WOS and author keywords: A case study of patient adherence research. *Journal of the Association for Information Science and Technology*, *67*(4), 967–972. https://doi.org/10.1002/asi.23437

Zins, C. (2007). Conceptions of information science. *Journal of the Association for Information Science and Technology*, *58*(3), 335–350.