

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Mapping a Research Field: Analyzing the Research Fronts in an Emerging Discipline

Gerardo Tibaná-Herrera,
María Teresa Fernández-Bajón and
Félix de Moya-Anegón

Additional information is available at the end of the chapter

<http://dx.doi.org/10.5772/intechopen.76731>

Abstract

The mapping overlay technique described in the scientific literature to analyze scientific domains must be complemented with procedures to identify and analyze the research fronts included in the cognitive structure of the represented domain. One possibility is the use of wordcloud maps to visually represent the cognitive structure of a discipline in any thematic domain, taking advantage of its capacity for abstraction and impact on the audience to stimulate new research processes. The case described in this chapter proposes an analysis of an emerging scientific discipline by using this combination of techniques (superposition and wordcloud) to explore its possibilities and limitations.

Keywords: bibliometric, mapping overlay, wordcloud, emerging field, e-learning, SCImago Journal & Country Rank

1. Introduction

The world scientific production analysis contributes, among many other things, to define the knowledge areas and subject categories that structure the generation of knowledge. Each classification system of scientific production defines its own areas and categories, which are mostly accepted by the scientific community that consults and feeds them. In this way, Scopus¹ classifies the works into 5 thematic clusters (life sciences, physical sciences, health sciences, social sciences and humanities), 27 knowledge areas and more than 300 subject categories.

¹<http://www.scopus.com>

Web of Science² does it in 3 knowledge areas (sciences, social sciences and arts and humanities) and 250 thematic categories.

Scopus currently has more than 70 million records and a defined group of metadata³ that are rigorously linked to each publication to describe its academic, social and geopolitical context. These two characteristics, having large volumes of structured information, are the inputs for the application of visualization techniques that generate new representations of knowledge, thus becoming powerful tools for science analysis.

The bibliometric data are very valuable to identify the scientific publications with the greatest impact in a given discipline, (i.e., Information Systems [1], Renewable Energy, Sustainability and Environment [2], to recognize different scientific fields and understand their internal dynamics and cognitive structure [3], either as an already consolidated research field or as an emerging discipline.

There are multiple methods and tools to visualize bibliometric information. For example, the distance-based, the graph-based or the time-based [4]. Mapping and clustering are also used to analyze the research fields of a scientific domain and the relationship between research fields and the evolution of the domain over time. As a tool, VOSViewer⁴ assures the comprehensive visualization of nodes labels on the map. These maps, called science maps, help to locate research results to explore collaborations and publication trends, to observe the evolution of a certain subject or discipline and for benchmarking activities between regions, countries, institutions, authors and disciplines [5]. However, the visualization must have the capacity to handle large amounts of data at a small and large scale. This reduces the visual search time, providing a better understanding of a complex data set. It also reveals relationships that otherwise would not be noticed, allowing a data set to be viewed simultaneously from several perspectives, aiding the formulation of hypotheses and being an effective source of communication [6].

Through the overlay of science maps, the research bodies can be located visually within the sciences, analyzing the scientific development of properly established disciplines, trends or emerging research topics that do not fit into traditional subject categories. This is achieved thanks to the existence or construction of a stable corpus on which another smaller body can be overlaid [7], producing intuitive comparisons, of greater interpretation and with the potential to be used in scientific analysis.

In its essence, science maps are matrices of similarity measures, calculated from the correlation between items of information present in the structure of scientific communication. In other words, they show the disciplinary structure of the sciences in terms of publications. The stable or base map is constructed with bibliographic data from a database that has a definite categorization of the sciences. The analysis made from the overlap will be conditioned by the size of the data selected for it.

²<http://clarivate.com/products/journal-citation-reports/>

³<https://www.elsevier.com/solutions/scopus/content>

⁴<http://www.vosviewer.com>

In the words of Guzmán, “we can say that the analysis of information with science maps, supported by metric information studies, allows graphically representing the relationships between documents published by specific disciplines or scientific fields. These show the sub-areas of research in which the discipline has been focused over the years in order to identify, analyze and visualize the intellectual structure, as well as the temporal evolution in which the analyzed disciplines are being developed.” [8].

Based on the above, science maps contribute to the identification of emerging disciplines by categorizing the publications that constitute their scientific communication channel [9].

However, a very select group of specialists usually carries out the analysis of these research products, since the results obtained are not easy enough to understand for most of the scientific community that is interested in knowing in detail the paths and trends that their discipline is taking. Faced with this need, other visualization techniques, such as wordclouds [10], infographics [11] and dashboards [12] have been positioned in virtual media as an alternative for the research results to achieve greater diffusion beyond the borders of scientific communication channels [13]. Wordclouds are used mostly to visualize a data set collected from surveys or forms. Among its advantages are: (a) its ability to abstract towards the essential, identifying and grouping existing patterns in writing [10], (b) they help to provide a general sense of the text (the same visceral response does not occur when looking at a text page) through the analysis of sentiment [14], (c) they provide a quick response on possible topics of interest and research for their community [15], (d) the visual representation of data generates impact among the audience, stimulating more questions than answers, and (e) they allow to share the results of the research in a way that does not require a deep understanding of the technicalities. Its link with the bibliometric analysis can be established considering the keywords field as the set of data collected from users (researchers) in a form (submit manuscript).

This study combines the use of the mapping overlay technique with the visualization of terms in wordclouds to represent the research fronts of a subject, in this case, the e-learning emerging discipline. The aim is to determine if this technique combination produces more intuitive, dynamic and easily accessible results for researchers and non-researchers.

2. Mapping a research field

To perform the research field mapping, we must first establish a body of documents to perform the bibliometric analysis, ensuring access to the bibliometric data of this set of publications. To analyze the e-learning case, we started with the methodology and findings of Tibaná-Herrera and others [9] for the subject categorization.

Secondly, the subject research fronts are identified, which determine the consolidation of the different tendencies over time that have contributed to the development and growth of the subject in scientific communications [16]. We propose the use of wordclouds composed of keywords [17], to visualize the research fronts of the field due to its representation capacity and rapid appropriation of the community to which it is presented.

2.1. Establishment of the body of documents

We start from the base that every research field has a set of scientific communications that contribute to the development of the subject. To identify these communications and analyze them, the subsequent steps can be followed:

Step 1. Definition of descriptors. It is about knowing all those terms present in the primary scientific literature with which the subject has been described. As expected, we start from a core term, which is generally the same as the research field. With this term, all the publications whose title, summary and keywords include the core term are identified in a comprehensible database.

E-learning case

- Core term: e-learning
- Data source: SCOPUS, database that indexes mostly journals and conference proceedings [18].

The search results should be refined according to the desired coverage degree in the analysis and the access availability of the bibliometric data.

E-learning case

- Publication type: Journal and Conference Proceeding
- Document type: Article, conference paper and review
- Analyzed timespan: 2012–2014. It corresponds to a period in which there is a stable worldwide production in e-learning, since in the previous period it was in constant growth and in the following period there was a significant decrease in production [19].
- Language: English.

The set of publications obtained can be used in its entirety or from a statistically representative sample.

E-learning case

- Results: 9291
- Representative sample: 2000 (21.6%)

Then, a bibliometric analysis based on keywords co-occurrence is carried out, aimed to determining the primary descriptors that are mostly present in the publications, their relationships and relevance, by means of the Visualization of Similarities (VoS) technique [20]. Additionally, they include secondary possible descriptors that reflect the same meaning, fruit of the linguistic similarities and/or acronyms or abbreviations that are used in the natural language. For example, when including the keywords of an article you can choose to use the *e-learning* or *elearning* descriptor [21].

E-learning case

- Keywords: 4521
- Primary descriptors: 51. E-learning, LMS, b-learning, online learning, Moodle, m-learning, ICT, learning objects, technology acceptance model, e-learning platform, adaptive learning, e-assessment, web-based learning, virtual learning environments, adult learning, informal learning, instructional design, SCORM, augmented reality, educational technology, intelligent tutoring systems, remote laboratory, simulation, learning analytics, learning environments, e-learning 2.0, teaching and learning, interactive learning environments, educational data mining, gamification, learning design, social learning, lifelong learning, metadata, MOOC, virtual classroom, labview, learning methods, personal learning environments, adaptive e-learning systems, computer-based learning, information literacy, virtual learning, Blackboard, continuing education, game-based learning, interactive learning, personalized learning, recommender systems, virtual laboratories, virtual reality.
- Secondary descriptors: 13. elearning, electronic learning, Learning management system, blearning, blended learning, mlearning, mobile learning, Information and communications technologies, eassessment, electronic assessment, VLE, Massive Open Online Courses and PLE.

Step 2. Correspondence of publications and descriptors. In a matrix containing all the indexed scientific publications and the primary and secondary descriptors identified, the number of articles published by the Conference Proceeding or the Journal with that descriptor in the title, abstract and keywords fields is recorded at each crossing. It is very important to use the same selection criteria described in the previous step to ensure information integrity. Then, the primary and secondary descriptors related to the same term are added, assuming that the sum reflects unique publications related to each other by the descriptors.

E-learning case

- Journals and conference proceedings included in the matrix: 12.923

Step 3. Percentage of participation in the subject (PP). It is the percentage of articles in the publication that are related to the subject during the timespan established in the initial criteria, this is done by taking the maximum number of articles per descriptor, bearing in mind that an article may be related to more than one descriptor.

E-learning case

Correspondence matrix description (**Figure 1**):

- 3.680 journals and conference proceedings do not have any publication related with any of the 64 descriptors.
- 7.801 journals and conference proceedings have a PP lower than 5%.

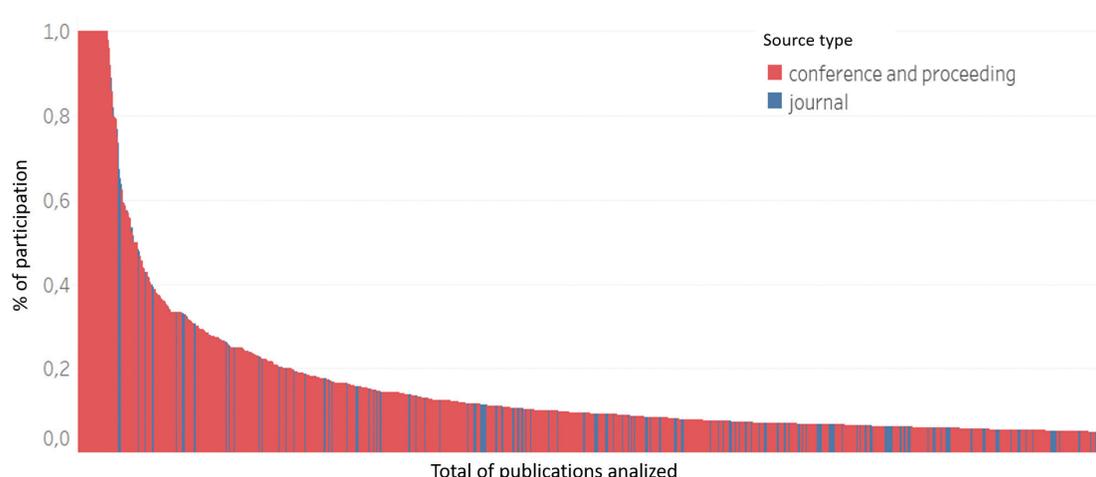


Figure 1. Percentage of participation (PP) of the term in journals and conferences. Source: [9].

Step 4. Cut-off point for the inclusion of publications in the analysis. You must determine the cut-off point over the PP from which the publications for the categorization of the thematic will be included. Other studies have classified publications among “pure”, “hybrid” and “unrelated” publications in a given subject [1] and on the determination of the core set of publications [21]. However, we believe that this value should be established through the combination between the maximum allowed error of the subject relation of the publication and the average PP of the total set of publications. The higher the cut-off point, the greater the precision in the selection of journals will be. Although, this precision means a reduced volume, and if not, a low cut-off point increases the error in the selection and its volume. Once the cut-off point is established, all publications that exceed this threshold are considered as the basic set of analysis of the emerging subject category.

E-learning case

- The set of publications must maintain an average PP higher than 50%, for which the cut-off point per publication was established at 25% (coinciding with the classification of pure and hybrid publications [1]).
- The cut-off point included 11 publications that were excluded because they defined other areas of knowledge in their scope.
- 82 journals and 137 conference proceedings that meet the criteria of the methodology were identified.

Step 5. Publication set analysis. The set of selected publications is analyzed under a bibliometric approach (a) to determine if it represents the existence of a scientific community that communicates its knowledge through these channels and (b) to recognize it as an emerging and distinctive scientific discipline that can be defined as a transversal thematic category [5]. For this, the mapping overlay technique [7] can be used, which facilitates the exploration of the knowledge bases of an emerging discipline and its evolutionary dynamics. This technique requires a base map on which to overlay a local map (thematic) and thus make comparisons. This overlap allows placing the discipline in the general topology of scientific knowledge and identifying whether a cluster effect occurs, which should be considered as evidence of the existence of a specific disciplinary field from the point of view of scientific communication guidelines followed by the researchers.

The relation degree of publications is established by the normalized value produced by the combination of citations, co-cites and coupling [22, 23]. In addition, this analysis can be enriched with the distribution by clusters that visualization tools perform, such as VOSViewer [24].

E-learning case

- The base map is a global map of science that includes the total number of publications indexed in SCOPUS, made up of 7 clusters, which in a clockwise and broad sense can be named as follows: Social Sciences (red), Psychology (light cyan), Medicine (green), Health Sciences (purple), Life Sciences (yellow), Physical Sciences (dark cyan) and Engineering and Computer Science (blue) (**Figure 2**).
- The composite indicator was arranged by SCImago Journal & Country Rank⁵.
- The local map that is overlaid on the global map of science is the set of 219 publications selected in the previous step (**Figure 3**).
- There is a cluster effect that shows a high cohesion among publications, which is sufficient evidence, in terms of scientific communication, that e-learning is a distinctive

⁵<http://www.scimagojr.com>

scientific discipline, since there is a network of relationships and interactions that are established between the authors and scientists who share thought structures, cooperation patterns, language and forms of communication.

- The publications distribution shows a main group in Social Sciences and other small groups in Computer Science and Psychology.

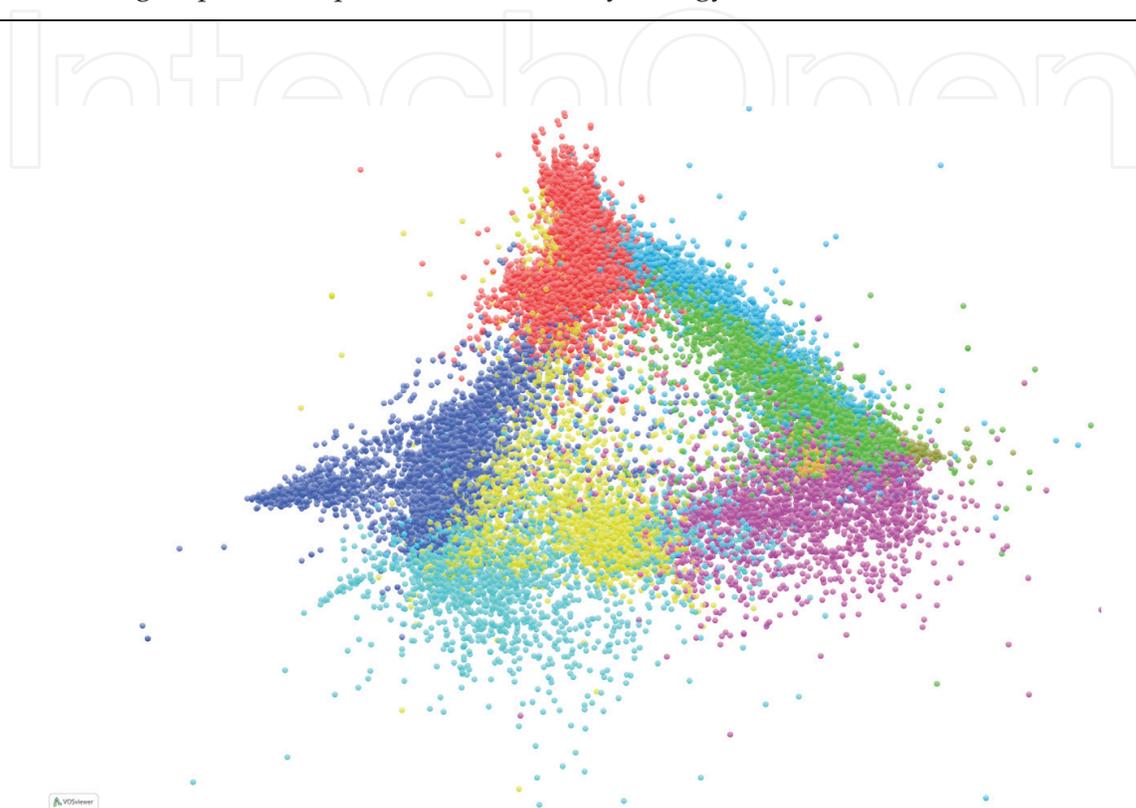


Figure 2. Global map of science based on SCOPUS and SCImago Journal & Country Rank using VOSviewer with its density map setting (Source: [9]).

2.2. Identification of research fronts

To identify the research fronts through the visualization of keywords in a wordcloud, it is necessary to identify the body of publications on which the analysis is going to be carried out (previous section). Then, all the keywords of the publications are extracted, keeping the same filters defined in the previous stages, with the confidence of finding a set of structured and well-defined terms. This technique provides value when the data has a treatment that ensures a correct interpretation. This is done through two tasks, being the first to refine the set of terms (which can be in the order of thousands) to obtain those that are mostly different and that can be visually represented without loss of information. The refinement process may include a minimum threshold of articles published by a journal or conference report to ensure that there is a volume and regularity guaranteed in the conceptual development of the thematic. It can also be refined by defining the number of terms to be displayed in the wordcloud.

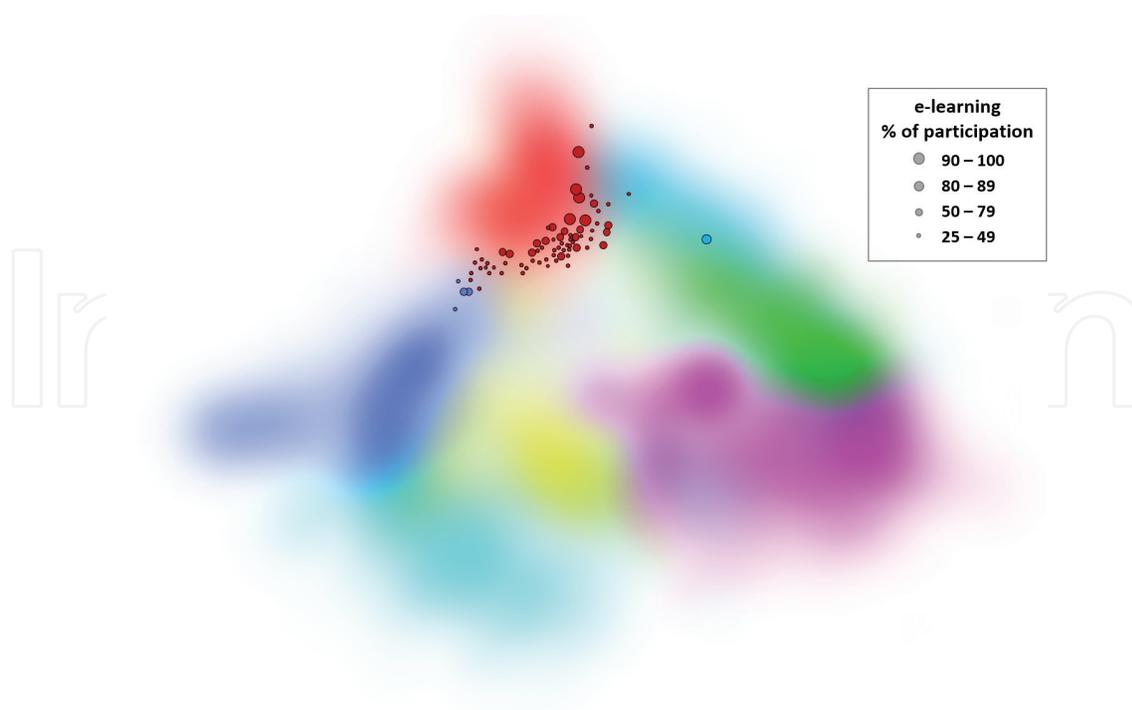


Figure 3. Distribution of publications related to the thematic, using the mapping overlay technique with VOSViewer in its density map configuration. The color of the publication indicates the area of knowledge in which it is superimposed and its size corresponds to the percentage of participation. The size of the selected publications has been modified for visual purposes (Source: [9]).

E-learning case

- Publication type: Journal and Conference Proceeding
- Document type: Article, conference paper and review
- Analyzed timespan: 2012–2014.
- Language: English.
- Minimum number of papers published by Journal/Conference Proceeding: 100
- Number of terms to display: 100

The second task is to configure the variables that determine the form of the wordcloud, among which are:

1. Keep each term with its own length. You can fall into the error of disaggregating terms, for example, the term *Information and Communication Technologies* should remain as one and not separate it into 3 or 4 parts.
2. Don't include terms in the visualization that correspond to the same name of the scientific field analyzed, places, dates, proper names, names of organizations and all others that don't contribute to the identification of research fronts.

3. Define simple shapes to represent the cloud. Today there are multiple wordcloud creation tools. Most of them allow to use a defined image for the cloud layout. It is recommended to use images without internal content, only frame, so that the words can be distributed inside without obstacles.
4. Select a Sans Serif font. The wordclouds are presented more frequently in digital media, in which a clean, non-blurred reading is sought, to avoid visual fatigue
5. Define an intention for the color usage. The visual representation should be as enriched as possible. Therefore, the color defined for each term must show its own characteristic. A good intention of color is to establish clusters of terms [23] that determine the main research fronts.

E-learning case

Examples of wordclouds.

Option 1:

Option 2:

Option 3:

Finally, by means of a rapid visual analysis of the generated wordcloud, the research fronts of the scientific field can be identified in a differentiated way.

E-learning case

- Based on the results shown in **Figures 4–6**, two significant clusters can be identified (**Table 1**):
- The most outstanding research fronts of e-learning are those that analyze the design and construction of interactive learning environments and teaching and learning strategies in the virtual modality

A limitation of wordclouds, that can affect the reader's interpretation, is the term length that can capture a quick attention being located in a central place of the visualization without having significant weight. However, this visualization technique is a powerful tool to abstract relevant information from large volumes of information, in addition, it can be used to observe the main trends of other bibliometric data. For example, journals and congresses with the greatest influence in the discipline or the institutions and countries that contribute the most to the discipline productivity.

publications for other bibliometric analyzes. Through this technique it was determined that e-learning has its scientific development mainly in the social sciences.

The visualization of the main keywords present in the set of publications of a discipline through wordclouds, allows to clearly identify the research fronts of this subject, by grouping the research topics and showing their relative weight in the scientific development of the discipline. In the case study, two main research fronts were identified in e-learning, interactive learning environments and teaching and learning strategies.

Acknowledgements

To SCImago Research Group for providing citation data for publications.

Conflict of interest

Not applicable.

Notes/Thanks/Other declarations

The data related to this research were obtained, on the one hand, from the access to SCOPUS and on the other, provided by SCImago Research Group. These are protected by licensing and copyright respectively.

Author details

Gerardo Tibaná-Herrera^{1,2*}, María Teresa Fernández-Bajón¹ and Félix de Moya-Anegón³

*Address all correspondence to: gtibana@gmail.com

1 Complutense University of Madrid, Madrid, Spain

2 SCImago Research Group, Bogotá, Colombia

3 SCImago Research Group, Madrid, Spain

References

- [1] Chan HC, Guness V, Kim HW. A method for identifying journals in a discipline: An application to information systems. *Information and Management [Internet]*. 2015;52(2): 239-246. DOI: 10.1016/j.im.2014.11.003

- [2] Romo Fernández LM, Guerrero Bote VP, Moya AF. Análisis de la producción científica española en energías renovables, sostenibilidad y medio ambiente (Scopus, 2003-2009) en el contexto mundial. *Investigación Bibliotecológica Archivonomía, Bibliotecología e Información*. 2013;**27**(60):125-151. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S0187358X13725462>
- [3] Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F. 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 [Internet]*. 2011;**5**(1):146-166. DOI: 10.1016/j.joi.2010.10.002
- [4] van Eck NJ, Waltman L. Visualizing bibliometric networks [Internet]. *Measuring Scholarly Impact*. 2014. pp. 285-320. Available from: http://link.springer.com/10.1007/978-3-319-10377-8_13
- [5] Leydesdorff L, de Moya-Anegón F, Guerrero-Bote VP. Journal maps, interactive overlays, and the measurement of interdisciplinarity on the basis of Scopus data (1996–2012). *Journal of the Association for Information Science and Technology [Internet]*. 2015 May; **66**(5):1001-1016. DOI: 10.1002/asi.23243
- [6] Börner K, Chen C, Boyack KW. Visualizing knowledge domains. *Annual Review of Information Science and Technology [Internet]*. 2005;**37**(1):179-255. DOI: 10.1002/aris.1440370106
- [7] Rafols I, Porter AL, Leydesdorff L. Science overlay maps: A new tool for research policy and library management. *Journal of the Association for Information Science and Technology [Internet]*. 2010 Sep;**61**(9):1871-1887. DOI: 10.1002/asi.21368
- [8] Guzmán Sánchez MV, Trujillo Cancino JL. Los mapas bibliométricos o mapas de la ciencia: una herramienta útil para desarrollar estudios métricos de información. *Bible University [Internet]*. 2013;**16**(2):95-108. Available from: <http://revistas.unam.mx/index.php/rbu/article/view/43851>
- [9] Tibaná-Herrera G, Fernández-Bajón MT, de Moya-Anegón F. Categorization of an emerging discipline in the world publication system (SCOPUS): E-learning. 2017 Oct 16 [cited 2018 Feb 1]. Available from: <http://arxiv.org/abs/1710.05723>
- [10] Flamary R, Anguera X, Oliver N. Spoken WordCloud: Clustering recurrent patterns in speech. In: 2011 9th International Workshop on Content-Based Multimedia Indexing (CBMI) [Internet]. IEEE; 2011. pp. 133-138. Available from: <http://ieeexplore.ieee.org/document/5972534/>
- [11] Alcívar M. Information visualisation as a resource for popularising the technical-biomedical aspects of the last Ebola virus epidemic: The case of the Spanish reference press. *Public Understanding of Science [Internet]*. Apr 10, 2018;**27**(3):365-381. Available from: <http://journals.sagepub.com/doi/10.1177/0963662517702047>
- [12] Verbert K, Duval E, Klerkx J, Govaerts S, Santos JL. Learning analytics dashboard applications. *American Behavioral Scientist [Internet]*. 2013 Oct 28;**57**(10):1500-1509. Available from: <http://journals.sagepub.com/doi/10.1177/0002764213479363>

- [13] Dinsmore A, Allen L, Dolby K. Alternative perspectives on impact: The potential of ALMs and Altmetrics to inform funders about research impact. *PLoS Biology* [Internet]. 2014; **12**(11):e1002003. DOI: 10.1371/journal.pbio.1002003
- [14] Bashri MFA, Kusumaningrum R. Sentiment analysis using latent Dirichlet allocation and topic polarity wordcloud visualization. In: 2017 5th International Conference on Information and Communication Technology (ICoICT7) [Internet]. IEEE; 2017. pp. 1-5. Available from: <http://ieeexplore.ieee.org/document/8074651/>
- [15] Jo Y, Kim E, Shin Y. Graphical keyword service for research papers with text-mining method. In: ICCDA '17 Proceedings of the International Conference on Compute and Data Analysis [Internet]. 2017. pp. 185-190. Available from: <https://dl.acm.org/citation.cfm?id=3093242&dl=ACM&coll=DL>
- [16] Pinto M. Viewing and exploring the subject area of information literacy assessment in higher education (2000–2011). *Scientometrics* [Internet]. 2015 Jan 15; **102**(1):227-245. Available from: <http://link.springer.com/10.1007/s11192-014-1440-2>
- [17] Cantos-Mateos G, Zulueta M-Á, Vargas-Quesada B, Chinchilla-Rodríguez Z. Estudio evolutivo de la investigación española con células madre. Visualización e identificación de las principales líneas de investigación. *El Profesional de la Información* [Internet]. 2014; **23**(3):259-271. Available from: <https://recyt.fecyt.es/index.php/EPI/article/view/epi.2014.may.06>
- [18] Leydesdorff L, De Moya-Anegón F, Guerrero-Bote VP. Journal maps on the basis of scopus data: A comparison with the journal citation reports of the ISI. *Journal of the American Society for Information Science and Technology*. 2010; **61**(2):352-369
- [19] Tibaná-Herrera G, Fernández-Bajón MT, de Moya-Anegón F. Global Analysis of the E-Learning Scientific Domain: A Declining Category? *Scientometrics* [Internet]. 2017 Dec 5. Available from: <http://link.springer.com/10.1007/s11192-017-2592-7>
- [20] Waltman L, van Eck NJ, Noyons ECM. A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics* [Internet]. 2010; **4**(4):629-635. DOI: 10.1016/j.joi.2010.07.002
- [21] Chiang JK, Kuo C-W, Yang Y-H. A bibliometric study of E-learning literature on SSCI database. In: International Conference on Technologies for E-learning and Digital Entertainment [Internet]. 2010. pp. 145-55. Available from: http://link.springer.com/10.1007/978-3-642-14533-9_15
- [22] Madhugiri VS, Ambekar S, Strom SF, Nanda A. A technique to identify core journals for neurosurgery using citation scatter analysis and the Bradford distribution across neurosurgery journals. *Journal of Neurosurgery* [Internet]. 2013; **119**(5):1274-1287. <http://thejns.org/doi/10.3171/2013.8.JNS122379> or <https://doi.org/10.3171/2013.8.JNS122379>
- [23] Hassan-Montero Y, Guerrero-Bote V, De-Moya-Anegón F. Graphical interface of the SCImago journal and country rank: An interactive approach to accessing bibliometric information. *El Profesional de la Información* [Internet]. 2014; **23**(3):272-278. Available from:

<http://www.scopus.com/inward/record.url?eid=2-s2.0-84903581415&partnerID=40&md5=cd75fe9502aa257c52bf0b9334e9ba07>

- [24] van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* [Internet]. 2010 Aug 31;84(2):523-538. DOI: 10.1007/s11192-009-0146-3

IntechOpen

IntechOpen