

Trabajo Fin de Grado

Facultad de Ciencias Experimentales



Grado en Química

Revisión Bibliográfica: Aplicación de la difusión molecular por Resonancia Magnética Nuclear en la estimación de peso molecular

Bibliographic Revision: Application of NMR Molecular Diffusion in the Prediction of Molecular Weight

Ghizlane Baghdadi

Curso académico 2019-2020

Fecha 03/11/2020

Tutores
Prof. Dr. Ignacio Fernández de las Nieves
Dr. Francisco M. Arrabal Campos
Departamento de Química y Física

**Bibliographic Revision: Application of NMR Molecular Diffusion in the
Prediction of Molecular Weight**

Memoria del Trabajo Fin de Grado en Química presentada por
Ghizlane Baghdadi

Almería, 7 de Noviembre de 2020

Fdo.: Ghizlane Baghdadi

Fdo.: Prof. Dr. Ignacio Fernández de las Nieves

**FERNANDEZ
DE LAS NIEVES
IGNACIO -
44288591K**

Firmado digitalmente por
FERNANDEZ DE LAS NIEVES
IGNACIO - 44288591K
Nombre de reconocimiento (DN):
c=ES,
serialNumber=IDCES-44288591K,
givenName=IGNACIO,
sn=FERNANDEZ DE LAS NIEVES,
cn=FERNANDEZ DE LAS NIEVES
IGNACIO - 44288591K
Fecha: 2020.11.07 11:14:23 +01'00'

Fdo.: Dr. Francisco M. Arrabal Campos

**ARRABAL
CAMPOS
FRANCISCO
MANUEL -
75259467V**

Firmado digitalmente
por ARRABAL
CAMPOS FRANCISCO
MANUEL - 75259467V
Fecha: 2020.11.07
11:17:47 +01'00'

Life is a relationship between molecules

Linus Pauling

*Asegurémonos, pues, merced a una investigación bibliográfica cuidadosa,
de la originalidad del hecho o idea que deseamos exponer,
y guardémonos además de dar a luz prematuramente el fruto de la observación
Cuando nuestro pensamiento fluctúa todavía entre conclusiones diversas
y no tenemos plena conciencia de haber dado en el blanco,
ello es señal de haber abandonado demasiado pronto el laboratorio.
Conducta prudente será volver a él y esperar a que,
bajo el influjo de nuevas observaciones, acaben de cristalizar nuestras ideas.*

Santiago Ramón y Cajal “Reglas y consejos”

Agradecimientos

Este Trabajo Fin de Grado ha sido posible gracias a la colaboración del equipo de investigación del profesor Ignacio Fernández de las Nieves. Quisiera agradecer especialmente a Josefa Leticia López Martínez (Pepa), por haberme dedicado tiempo, paciencia y esfuerzo.

MEMORIA

ÍNDICE

1. ABSTRACT	3
2. RESUMEN	3
3. INTRODUCTION	5
4. OBJECTIVES	17
5. METHODS AND MATERIALS.....	17
5.1. Data collection applying automated scripts	18
5.2. Comparison between databases used within this TFG	19
5.3. Parse and refine of raw data	19
5.4. Data analysis and visualization	21
5.4.1. Gephi	21
5.4.2. VOSviewer.....	22
5.4.3. CitNetExplorer	22
6. RESULTS AND DISCUSSION	23
6.1. Classification of documents by type of documents	24
6.2. Classification of documents by subject area	24
6.3. Classification of documents by publisher	25
6.4. Classification of documents by authors	26
6.5. Classification of documents by country or territory	26
6.6. Correlations between authors and year of publication	27
6.7. Correlations between authors and citations.....	29
6.8. Classification of documents by index keywords	32
6.10. Limitations encountered during the performance of this TFG	36
7. CONCLUSIONS	36
8. REFERENCES.....	37
9. LIST OF ABBREVIATIONS	40
ANEXO	41

1. ABSTRACT

A bibliometric study has been conducted on the application of molecular diffusion using Nuclear Magnetic Resonance (NMR) in the prediction of molecular weight. Consolidated databases such as Scopus and WebOfScience have been used to afford this analysis. The use of programs such as Gephi, VOSviewer, OpenRefine, Table2Net and CitNetExplorer have allowed us to analyze the set of bibliographic data and obtain graphs and density maps based on co-authoring, connectivity, co-citation and co-citation of keywords. The analysis of data developed in this work has yielded specific scientific categories where all the documents have been grouped.

Keywords: *DOSY, PGSE, NMR, molecular difusión, molecular weight*

2. RESUMEN

Se ha realizado un estudio bibliométrico acerca del uso de la difusión molecular mediante Resonancia Magnética Nuclear (RMN) en la estimación de peso molecular. Para ello se han empleado bases de datos consolidadas como Scopus y WebOfScience. El uso de programas como Gephi, VOSviewer, OpenRefine, Table2Net y CitNetExplorer ha permitido analizar el set de datos bibliográfico y obtener grafos y mapas de densidad basados en coautoría, conectividades, co-citación y co-ocurrencia de palabras clave. El análisis de datos elaborado en este trabajo ha permitido confeccionar una agrupación de todos los documentos científicos en categorías específicas.

Palabras clave: *DOSY, PGSE, RMN, difusión molecular, peso molecular*

3. INTRODUCTION

The word bibliometrics was introduced by Pritchard¹ who replaced the previous term "statistical literature" that was used with the intention to measure bibliographic information from scientific publications.

Bibliometrics, as a scientific activity, refers to the nature and the way in which information is presented quantitatively. Its main body of study is based on the scientific publications and its main goal is to assess the scientific activity and impact of the manuscripts and journals in which they are published, determining on the latter their coverage and quality. Today, it is accepted in the scientific community as an analytical tool that helps to verify and explore the research represented of every single individual, or institution. Thus, it is focused on the study of the development, growth pattern, and spread of any discipline of research. In this sense, it can be viewed as a statistic approach of bibliographic computing that estimates and measures the expansion of a subject. This end-of-degree work will use the bibliometrics full potential to try to address the set of publications that encompass a specific scientific topic and attempt to classify them. Bibliometric analysis is based on a set of indicators that supply data on the investigation process, its extent, its development, transparency, and network structure. It includes not only descriptive statistics, but also network analysis on keywords, texts, quotations, authors, institutions, and their relationships. Elements such as frequency, relationship, centrality on documents, authors, institutions, and countries are investigated. Researchers use bibliometrics to explore publication trends, the knowledge body of a particular area, citation patterns, co-authoring networks, and the impact of scientific production. It is considered an emerging research discipline in the library and information sciences area.²

Bibliometric analysis allows, among other things, to:

- ∞ Know the growth and evolution of academic results related to a certain subject in a certain period of time.
- ∞ Identify changes and evolution of research interests in scientific production.
- ∞ Know the themes or fronts that drive the investigation of a certain discipline.
- ∞ Identify the main contributors in the production and impact of a certain topic.
- ∞ Measure the usefulness of research dissemination and communication services.
- ∞ Identify central publications of a certain discipline.
- ∞ Formulate procurement policies in research-oriented libraries and institutions.
- ∞ Study the dispersion and obsolescence of scientific literature.

In addition, bibliometric studies, allow an objective quantification of knowledge, help to summarize the research reported in the scientific literature, by measuring their indicators and are used in the current boom of information and knowledge through the consolidation of bibliographic databases. On these, the stored information can be evaluated, synthesized, and analyzed and can be understood as macro visions of research, as it generates multidimensional visions on various units of analysis.

Among its main tools are mathematical and statistical methods, advanced information retrieval routines, data and text mining methods and techniques for viewing and representing information that allow the user to analyze sets of documents.

grouped by region, by thematic, by authors, keywords, etc. The result of these analyses generates a very useful framework for decision-making in both research and management, assisting in resource management and documentation service planning.

Given that the selection of the source of information is key in these bibliometric studies, the SCOPUS database was used in this TFG, because of the advantages it offers over the other indexes and databases mainly in terms of coverage. Scopus owned by Elsevier and is one of the most known citation databases in the scientific community. As it assembles highly evolved techniques presenting an innovative technology with latest analytic tools to make the research easier for the authors, the librarians, and even for students. One of its advantages is that it continually updates the content including researchers and institution profiles, optimizing the link between the institutions, the authors as well as the topics. With more than 5000 publishers, Scopus indexes covers 24600 active serial titles and over 194000 books. It therefore presents 16 million authors and 70000 institution profiles. We may conclude that Scopus brings out requisite citation results, Scientifics profiles exerting ultimate metadata to make sure the result is satisfying and covering all researcher necessities.³ It includes titles from all geographic regions, even if the content is not in English with just the requirement that at least the summary should be in this language. About 22% of all indexed documents are non- English languages, adding a total of 40 languages.

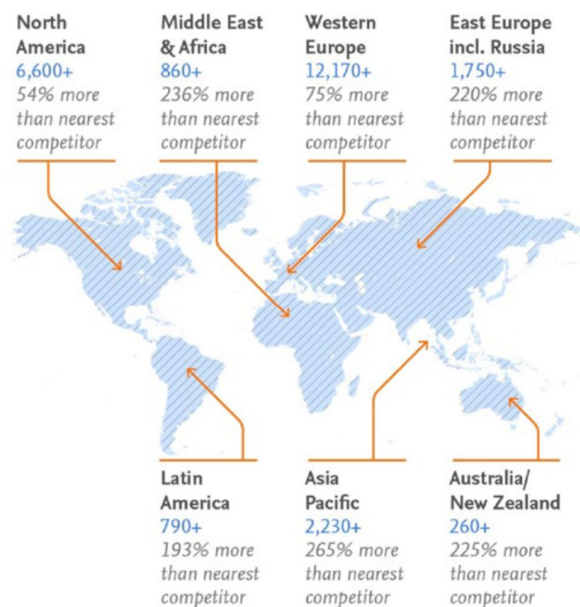


Figure 1. Geographical representation of the number of active titles indexed in Scopus.³

This coverage is not only regional, there is also a major advantage in covering the ranges of knowledge, where bibliographic content is available in four major areas: life sciences, physical sciences, health sciences and social sciences and humanities. These are then classified into 27 subareas of knowledge and then into over than 300 knowledge categories. **Figure 2** shows the distribution of the more than 25000 indexed titles in these four major areas.

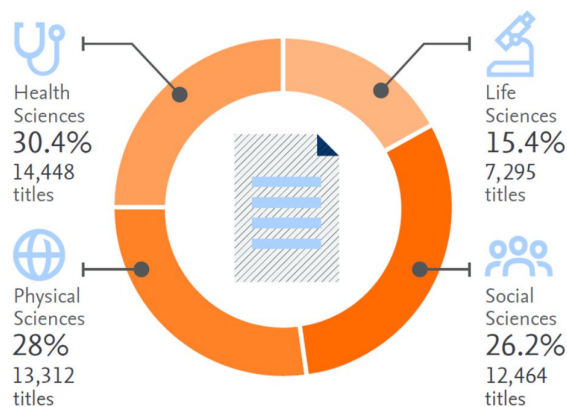


Figure 2. Active titles organized by main subject clusters. Note: A title can fall in more than one subject area.³

This advantage makes Scopus the ideal candidate when it comes to requiring a database that contains information to sustain a global analysis of knowledge domains. The use of this database as a source of information is at the core of this TFG. Therefore, in addition to a consolidated database, a solid and stable documentary basis on which to carry out our bibliometric analyses was required. This basis is based on a set of journals indexed in SCOPUS that mostly publish works related to molecular diffusion and represent a trusted communication channel among the scientific community. The current study is supported by the existence of this scientific communication that contribute to the progress of the topic to be studied, which is being reviewed by academic peer and deposited into databases of international recognition.

We have followed the next methodology.

Step 1. Determination of the descriptors. state a list over all the terminology that describes the theme of the term core "molecular diffusion." This is reached through the bibliometric analysis of articles include the term core. It begins with the definition of the type of information, in this case the primary literature is held to be an essential and fundamental reference of knowledge in the scientific world. As mentioned previously, the data resource was Scopus as the database that mostly indexes journals and conferences proceedings. The search results are refined by source type (journals), by primary literature (article and review), the language (English), then the time period for the analysis is selected (from the beginning until February 2020), and finally, a representative sample of the documents is selected where we perform the bibliometric analysis based on the co-occurrence of keywords, with the objective of establishing primary descriptors that are mostly present in articles, their relationships and relevance by means of different techniques of visualizations. As it is discussed further below, the Visualization of Similarities (VoS) application by Waltman *et al.*⁴, as shown by Cobo *et al.*,⁵ provide a very accurate look at how a document corpus is described and linked. Based on the set of primary descriptors, new descriptors are included as result of linguistic similarities or acronyms or abbreviations used in natural language, for example, PGSE or DOSY. These new descriptors, that reflect the same meaning as the one provided by the author, are called Secondary Descriptors.

Step 2. Correspondence of publications and descriptors. Build a matrix of articles volume for each descriptor (primary and secondary) and each publication indexed in the database. Using the same selection criteria described in the previous step, a query is made to the database for each of the descriptors that have appeared thus determining

the number of articles of each descriptor. Finally, the primary and secondary descriptors of each term are added, assuming that the sum reflects unique works related by descriptors.

Step 3. Analysis of the set of publications. The selected journals are analyzed under a bibliometric view to determine if they show a distinctive scientific discipline that can be delimited as a cross thematic category. The base map will be a global map of science that includes the total of journals indexed in Scopus. The relationship degree of publications is established by the normalized value produced by the combination of cites, co-cites and coupling and finally, this analysis is enriched with the clustering performed by for instance VOSviewer, CitNetExplorer or Gephi (see below). The local map that will be overlaid on the global map of science is the set of journals and conference proceedings selected in the previous step.

In summary, the methodology followed in this TFG is derived on the principle that an important presence of field-specific descriptors in the items of an article is directly proportional to the number of interactions by citation, co-citation, and coupling of a publication with others that would form part of the discipline cluster. There are multiple methods and tools for visualizing bibliometric networks, such as distance-based, graph-based or time-based.⁶ Mapping and clustering are also used to respond to concerns about the main fields of research in a scientific domain. As a tool, Gephi and VOSviewer assure the comprehensive visualization of node labels on the map.

The global science map that we have employed is the one constructed using SCImago,⁷ guarantees to normalize values for each of the publications to be visualized. As can be seen in **Figures 3** and **4**, the map is composed of seven clusters, which in a clockwise and wide sense can be denominated as: social sciences (green), psychology (orange), medicine (clear cyan), health professions (green), life sciences (yellow), physical sciences and engineering (dark cyan) and computer science (blue). **Figure 3** illustrates journals included in Q1 to Q4 quartiles, whereas **Figure 4** is just focused on journals published in Q1.

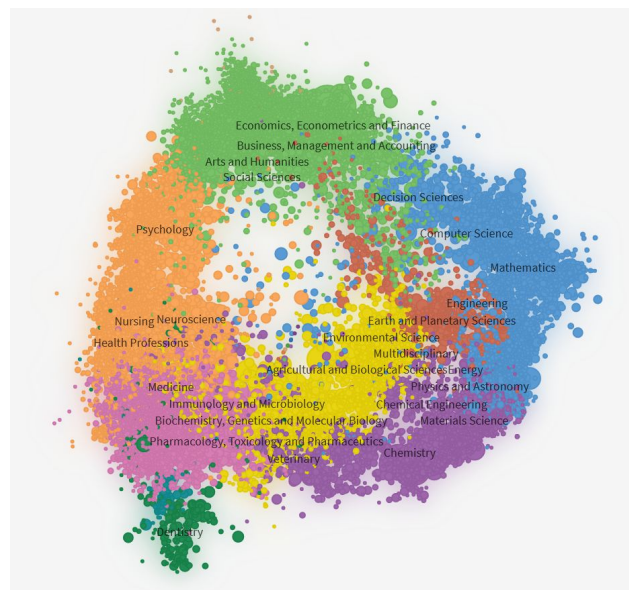


Figure 3. Global map of science based on SCImago Journal & Country Rank including Q1 to Q4 journals. The Subject Areas were selected as labels, and the values from the last SJR available (2016) as the size of the nodes.



Figure 4. Global map of science based on SCImago Journal & Country Rank including in Q1 journals. The Subject Areas were selected as labels, and the values from the last SJR available (2016) as the size of the nodes.

Figures 5 and 6 present the overlap of the previous maps with the journals included in the areas of Chemistry and in the subject category of Organic chemistry, respectively. The color indicates the area of knowledge in which the publication is superimposed, and its size corresponds to the percentage of participation within the SJR.

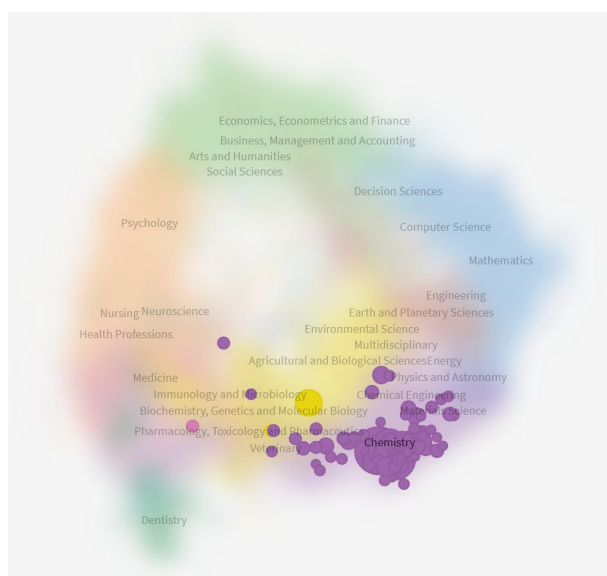


Figure 5. Overlay of the journals of the Area of Chemistry with the Q1 SJR-2016 as node size.

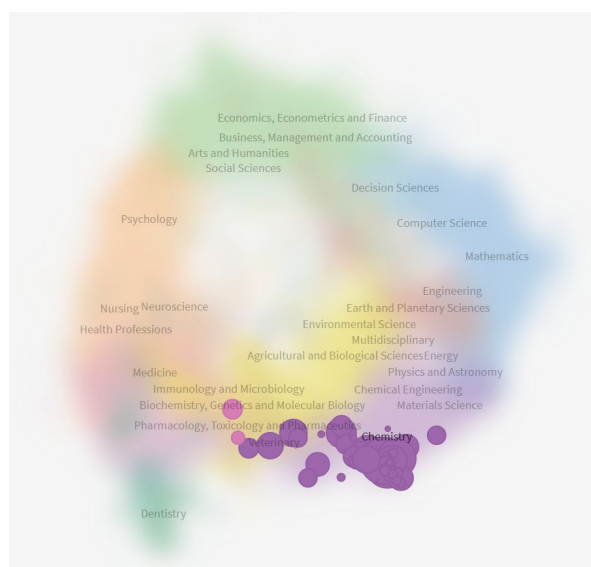


Figure 6. Overlay of the journals of the Subject category of Organic Chemistry with the Q1-Q4 *SJR-2016* as node size.

To refine the information obtained from Scopus the software OpenRefine is frequently used. It was formerly called Freebase Gridworks and later Google Refine. It is an open access software projected to operate files with massive data, allowing the user to refine the disordered data by cleaning it, converting it to different format, also stretching the data with web services.⁸ One of its main advantages is the privacy, as it remains the data confidential on the computer keeping the data in the same operator unless the user decides to share its work. An OpenRefine project is based on a table where the user can filter the rows to display them using facets that define filter criteria. Actions carried out in an OpenRefine project on one data set can be stored and reproduced in a different data set. These actions are carried out on the visible rows of the table, so actions such as designing a new list taking in consideration the data from other columns, converting every cell in a row into a single column, etc. are all possible. Transformations are performed only once, and the formulas used for these transformations are not stored in cells or in spreadsheets.

Gephi has a cooperative approach to the development of sophisticated systems and networks, by creating an interactive and dynamic visualized graph, endowed with a three-dimensional render engine to display real-time evolving network. It was originated by students at the UTC University in France and gives the opportunity for the users to study and distribute the software by adding free tools to the graphs, and therefore creating an open collaboration. In Gephi, modularity class, which is a measure of the presence of distinct communities within the network, is calculated using the Louvain Method, an iterative algorithm that builds a “community of communities” in part by assessing the gain or loss in modularity by moving individual nodes between communities and producing a particular number of modularity classes that best represents the organization of the network.⁹ Graph data might include information about any type of network or complex structure where no programming skills are required.¹⁰ Importantly, the Gephi program allowed us to use columns to bring together the nodes and edges by building a network from a table,¹¹ and supports graph data in a variety of formats, including comma delimited (CSV). Graph files include three types of data: nodes, edges, and attributes. Nodes are individual data

points, attributes describe nodes, and edges connect nodes. Gephi can be obtained from <https://gephi.org/> and can be used freely for any purpose.

VOSviewer visualization framework, is a free access software tool that aim to transform the bibliometric data to a more creative visualizable graphs and networks. VOSviewer is usually used to develop examining maps due to the density visualization, overlay visualization, and network visualization, showing more details for a better analysis of the data by zooming and scrolling. This software tools build its network by creating its own file using bibliographic database obtained from Scopus, PubMed, or WebOfScience. Thus, it allowed us to create a map that help on visualizing authors, countries, authors, keywords, or years of publication.

The principle functions of VOSviewer may be gathered as below:

- ∞ Creating maps based on network data. Maps can be created based directly on the adjacency matrix of a network, but it is also possible to create maps of scientific publications, scientific journals, researchers, research organizations, countries, or keywords based on co-authorship, co-occurrence, citation, bibliographic coupling, or co-citation networks extracted from WebOfScience, Scopus, PubMed, or RIS files. Term maps can be created directly based on a text corpus. Maps are created using the VOS layout and clustering tools, by this way provide a unified framework.⁴
- ∞ Visualizing and exploring maps. Where zooming and scrolling functionality empowers maps to be explored in full detail, being a fundamental detail when working with large maps containing hundreds or even thousands of items, providing this way the user to choose between visualizing by density or network visualization.

VOSviewer can be attained from www.vosviewer.com and can be operated freely for any aspiration. More information, including a step-by-step tutorial, can also be found in a more recent book chapter by Van Eck et al.¹² As in Gephi, the VOSviewer software allows for the use of different colors to indicate clusters of objects. An interestingly, it allows to delete or merge terms that may be closely related to term cluster denoted by the same cluster color. According to the user manual, the proximity of the terms can be interpreted as an indication of their relatedness.



Figure 6. Main software packages used within this TFG.

Finally, CitNetExplorer is a new software tool destined to analyze and visualize direct citation networks. CitNetExplorer, being an abbreviation of ‘citation network explorer’, can handle large citation networks. Further, CitNetExplorer afford practical functionality for drilling down into a citation network, by allowing users to start at the level of a full network consisting of several millions of publications and to then

gradually drill down into this network until a small subnetwork has been reached including no more than, for instance 100 publications, all dealing with a specific topic of interest. In fact, one can see that CitNetExplorer shares several concepts from the VOSviewer tool. This applies in particular to certain features related to visualization (e.g., smart labeling) and user interaction (e.g., zooming and panning).¹³ CitNetExplorer can be downloaded from www.citnetexplorer.nl and can be used freely for any purpose.

By involving the above-mentioned software tools in this TFG, we present herein the results of a bibliometric study based on literature screened from 1989 until 2020 over the topic *Applications of NMR Molecular Diffusion in Molecular Weight Prediction*. Our dataset, as it is described further below, is based on 546 article references.

Since NMR spectroscopy is therefore the heart of this TFG, it is relevant to illustrate in some paragraphs the functioning of this method for a better understanding of the topic.

NMR, UV-Visible and IR spectroscopies together with MS spectrometry are probably the most used analytical tools that allow chemists to obtain structural insights.¹⁴ In terms of NMR, thanks to the advances that have been produced in recent decades, development of superconducting magnets and Fourier's mathematical methods in the 1970s, experiments based on multi-pulse sequences and multidimensional experiments in the 80s, magnetic field gradients in the 1990s, and cryoprobes a decade later, NMR has become a powerful technique from the viewpoints of structure characterization and quantification. Probably the application to structural determination is the best-known aspect of NMR in the area of organic chemistry. However, other applications focused on dynamics and interactions are also important due to the information at the atomic level they can provide and for the functional, mechanistic or design implications that can be obtained from their results. This is relevant since the biological function of a given drug, for example, depends greatly on the interactions it can establish with other macromolecules such as proteins, nucleic acids, small ligands or lipids. But in addition, these interaction processes are clearly determined by the mobility and dynamics of the molecules involved.¹⁵ This TFG addresses one of the areas in which high-resolution NMR has been applied in the last decades: the study of the phenomenon of molecular translational diffusion in solution with special emphasis on its application to molecular weight prediction.

The idea of determining molecular translational diffusion by NMR is in principle very intuitive. When applying a diffusion experiment, signals belonging to smaller molecules attenuate faster than those belonging to larger molecules, because they move faster in solution. The basis for the resolution of the diffusion coefficients (D) is the so-called spin-echo sequence, known as Pulsed Field Gradient Spin-Echo or PGSE, which was proposed by Stejskal and Tanner in 1965. This sequence is one of the most important in NMR and is the basis of the well-known application of the image NMR (MRI) technique applied daily in the clinical and biomedical world. In the early 90s, thanks to the advances in the design of high-resolution NMR probeheads incorporating stable magnetic field gradients, the study of molecular diffusion properties became better known and widely used in the chemical and biochemical world. Magnetic field gradients represent a very effective alternative to conventional phase cycles used for the selection of coherence transfer paths and are the basis of almost all the NMR sequences routinely employed nowadays.

The concept of diffusion is used to describe many different physical processes. In NMR, diffusion may refer to spin diffusion, rotational diffusion or translational diffusion. In this TFG, using the diffusion term will only refer to molecular diffusion, a form of translational diffusion due to the Brownian movement of dissolved molecules that occurs in absence of any external force (such as concentration or electric field gradients). In the movement associated to molecular diffusion, the translation of molecules is due only to their kinetic energy. As they move, the molecules collide with each other and change the direction of movement. Over time, the movement of a molecule follows a random path that results in a property known as the diffusion coefficient, D . In other words, the diffusion coefficient describes the distance a molecule moves in a specific medium over a specific time interval. Therefore, the diffusion coefficient is individually defined for each molecule in a given solvent and at a specific temperature.

The D -value therefore depends on three factors: i) size and shape of the solute, ii) temperature, and iii) viscosity of the solvent. Increasing the size of the solute or solvent viscosity makes it difficult to diffuse, while increasing the temperature accelerates it. The units of the D -coefficients are $\text{m}^2 \text{s}^{-1}$. For the same solvent, the higher the molecular mass the lower the D value. For the same compound, the D value decreases by increasing the viscosity of the medium. In some cases, non-expected values may be due to evidence of existing different solvation forms, effects of ion-pairing, different aggregation states, i.e. monomer, dimer, tetramer, etc., existence of hydrogen bonding between species, etc.

Today the measure of D -coefficients is considered a powerful tool in NMR and a good alternative to the classical strategies and experiments normally used for structural characterization and dynamic effects.

In an ideal gas, a molecule will move at an average rate dependent on its mass and temperature. In a liquid, the molecule will move at the same average speed, but the distance traveled will be restricted by the molecules of the solvent around it. This constraint, represented by the frictional force F_f , is proportional to:

$$F = -fv \quad [1]$$

where f is the friction coefficient and v the rate. For a spherically shaped object moving in a viscosity liquid, the Stokes' law predicts that the frictional force of a sphere complies with:

$$F = -6\pi\eta rv \quad [2]$$

where r is the radius of the sphere. In this case, the coefficient of friction is defined by:

$$f = 6\pi\eta r \quad [3]$$

Considering the relationship between a molecule's kinetic energy and its friction, the D value can be obtained from the Stokes-Einstein equation:

$$D = k_B T/f = k_B T / 6\pi\eta r \quad [4]$$

where k_B is the Boltzmann constant ($1.38066 \times 10^{-23} \text{ kg m}^2 \text{ K}^{-1} \text{ s}^{-2}$), T the temperature (in K) and the viscosity of the solution (in $\text{N s m}^{-2} \text{ s}^{-1} \text{ m}^{-1}$). This equation was originally developed for colloidal spherical particles, where the Brownian movement is well defined. However, it is also an approximation for small (non-spherical) particles. The use of the Stokes-Einstein equation to predict diffusion can be improved by replacing the radius of the molecule, r , by its effective radius, r_{eff} , also called hydrodynamic

radio, r_h . The r_h value can be used to compensate for factors such as non-spherical molecule shapes and solvation spheres.

The interest in determining the rate of translation at which a molecule diffuses is because this movement is related to important molecular properties. For example, D -values can be used to predict the molecular mass (M) of a compound where its application can be found in various areas such as statistical distribution of molecular masses in polymers,¹⁶ characterization of additives in polymers,¹⁷ analysis of complex mixtures of hydrocarbons¹⁸ or in inorganic or coordination chemistry.¹⁹ This is possible because M depends, in a first approximation of:

$$M = (k_B T / 6\pi\eta F_P D)^3 \quad [5]$$

where F_P is the so-called form factor or Perrin factor that depends on the friction coefficient:

$$F_P = f/f_0 \quad [6]$$

where f and f_0 is the friction coefficient of the molecule and sphere, respectively.

The molecular mass of a compound can be estimated using calibration curves. As an example, **Figure 7** shows the dependence of D based on the reciprocal of the cubic root of M for a battery of organic compounds. As it is observed, the trend is that the higher the molecular mass, the lower the value of the diffusion coefficient. The observed deviations are due to the different shapes of the molecules, i.e. the more similar the compounds are in terms of structure and shape, less the deviation is observed. The implementation of calibration curves for estimating M has been used for example in proteins,²⁰ polymers²¹ and complex mixtures of agri-food origin.²²

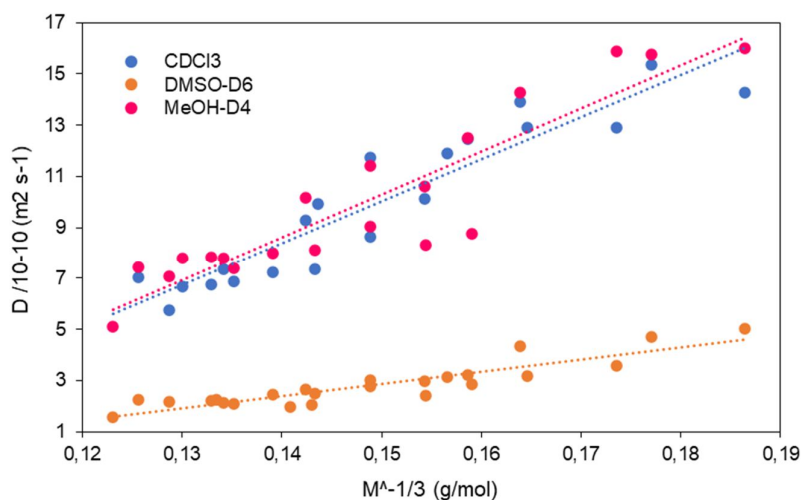


Figure 7. Linear relationship between the diffusion coefficient and the reciprocal of the cubic root of the molecular mass of several organic compounds in three different solvents. The D -values were extracted from reference 23.

The rate at which a molecule diffuses can also be used to study molecular interactions. The concept is defined on the principle of the D value of one molecule is adjusted after the insertion of another molecule whenever there is an interaction among both sides. This approach can be used both qualitatively, to identify compounds that bind to a specific receptor in NMR screening or in ligand-host interactions, and

quantitatively, to determine association constants in ion-pairing or aggregate formation.¹⁹

In terms of the pulse sequences used in diffusion NMR, a brief description will be given. The spin echo pulse sequence (PGSE) is the basic experiment used to determine D-values. In theory, although the experiment is able of measuring both rotational and translational diffusion, it is often used to measure the latter. In the bibliography you can find excellent review articles with fundamentals, theoretical bases and applications.²⁴

For the experimental determination, a series of ¹H NMR spectra is acquired using the PGSE-based sequence where the force of the magnetic field gradient (G_z) is progressively increased and the attenuation of the intensity of the signals in each spectrum is analyzed. The relationship between the observed signal strength (A) obtained when using gradients is a decreasing Gaussian function described by the following equation:

$$E_{diff} = e^{-D\gamma_{eff}^2\delta^2\sigma^2G^2\Delta'} \quad [7]$$

Where D is the translational diffusion coefficient of the molecule to which the monitored signal belongs, γ_{eff} represents a linear combination of the gyromagnetic ratios of the nuclei studied depending on the coherence transfer pathway, δ is the PFG duration, and σ is the gradient shape factor. The diffusion delay Δ is the time between the two PFG pulses in which the molecular diffusion can induce its effect, while Δ' is this same delay corrected by an amount that depends on the specific pulse sequence and gradient shape used.²⁵

Usually, the most convenient parameter to vary is the strength of the G_z gradient, since the change in the intensity of the signals is only due to the diffusion process. But if we analyze the equation, the diffusion time, Δ , could be also varied instead of G_z. In addition, the duration of the gradient, δ , could also be varied instead of G_z, but in this case the time when magnetization is in the transverse xy plane should be kept constant throughout the series of experiments, otherwise the signal strength would be affected by both diffusion and cross relaxation.

Although all experiments that use gradients for coherence selection are sensitive to diffusion, some are better than others for getting a correct measure of D-values. Different pulse sequences for D determination have been described in the literature,²⁴ all of which are always trying to improve the accuracy of measurements. The most common ones used in the works included in the bibliometric analysis of this TFG are shown in **Table 1**.

The eddy currents mentioned in some of the sequences provided in **Table 1**, are induced currents in the magnetic components of the probe, which come from a coupling between the gradient coil and the main magnetic field. This interaction causes lock frequency distortions that damage the shape of the signals detected in such experiments. As signal distortion increases as the amplitude and duration of the applied gradient increases, the attenuation in the signal height is faster than expected by the molecules' own Brownian diffusion, leading to erroneous D estimates. The incorporation of increasingly shielded field gradients into high-resolution probe heads has significantly reduced these distortions. However, even with the best technology, the distortion in the signals resulting from applying field gradients in the basic sequence STE substantially deteriorates the result of the experiments. That is why

specific sequences have been developed capable of minimizing these effects during the acquisition. The two most relevant options have been the incorporation of a polarization storage delay, usually called T_e , at the end of the sequence (LED in **Table 1**) that allows these currents attenuate their self, and the use of bipolar pair pulses (BPP in **Table 1**).

Table 1. Secuencias de pulsos para experimentos de difusión.

Acronym	Name of the experiment	Bruker sequence	Reference
SE	Pulsed field gradient spin-echo	zggpse	26
STE	Stimulated spin-echo	stegp1s1d	27, 28
BPPSTE	Stimulated spin-echo with bipolar pair pulses	stebpgpls1d	29
STE-LED	Stimulated spin-echo with longitudinal eddy current delay	ledgp2s1d	30
BPPSTE-LED	Stimulated spin-echo with bipolar pair pulses and longitudinal eddy current delay	ledbpgp2s1d	29
DSTE	Double stimulated spin-echo	dstegp3s1d	31

Convection is one of the most important problems affecting diffusion measurements. These convection currents originate when there is a temperature gradient within the sample. Temperature is therefore key in conducting diffusion experiments. Poor control of it involves the creation of temperature gradients along the NMR tube which, based on the viscosity of the solvent used and the temperature range, may generate convection movements such as the one shown in **Figure 8**.

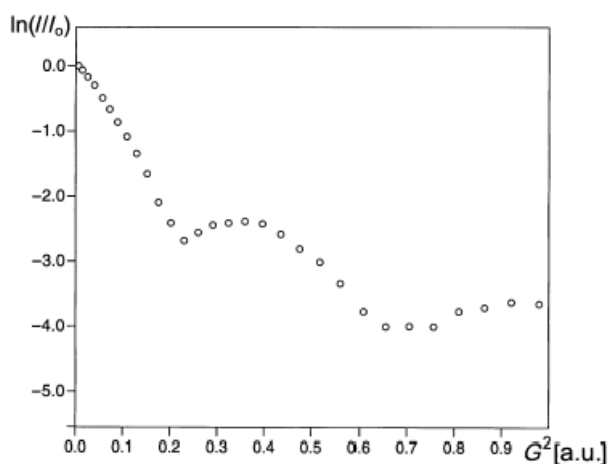


Figure 8. Problems associated with convection. ^1H PGSE ($\delta = 2$ ms, $\Delta = 68$ ms).³²

The effect is the same as that of diffusion, the molecules are moved in addition to their own Brownian movement, by the effect of convection currents, which therefore leads to erroneous values of D . A similar effect is observed when using low viscosity

solvents, i.e. chloroform, and the temperature used is close to its boiling point. Without due precautions, the results obtained under these circumstances are like those shown in **Figure 8**. DSTE sequences (**Table 1**) suppress convection effects. This so-called double-stimulated approach is based on duplicating the STE stimulated sequence, either with LED period, with bipolar pulses or both at the same time.³³

In this TFG the term DOSY, that is a format in which the results are represented in a pseudo-2D format with the conventional proton spectrum on the x-axis and the value of the diffusion coefficient on the y-axis, has also been included in the searches. In this sense, the final goal of the current work is to classify our dataset of articles in topics, categories, or scientific terms by using the different bibliometric tools described above. This classification would be possible by defining clusters and analyzing the relationship between authors, countries, institutions, year of publication, number of citations, publishers, subject area, funding sponsors, etc. It is important to mention that in this TFG only research articles and reviews will be considered, and in the searches, we will refine the resulting datasets by removing book chapters, conferences papers, proceedings, and undefined document types.

4. OBJECTIVES

The objectives of the current TFG are divided in the following bullets:

- ∞ Use of bibliometric tools for the evaluation of the topics included in the analyzed dataset of references.
- ∞ The use of different subject area referring to the analyzed topics.
- ∞ Analysis of the area of knowledge by using established databases such as Scopus and WebOfScience.
- ∞ Revision of the accessibility of the cited articles and reviews in the employed databases.
- ∞ The development of graphs and density maps based on co-authorship, bibliographic coupling, keyword co-occurrence and co-citation.
- ∞ Analysis of the whole dataset based on the geographical distributions of their authors.
- ∞ And finally, a classification of the different reviews and articles in scientific categories by using the index co-occurrence keywords.

5. METHODS AND MATERIALS

In this TFG, the methodology is focused in the analysis of the totality of the articles and reviews related to the selected keywords, by grouping the articles in several fields, identify scientific collaboration networks by citation or co-citation, gather the articles by years, country of origin, keywords, subject area, as well as funding sponsors. In addition, for a better understanding of the topics, we will visualize bibliometric networks and statistics based on diagrams with the help of key software tools. For this purpose, two main databases are considered: Scopus and WebOfScience. Then, Gephi, CitNetExplorer and VOSviewer are employed in order to help on the visualization of the different fields-sections, making an easy discernment of different cluster over each graph.

5.1. Data collection applying automated scripts

The flow digram pictorialized in **Figure 9** illustrates a step by step representation of the sequences involved in the process of extraction of information by Scopus using the Research Network Bot (ResNetBOT). The performance of this specific bot, restores data for divergent analyses, and it is possible to be divided into three parts:^{34,35}

1. Getting data on paperwork including ‘DOSY and PGSE NMR’ in keywords or in the title. It is important to mention that the use of keywords in scientific documents is by far the most frequent research topics in a variety of fields of organic chemistry. The information is usually visualized graphically, where the size of the nodes is proportional to the h index and the lines between connected nodes point to the citations.
2. Collecting information from authors of the documents cited in step (1), by covering all data about authors in the Scopus database, such as: author identification number, affiliations, publications and dates, number of citations, h-index, among other items.
3. Processed to acquire the collaborative network of authors. The information stored refers to the number of collaborations between co-authors and their affiliations, country and city. The final step would be visualizing the information collected in a graph where the size of the nodes will indicate the scientific production according to the country and will represent how strong the collaboration is with one another. Thus, ResNetBOT seek for the iterative method for any authors that appears in the search, looking for information that make the list understandable from the name of the institution with whom the author works, the current affiliation, city, country, till the number of author-co-author collaborations.

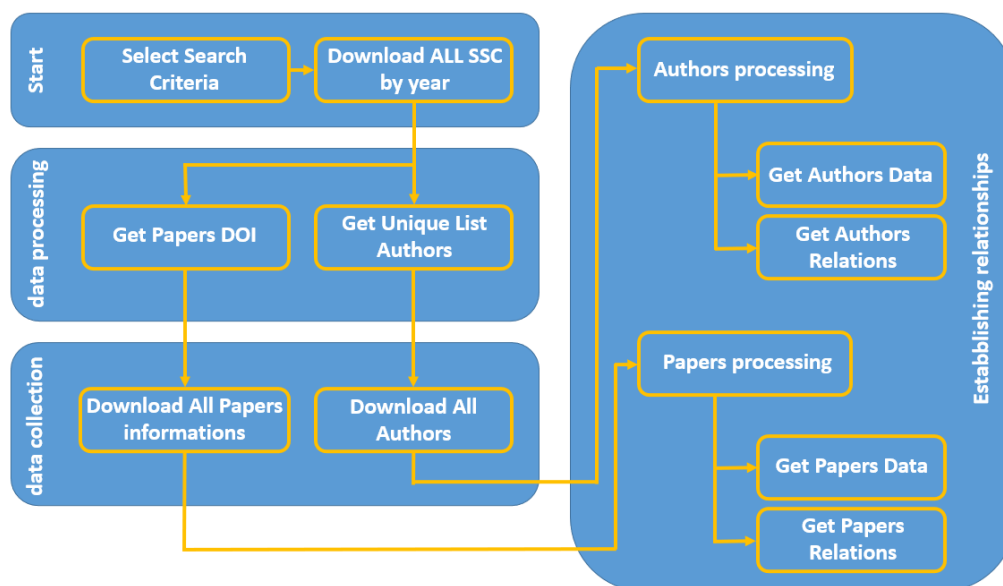


Figure 9. Representation of the sequence of steps involved in the process of extraction of information for ResNetBot.³⁴

5.2. Comparison between databases used within this TFG

We decided to extract the set of references from two of the most powerful databases that exist in the market: Scopus and WebOfScience. In the present section, we present the advantages and drawbacks for each of them, which are summarized in Table 2.

- ∞ Scopus was launched in 2004 by Elsevier in the Netherlands and includes 12850 including open access publications with a total of 30 languages, whereas WebOfScience was launched as well in 2004 by Thomson Scientific and Health Care Corporation in the US and enjoys a total of 8700 publications with more than 45 languages.
- ∞ Both databases focus their attention in the scientific yard, as well as a private access database requiring a payment account.
- ∞ Scopus handle a vast publication range, collecting keywords searching and citation analysis of publications since the 1966. Articles containing a citation analysis are only those published after 1995, whereas in the case of WebOfScience these articles are registered since 1900.
- ∞ About the actualization of both databases, Scopus update its searching results once or twice per week whereas WebOfScience does it only once per week.
- ∞ Scopus covers around 20% more coverage than WebOfScience. However, considering the visualization aspect, WebOfScience creates more understanding graphs with probably more completed details.
- ∞ Using the same character of research, Scopus is the database presenting advanced results as it offers an extended number of publications in comparison with WebOfScience. In addition, WebOfScience does not extend data to open access publications.
- ∞ The two databases make room for published articles as both of them are regularly updating for printed articles but not the case for online early ones.

5.3. Parse and refine of raw data

Scopus is a great source to collect the largest part of information in relation to our topic. We requested the bot to retrieve all details that the program allows. The output file usually contains discrepancies, as it is a common mistake for the majority of databases with a large amount of information. The most frequent problems are wrong authorID, multiple authorIDs, and wrong affiliationID.³⁴

Due to the inconsistencies in the information collected, we decided to use a specialized tool in refining data. OpenRefine software is an accessible tool for any scientific researcher as it provides some great algorithms. It can be obtained from www.openrefine.org. We have used the version OpenRefine 3.3, selecting the Mac Kit. First, the files are uploaded to the platform, which supports extensions TSV, CSV, *SV, Excel (.xls, .xlsx), JSON, XML, RDF like XML and Google Data. In our case, we uploaded the file with the CSV extension downloaded from Scopus (Figure A1). Then, we created the project where we can select the way in which the program analyzes the data in the CSV file (Figure A2). We select the text facet algorithm for refining. When build from a column, the text facet assemble matching cells into that column through rows and exhibit the number of rows in each group. For instance, we

can select a row with text filter (Figure A3), it filters the data table on the other side (Figure A4), showing only the rows with that selected word (e.g.: ACS).

Table 2. Comparison of the characteristics of Scopus and WebOfScience database.³⁶

Characteristic	Scopus	WebOfScience
Official inauguration	2004	2004
No. of journals	12.850 (500 open access)	8.700
Languages	English (plus 30 other languages)	English (plus 45 other languages)
Focus (field)	Physical sciences, health sciences, life sciences, social sciences	Science, technology, social sciences, arts and humanities
Period covered	1966-present	1900-present
Databases covered	100% Medline, Embase, Compendex, World textile index, Fluidex, Geobase, Biobase	Science citation index expanded, social sciences citation index, arts and humanities citation index, index chemistry, current chemical reactions
Keywords allowed	30	15
Search		
Abstracts/ Authors/ Citations/Patents	(+)	(+)
Uses	Links of full-text articles and other library resources	Links to full-text, links to related articles
Updating	Once to twice by week	Weekly
Developer/owner (country)	Elsevier (Netherlands)	Thomson Scientific and Health Care Corporation (US)
Citation analysis	Total number of articles on a topic or by an individual author cited in other articles	Total number of articles on a topic or by an individual author cited in other articles

When words or sentences refer to the same concept, but they are written slightly different, such as “molecular weight”, “Molecular weight” or “Molecular Weight”, the program replaces all of them by just one, so then, only one term is found along all the set of references (Figure A5 and A6). It uses algorithms like “key collision methods”

as well as “nearest neighbor methods”. The last step would be exporting the refined document (Figure A7). The resulting file will be then processed by other software’s to generate final graphs or correlative figures. **Figure 10** shows how the OpenRefine software applied on the imported Scopus file worked for us, and in where one could find the type of files that the tool can import, can export, together with some other services employed by the software.

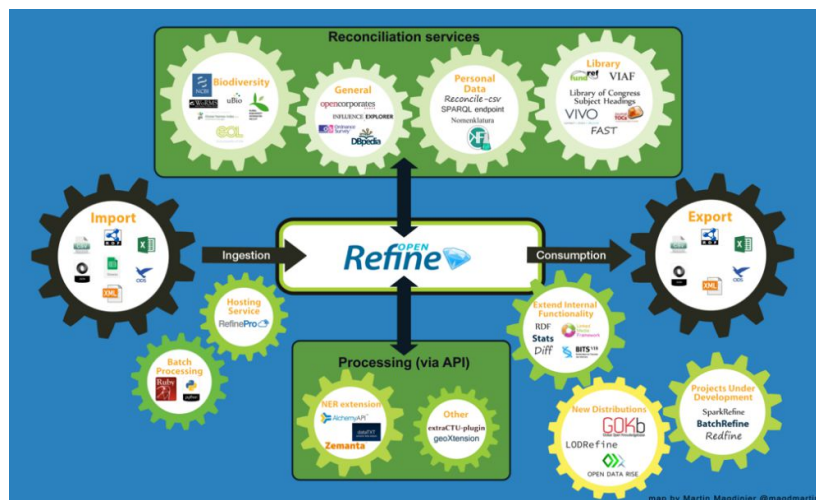


Figure 10. A visualized graph describing the functions of the software OpenRefine.³⁷

5.4. Data analysis and visualization

To visually analyze the information already polished with OpenRefine, the tools employed in this TFG were Gephi, VOSviewer and CitNetExplorer.

5.4.1. Gephi

The software Gephi can be obtained from www.gephi.org. We have used the version 0.9.2. We then downloaded the .dmg file for Mac and drag the Gephi application to the applications folder. The platform supports extensions GEXF, GDF, GML, GraphML, Pajek NET, GraphViz DOT, CSV, UCINET DL, Tulip TPL, Netdraw VNA, Spreadsheet.

In our case, we execute the tool Table2Net³⁸ (Figure A8) to generate GEXF file from the CSV file obtained from OpenRefine. After uploading the CSV file, the tool gives the option to select the Network type. We select single type of nodes linked by common values in another columns, by doing that, “Normal” Network is chosen (Figure A9). Followed by deciding which column defines the nodes, nodes attributes, which column defines the links, links attributes and if desired, additional settings like time series or edge weight (Figure A10). If there are multiple items per cell, it should be specified the separator. In our case, the use of those characteristics was dependent mainly on the type of graph needed. As an example, we have defined authors as nodes and years as links. Several authors appear in the same cell separated by commas; therefore, we specified the separator for the program to identify each author as a node. However, for the years it is not necessary because only one appears per cell. In the last step, the tool allows to build and download the GEXF file (Figure A10).

Once the GEXF file is created, we returned to the software Gephi and already in there, we click on "File" and then "Open". A window appears where we select an "Undirected" type of graph in a new workspace (Figure A11). To obtain the desired results, the program promotes a great variety of algorithms and tools (Figure A11) giving the user the chance to engender some unique graphs by performing and experimenting colors, adjusting the size or shape of edges and nodes, making three-dimensional graphs, as well as using statistical metrics to interpret proportional and topological characteristic nodes, etc. It also allows you to work viewing the data in the "Data Laboratory" tab (Figure A12) and preview the final result of the graph (Figure A13). To export the final graph, there are various archive types to select from SVG, PDF, PNG. Another easy way to save the graph is by using the option screenshot, as Gephi allows this option by clicking on the camera icon.

5.4.2. VOSviewer

For more specific graphs, we explored VOSviewer, once the version 1.6.13 for Mac is available from www.VOSviewer.com.

An advantage of this alternative tool is that can bring attractive graphs, most notably when it combines and connect authors, countries, keywords, most cited authors, etc. In contrast to Gephi, VOSviewer does not need any particular type of file, as the CSV file generated in Scopus is perfectly read by the system. By that, we simply click on "Create ", then "create a map based on bibliographic data" (Figure A14), we choose to read the data from bibliographic database files since Scopus appears in that category (Figure A15). Then, we selected the file downloaded from Scopus and we click on "Next" (Figure A16). In the following step, we select the type of analysis and the counting method needed for that specific graph (Figure A17), introduce the minimum number of documents and citations of the nodes (Figure A18), and the final design of our graph appears in the platform by clicking at "Finish" (Figure A19).

For the visualization of the graph, VOSviewer offer to the user three possibilities: "Network Visualization", "Overlay Visualization", and "Density Visualization", each of one developed for a specific type of graph (Figure A20a-c.). In "Items", the software groups the clusters depending on the relationship they build by co-citations or link of strength (Figures A21). In "Analysis", it presents interesting tools giving the opportunity to make the graph as a unique piece. Those tools are the method of normalization, advance parameters for layout as well as for clustering (Figures A22). Other buttons for rotating the graph, scale of visualization, labels and lines are considered important tools for their ability to show clearly the words (authors, countries, keywords). Coloring is also provided which are important in order to detect easily all the established clusters of our graph. To export the created graph (Figure A23), we have two options. We can Screenshot the graph like in Gephi, or we may save it as different type of file (PNG, BMP, EMF, EPS, GIF, JPG, PDF, SWF, SVG, TIFF, PDF).

5.4.3. CitNetExplorer

CitNetExplorer is the third software used in our bibliometric study for visualizing the scientific network. Not only VOSviewer but also CitNetExplorer is developed by Dr. Nees Jan van Eck at Leiden University. The version 1.0.0 for other systems like Mac is downloaded from www.citnetexplorer.nl. When we launch the program, a new window is directly open (Figure A21) in which we should select only files created

earlier in the WebOfScience database, due to the incapacity of this program to read files from any other databases such as Scopus. When you download the file from WebOfScience, it is important to save it selecting the “delimitado por tabulador (win)” option, otherwise the CitNetExplorer software does not read the file properly and it won't be run. Once chosen the appropriate and selected the minimum number of citations (Figure A21), the graph is created in "Citation network". We may observe that in the timeline it appears the years of publications, and in the horizontal access is represented the highly cited authors in that year (Figure A22).

We then start to modify the graph by using algorithms to define and colored the obtained clusters in the "analysis, clustering" section (Figure A23, A24). To generate the final graph, we export it by using the same method of saving applied in VOSviewer, as both softwares are developed in the same way.

6. RESULTS AND DISCUSSION

The bibliometric analysis that is pursued in this TFG is based on the applications of NMR molecular diffusion in the prediction of molecular weights. All the visualized graphs, tables of data, are created by Scopus, Gephi, VOSviewer and CitNetExplorer.

For this purpose, we have combined some of the operators to create a dataset with the maximum number of documents. The two operators used in the search were: (i) AND: is used for a result that covers all terms and appellations that perhaps far removed; (ii) OR: we use this operator when the research requires the use of one or multiple terms e.g. synonyms, abbreviation or alternate spellings. As a result, any document in the database that includes whichever of the words will be included.

The following figure illustrates how those two operators have been used in one of our searches.

Document search Compare sources >

Documents
 Authors
 Affiliations
 Advanced Search tips ⓘ

Search Article title, Abstract, Keywords + -

dosy or pgse and nmr and molecular and weight or diffusion x

E.g., "Cognitive architectures" AND robots

> Limit
Reset form Search Q

3 TITLE-ABS-KEY (dosy OR pgse AND nmr AND molecular AND weight OR diffusion) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "re")) AND (LIMIT-TO (SRCTYPE , "j"))

2 TITLE-ABS-KEY (dosy OR pgse AND nmr AND molecular AND weight OR diffusion) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "re"))

1 TITLE-ABS-KEY (dosy OR pgse AND nmr AND molecular AND weight OR diffusion)

Figure 11. Example of the operators employed in a typical search.

From the resulting set of references representing 628 documents, we had to exclude some of them. First, we limited the documents to articles and reviews, rejecting the

rest of formats. The resulting dataset was reduced down to 587 documents. Then, we exclude the book series in source type, resulting in a final dataset based on 564 documents.

6.1. Classification of documents by type of documents

To classify all the documents along the TFG topic, as a first step we selected all the documents appearing between 1989-2020, collecting as mentioned previously, more than 600 documents of the type of articles, conference papers, reviews, books and book chapters (**Figure 12**). Then, we refined the documents to only focus in articles and reviews rendering a total of 564 documents in which 551 were articles (91.1%) and 13 were reviews (2.1 %). This type of graph is directly obtained from Scopus.

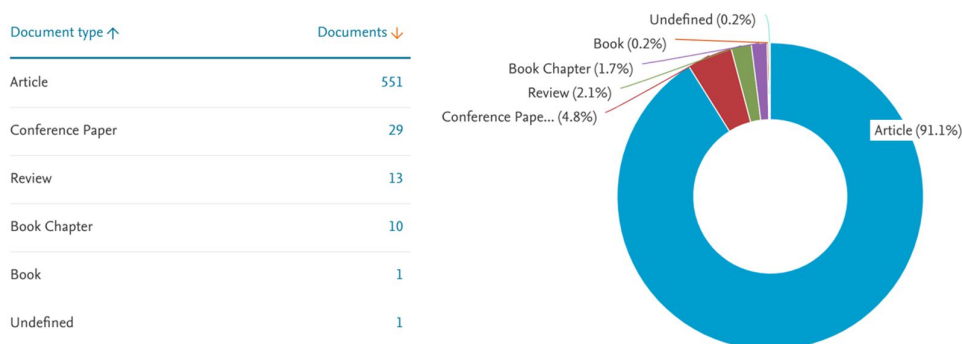


Figure 12. Showing all types of documents related to this TFG. Created by Scopus.

6.2. Classification of documents by subject area

The application of NMR molecular diffusion in the prediction of molecular weight is studied beyond the field of chemistry, such as materials science being located in the second place with a total of 89 documents (15.7%), just behind Chemistry that presents the vastest section with 213 documents (37.7%). Chemical Engineering with 73 documents, and then Biochemistry, Genetics and Molecular Biology with 72 documents occupy the third and fourth position in the global ranking. Then we have the rest of subject areas, for instance, Physics and Astronomy, Pharmacology, Toxicology and Pharmaceutics, Engineering, Environmental Science, Agricultural and Biological Sciences, Energy. In **Figure 13**, we can visualize the contribution of each subject area.

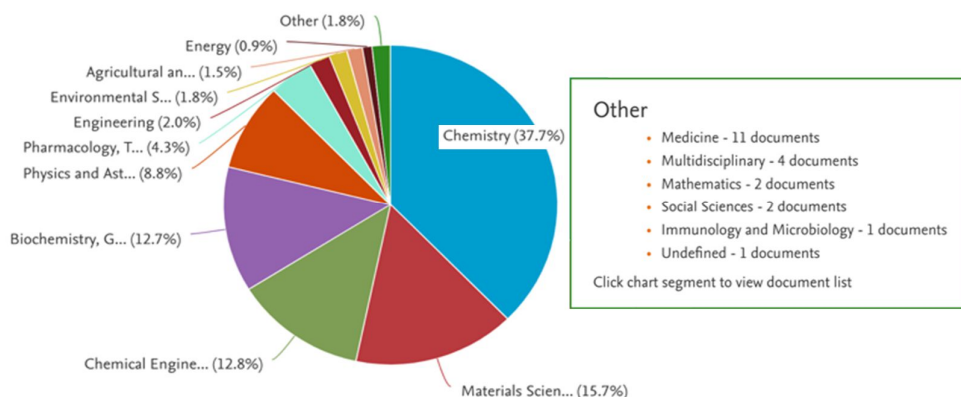


Figure 13. Distribution of documents based on different subject areas. Created by Scopus.

6.3. Classification of documents by publisher

In this section we have created a network graph using the Gephi software as it presents the most cited clusters and their inter-connection. The nodes represent the authors, and the labels illustrate the publishers. To build this graph, we first distributed the nodes. The two algorithms used were "Expansion" and then "Fruchterman Reingold". We also had to reorganize the clusters to detect communities above 0.496, which means that we had a large number of small groups that we had to eliminate in "Modularity Class" as they did not belong to any of the groups. Then, in the same partition we colored our clusters as shown in **Figure 14**.

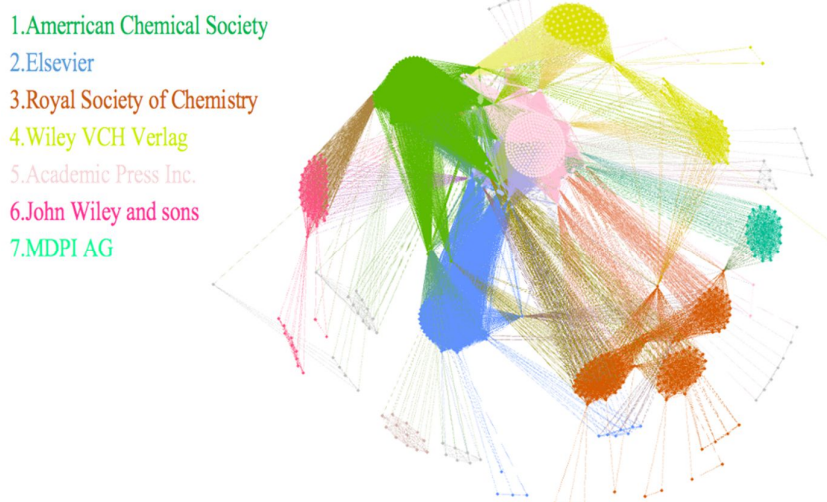


Figure 14. Clustering of the most cited publishers created with Gephi-0.9.2.

Figure 15 illustrates the major publishers detected in Gephi including the specific percentages. The main publisher is American Chemical Society with 68 publications (32%), then Elsevier with 47 documents (22%) related to our topic, and then in third place the Royal Society of Chemistry with 42 publications (20%). It is interesting to observe how the new open access editorial of MDPI created in 1996 is becoming important among them occupying the seventh position. The first 6 publishers accounts for 214 documents, and it is worth mentioning that there are 33 other publishers, representing a 19 %, with a lower number of documents related to our topic. Among these we have to mention MDPI, Blackwell Publishing Ltd, Nature Publishing Group, Taylor and Francis and Springer New York. Importantly, we detected around 295 documents from our dataset with no publisher information, representing a 52.3%.

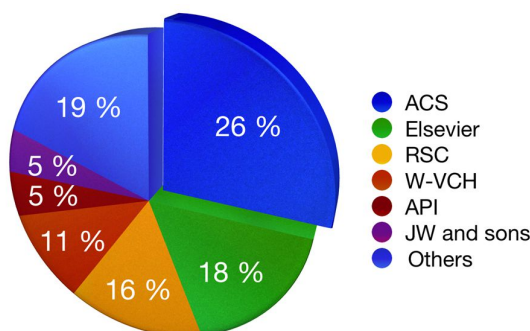


Figure 15. Major publishers found all over the articles and reviews of our dataset.

6.4. Classification of documents by authors

In this section we will depict the 15 authors with the highest number of publications. As shown in **Figure 16**, Pregosin, P.S. is the main author in this specific field with a total of 14 publications. Morris, G.A. and Price, W.S. share the second position in our list, both rendering 13 publications. Nilsson, M. occupies the third position with a total of 11 documents. Interestingly, there is one Spanish researcher located at the fifth position with 9 contributions, i.e. Fernández, I. which currently works at Universidad de Almería.

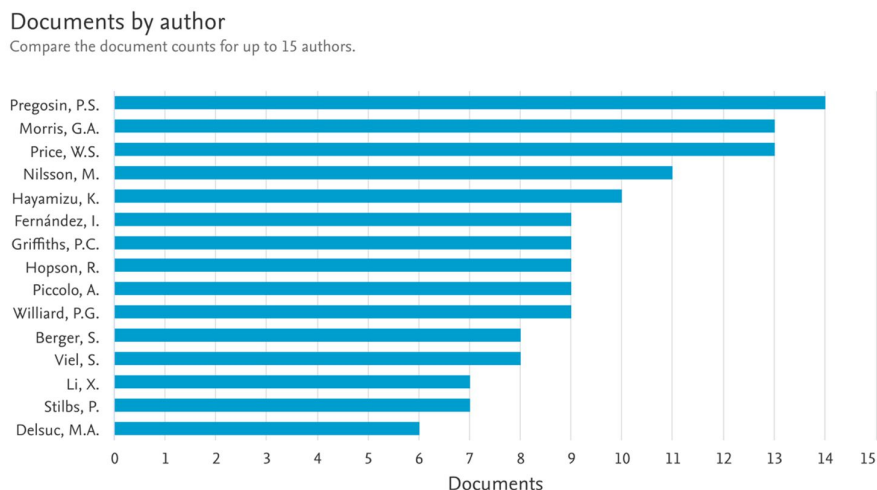


Figure 16. Graph representing a list of the 15 authors with most publications related to the topics of this TFG. Obtained from Scopus.

6.5. Classification of documents by country or territory

The classification of countries represents an important feature in this TFG as it indicates a slight evidence of the interest of each country in this scientific discipline.

As expected, United States dominate this yard, introducing around 100 documents with an impressive stock of 3470 citations. Then Europe (France, Italy, Germany) occupies the second place with 68, 60 and 58 documents, respectively. They also reached around 1600 citations each. Spain hits the seventh place with a total of 35 documents with 898 citations, which is considered a great achievement in comparison with its few publications. These main countries together with the rest are shown in **Figures 17** and **18**, created by Scopus and VOSviewer, respectively. **Figure 17** is one of the many graphs that Scopus database offers in its plugins as we reached this classification by using the algorithm "Analyze search results". To create **Figure 18**, and since we are analyzing the documents and their citations in relation with territories, we first selected the .CSV and then in "Type of analysis" we chose citation and in "Unit of analysis" we chose countries. In the next window we chose the desired threshold. In here, we selected 8 the minimum number of documents of a country, and 50 citations as the minimum number of citations of a country. By doing this, of the 57 countries, 23 meet the thresholds. Then, we visualized the graph by selecting the algorithm "Density Visualization" and "Cluster density", since we wanted to analyze the impact of each country in relation to our TFG topic.

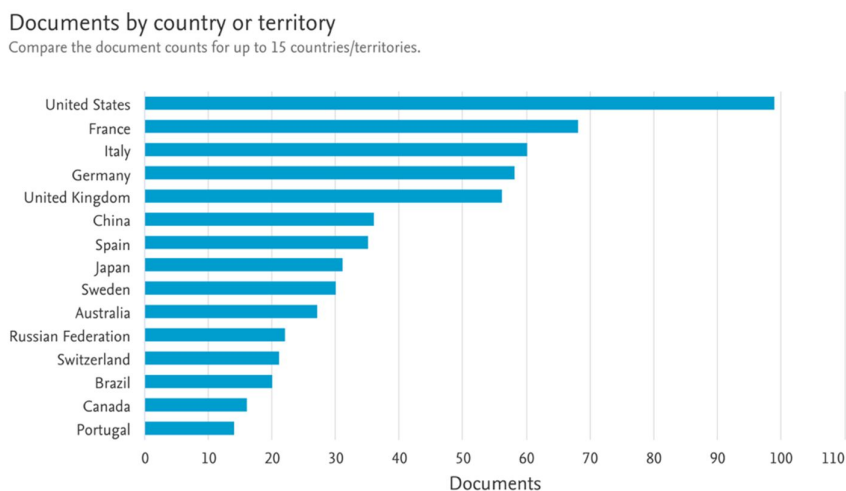


Figure 17. Comparison of the 15 countries/territories with the major number of documents according to the TFG topic. Created in Scopus.

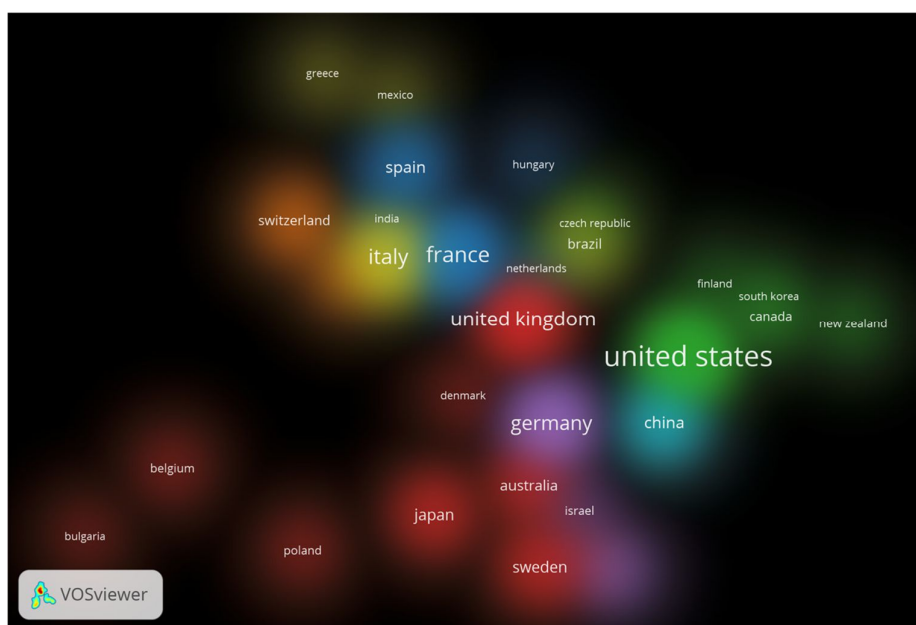


Figure 18. A layout of density map visualization of countries/territories within the documents related to TFG topic.

Interestingly, the closer the countries the higher their co-citations. The fact that United States are closer to China, Germany, Canada or United Kingdom than to other countries such as Bulgaria or Belgium, implies that the documents published by United States researchers cite documents preferentially from these former countries.

6.6. Correlations between authors and year of publication

According to WebOfScience, the community of scientist released the first article related to our TFG in the early 20st century, more specifically in 1905, being written by Einstein, A. After this work, no publication appeared until 1950, where Hahn, E. L. published one of the seminal works in the diffusion NMR discipline. **Figure 19** shows the importance of this article as one of the most cited documents in the scientific

community, since from this publication the whole NMR diffusion methodology and theory had been grown up.

The increment of documents was significant in the 21st century, and the possibility of accessing to more sophisticated instruments is probably one of the main reasons. The paper published by Stejskal, E. O. in 1965 is also one of the most cited. In the correlation given in **Figure 19**, papers published in 2020 are not included in the analysis as those have not been cited yet. Interestingly, other authors already introduced appear with many connecting lines such as Pregosin, Price or Fernández.

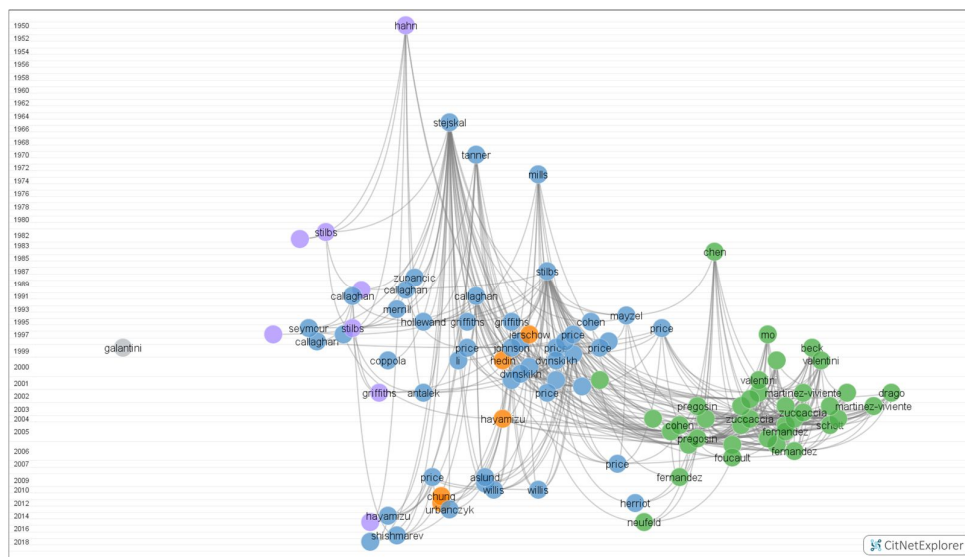


Figure 19. Visualized graph of citations of authors released by years.

Figure 19 also provides an interesting feature. From the figure one can observe that there are two main clusters, a green one and a blue one. This specific clustering is based on co-citations, what implies that the authors in green are usually cited to each other, and the same happens with the authors found in the other cluster. The fact of having two main clusters is usually associated with two different fields of applications, however, there are connections between both them albeit they are only few.

Besides, in **Figure 20**, we have created a graph with Gephi in which we present the clusters in which the authors are classified based on different period times, and further, how these authors are linked between each other. In this graph, the dataset from WebOfScience was used, in where again, the last 30 years were the decades with the most concentration of publications. Clusters 13 and 14 that belongs to the earlier publications and the most recent ones, respectively, have almost no representation in the graph since their number is rather small with respect to the rest of clusters, and because their citations are also very limited.

For its design, we used the layouts obtained from the “Expansion” and “Fruchterman Reingold” plugins, in which the nodes represent authors and the edges present years. To make the graph more understandable, we made some tuning in the modularity. As a result, we were able of deducing 14 different communities or clusters. For coloring those clusters, we used a partition on attribute, by selecting "Modularity Class", and in this way the percentage of communities appears so then we finally applied the algorithm and colored them.

The three largest clusters (red, green and light blue) include the authors who have published between 1992 and 2010, (36.79% of the total). Interestingly, the following cluster corresponds only to the authors who have published in 2017 (10.16% of the total), the year in which more publications have been registered in our search (43 documents). On the contrary, the year with less authors is 1989 with only one article and 0.1% (2 authors), which is the cluster 14 in light grey.

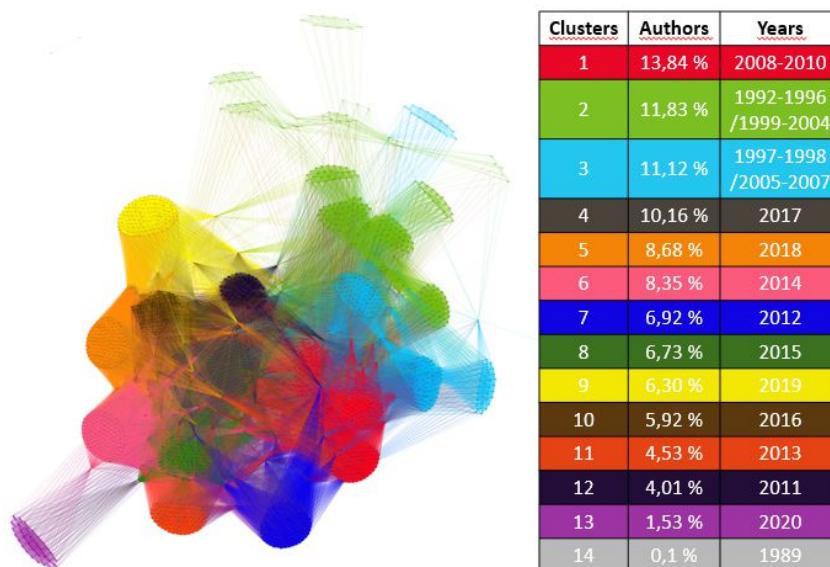


Figure 20. Network of author communities or clusters along and their publications between the range of years between 1989 to 2020.

6.7. Correlations between authors and citations

An essential part of any bibliometric study is the understanding of the connections between authors. From these connections the collaborations between research groups are evident and give a real picture of which groups are working in similar disciplines or topics. Interestingly, co-citations are turning to have a significant influence on increasing the impact of any scientific discipline.

We have highlighted in the following graphs and tables the most influential scientists with the highest number of citations. For instance, **Table 3** shows the top 20 most cited scientist and their total link strength given as the number of documents in where at least two keywords occur together.

As it is shown, Pregosin P.S. is the leader in this section, with the highest total link strength and number of documents, but on the contrary, is located in the second place in terms of citations (680 citations). In the first place of citations is Hayamizu K. with an overall of 849 citations on its 10 documents. In second place of total link strength and cited documents appears Morris G. A. with 420 and 13, respectively. The third position in this ranking goes to Price W. S. with 320 citations and 74 total of link strength. The first Spanish author is Fernández I. located in seventh position in the list but overcoming both Morris and Price in the total number of citations, and also defeating Price in total link strength.

Figure 21 illustrates the citations/documents ratio in where the top three researchers with highest numbers Morris K. F., Hayamizu K. and Fernández I. are the three of them represented in red. Among the three of them, Morris is probably the most

recognized author with not a very large number of documents but very cited in the diffusion NMR community.

Table 3. The top 20 list of authors with highest number of citations.

Author	Documents (D)	Citations (C)	D/C	Total link strength
Pregosin P. S.	14	680	48.6	489
Morris G. A.	13	493	37.9	420
Price W. S.	13	320	24.6	74
Nilsson M.	11	420	38.2	416
Hayamizu K.	10	849	84.9	57
Hopson R.	9	460	51.1	418
Williard P.G.	9	460	51.1	418
Fernández I.	9	495	55.0	291
Li X.	9	322	35.8	226
Piccolo A.	9	426	47.3	86
Griffiths P. C.	9	155	17.2	61
Viel S.	8	281	35.1	300
Berger S.	8	435	54.4	177
Stilbs P.	7	353	50.4	143
Shestakova P.	6	76	12.7	114
Kuchel P. W.	6	105	17.5	21
Morris K. F.	5	666	133.2	322
Li W.	5	199	39.8	200
Kumar P. G. A.	5	207	41.4	116

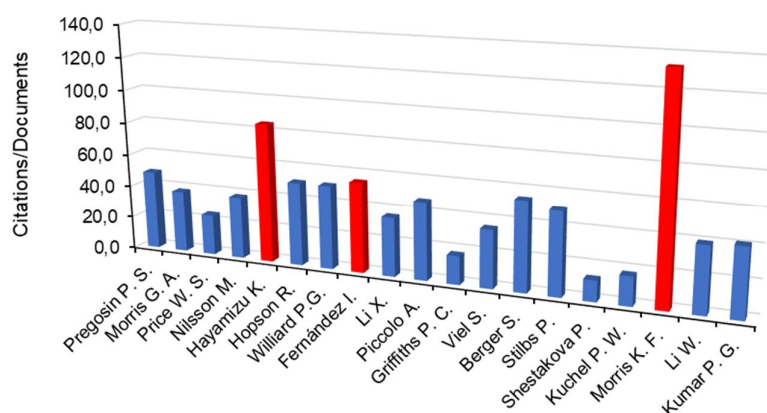


Figure 21. Citations/documents ratio along the top authors given in Table 3. In red are represented the top three.

Figure 22 provides a clustering of authors vs citations created in the VOSviewer and Gephi software. In **Figure 22 (top)**, we have selected "Network visualization" to generate the graph in where the size of the nodes is associated with the number of documents of each author. **Figure 22 (bottom)** was created using the Force Atlas 2 distribution. In the latter, the clusters are colored by differentiating them with "Modularity Class". We obtained 125 communities, so we discarded those that represented less than 0.1% of the total number of authors (as they occupy the lowest position in our classification with poor edges). Then, we made some adjustments in the 'Display Node Label' to add names of the authors playing with the size of the text.

The clusters shown with the highest density in yellow are the most cited authors as explained earlier.

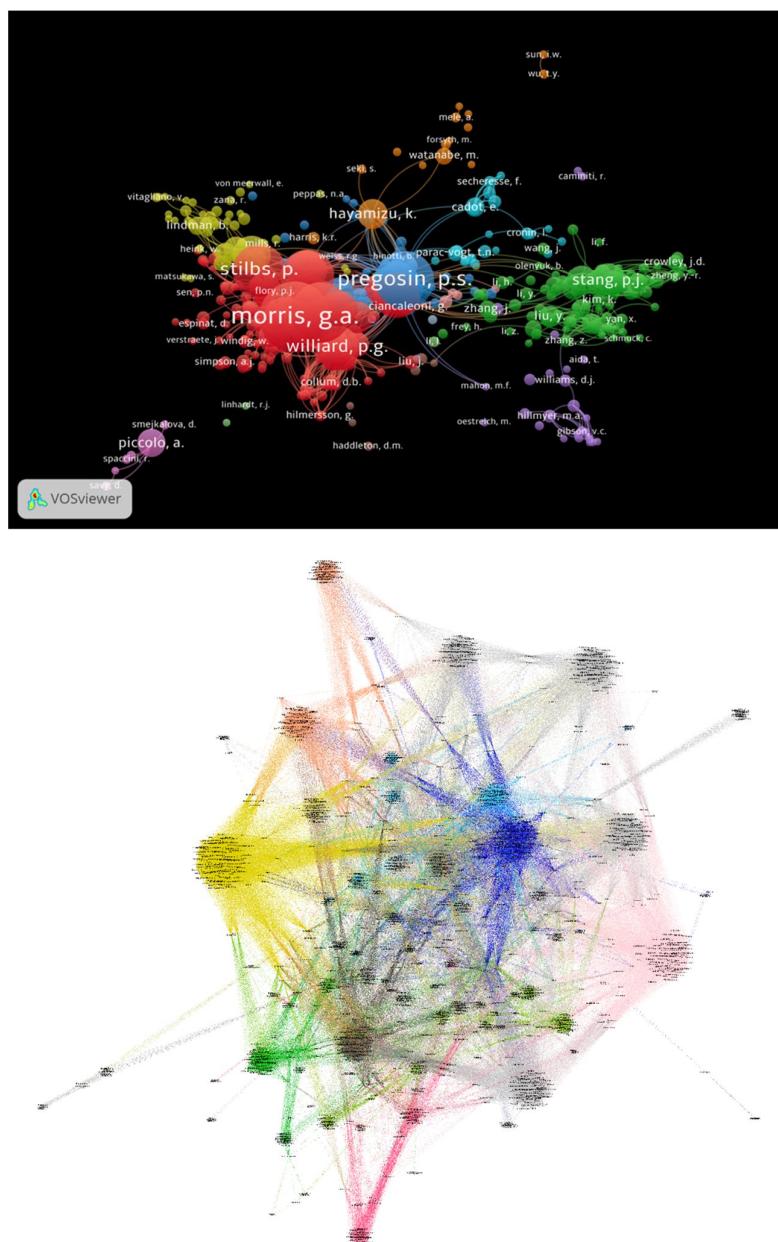


Figure 22. VOSviewer density map (top) and Gephi Layout (bottom) built on cited authors. In the latter clusters are given as nodes and citations as edges.

6.8. Classification of documents by index keywords

Keywords are considered a crucial index variable since they qualify the documents to appear in the above databases and allow to classify them. In addition, they help the searcher to contextualize the publication along with the field, research focus and subject.

Using the index keywords included in our dataset, permitted us to create some visualization graphs in VOSviewer. **Figure 23** shows the most used keywords by authors in their documents, as each publication contain at least five keywords in the abstract to guide the reader to better understand the aim of the work. This visualization map detects the connection between keywords, in a way in where, for instance, the node “molecular dynamics” or “polymers” are the most cited based on the their size, as we can observe in the blue clusters of the graph, which have a considerable amount of connections with the rest of clusters (ranked as 92 links), together with the keywords “water” (93 links) from the yellow cluster, or the keyword “methodology” (63 links) that appears in the red cluster. These keywords are shown to be the most used in publications involving PGSE and DOSY NMR diffusion-based spectroscopic methodologies.

For the creation of this graph, we had to configure the algorithms "Clustering" to a "Minimum cluster size" of 12, and the algorithm "Resolution" which was increases to 2. We also selected "Merge small clusters" to make the small cluster merged into bigger clusters. For the visualization, we leave the default setup except changing the background color into black. We thus obtained 8 clusters out of the 143 items introduced to the system.

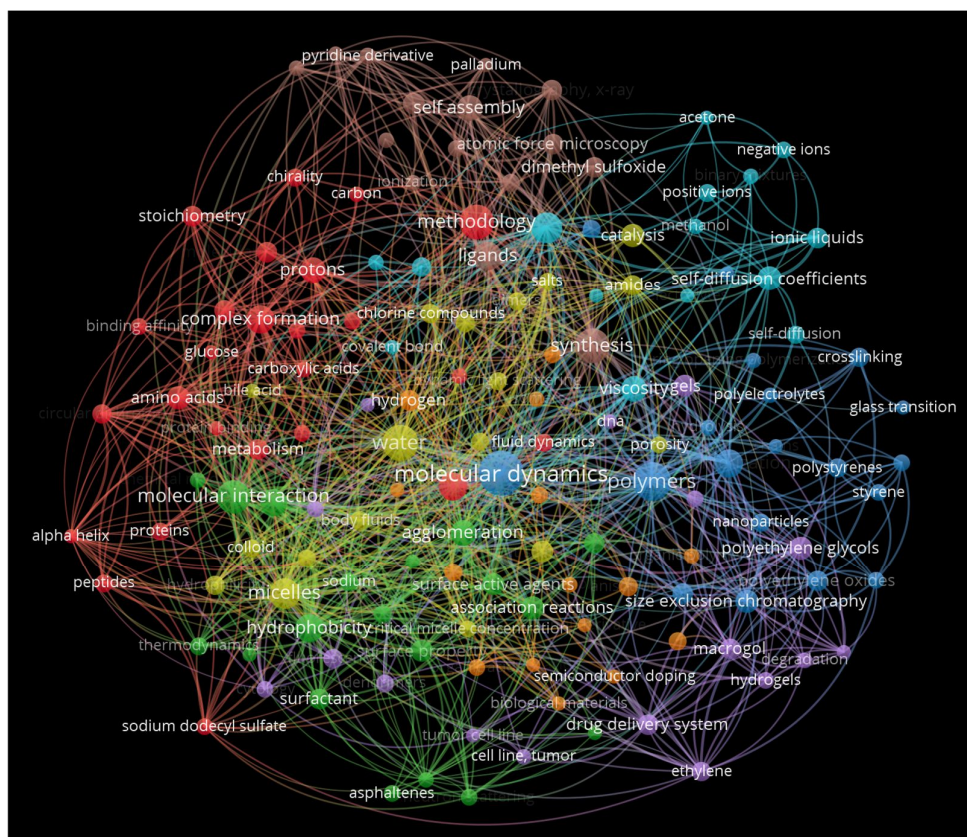


Figure 23. Most cited index keywords found in our database.

As can be deduced all the clusters are somehow connected to each other, where we have shown all the keywords that at least have 10 links to other clusters.

Cluster 1 (in red) is represented in **Figure 24** and includes keywords with the major number of links. In this cluster the keywords “Hydrodynamics”, “Complex formation”, “Amino acids” or “Proteins” are probably the one that mostly represent the field and subjects of the related publications. Importantly, this cluster is directly connected to cluster 2 (in green) and cluster 4 (in yellow).

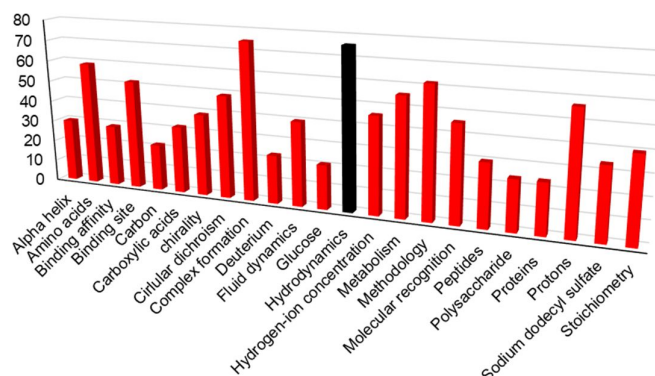


Figure 24. Keywords included in Cluster 1. In black is given the most representative.

Clusters 2 (in green) and cluster 4 (in yellow) are shown in **Figure 25**. Probably the most representative keywords for each group is “Molecular interaction” and “Micelles”, respectively, what clearly explains why are connected the previous cluster, since they share most of the keywords.

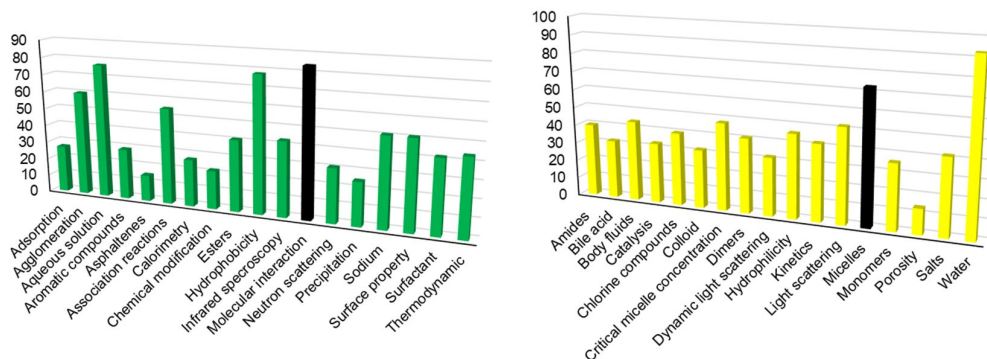


Figure 25. Keywords included in Cluster 2 (in green) and cluster 4 (in yellow). In black are given the most representative.

Cluster 3 (in dark blue) provides the keywords with the highest number of connections (to the green, yellow, lilac, orange, light blue, and even the brown cluster), what explains why it is situated in the middle of the density map shown in **Figure 23**. Probably the keyword that mostly represents the research field of this cluster would be “Polymers” which have been marked in black in **Figure 26**. Other important keywords are “Molecular dynamics” and “polymerization” which are clearly associated to almost every publication related within the study of polymers.

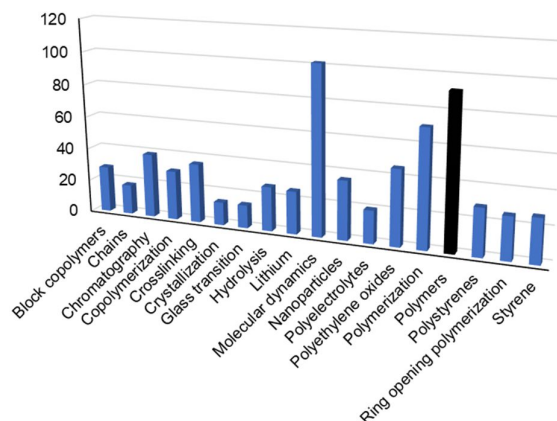


Figure 26. Keywords included in Cluster 3. In black is given the most representative.

In cluster 5 (in lilac), "Drug delivery system" is probably the top keyword as it resumes the rest of keywords shown in this community. It is linked to only 3 clusters, i.e. orange, blue and green. In the same way, cluster 6 (in light blue) has in general a low number of links, however, we could deduce that with together with the rest of keywords addresses at least partially the supramolecular chemistry field with keywords such as "beta-cyclodextrins", "cyclodextrins" and "hydrogen bonds". Both clusters are represented in **Figure 27**.

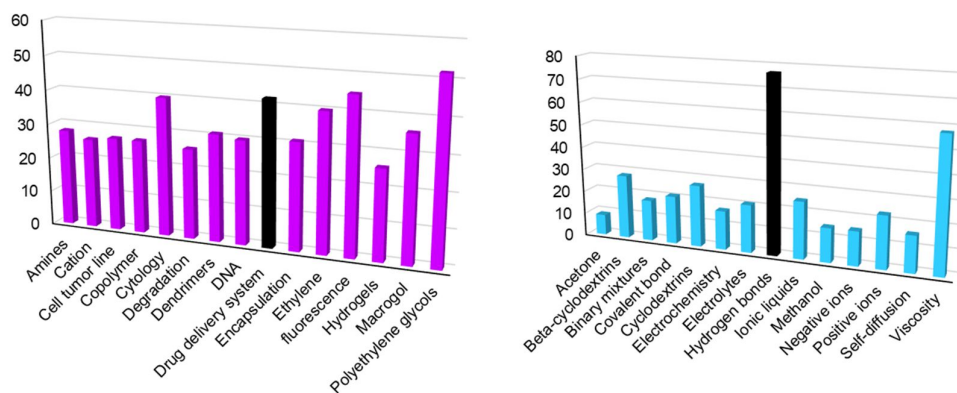


Figure 27. Keywords included in Cluster 5 (in lilac) and cluster 6 (in light blue). In black are given the most representative.

In cluster 8 (in brown), we identified a powerful keyword who probably represents the rest of them which is "organometallic compounds", but due to its low number of links to other communities, it is somehow isolated in the top of the density map of **Figure 28**. Their main connections are with clusters red, light blue and yellow. Unfortunately, in cluster 7 (in orange), all the keywords found are difficult to represent with just one or two keywords and this is the reason why in **Figure 23** are located somehow diluted and connected with almost every cluster. The keywords "Hydrogen", "Organic acids" and "Organic compounds" account for the multi-disciplinarity of this cluster.

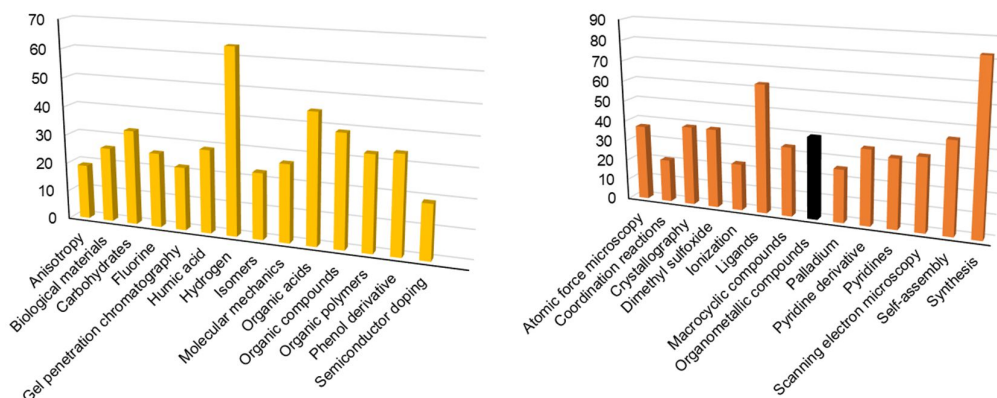


Figure 28. Keywords included in Cluster 7 (in orange) and cluster 8 (in brown). In black are given the most representative.

The keywords contained on each of the eight clusters are given in **Table A1** and **A2** in the annex. Altogether, **Figure 29** provides density maps including the main index keywords for clusters in blue (cluster 3), red (cluster 1), and brown (cluster 8). We have chosen these four clusters due to their major importance and impact in the organic chemistry field. To build these maps we have used Scopus to extract the documents corresponding to each group, i.e. the keyword "Polymers" was found in 175 documents, the keyword "Proteins" was localized in 54 documents, and finally the keyword "Organometallic compounds", was found in an overall of 22 documents.

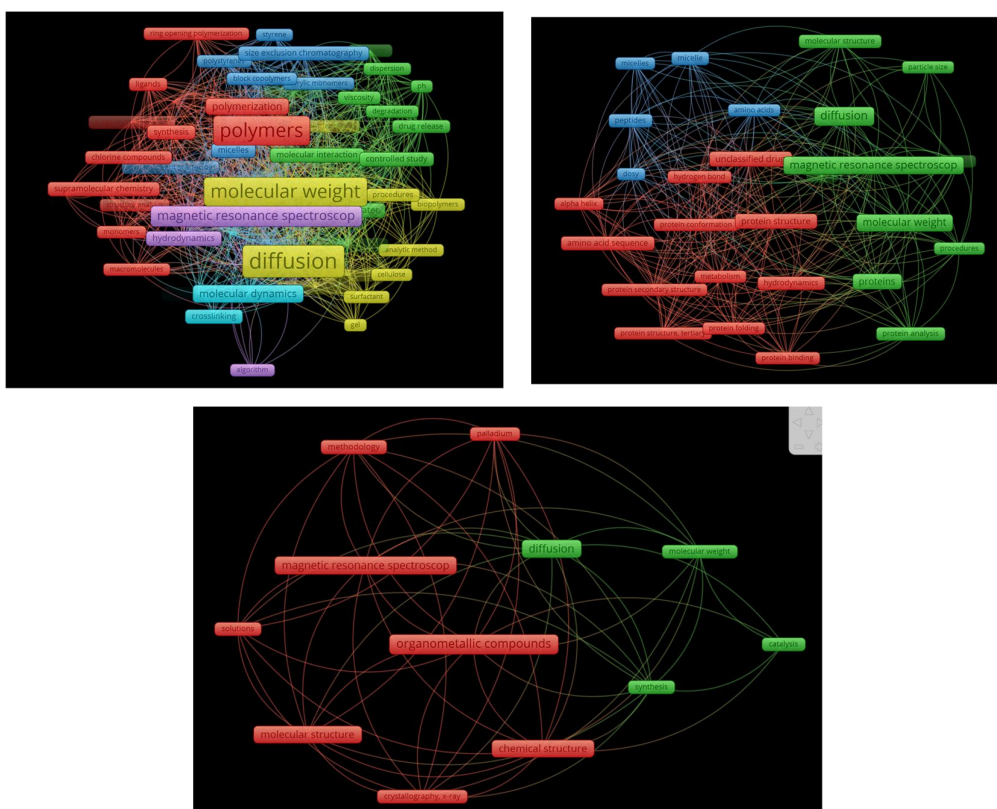


Figure 29. Density maps of main keywords found in the 175 documents for cluster 3 (blue), 54 documents for cluster 1 (red) and 22 documents for cluster 8 (brown).

The provided clusters account for almost an 87.6 % of the documents of our dataset. However, there is still a 12.4 % of publications that have not been classified. As an example of this group we have extracted the document titled "Relationship between Li^+ diffusion and ion conduction for single-crystal and powder garnet-type electrolytes studied by ^7Li PGSE NMR spectroscopy" of Hayamizu, K. et al. which belongs to the field of Physics and Astronomy, and its index keywords show a strong relationship with Solid-State Batteries, Solid Electrolytes, and Garnets, and consequently was not included in any of the above-mentioned clusters.

6.10. Limitations encountered during the performance of this TFG

While elaborating this work, we have found several issues that are worth to mention herein and that have been summarized in the following points:

- ∞ Gephi is a powerful statistical program but a weak tool for analyzing and graphing a big dataset with thousands of nodes and labels or edges, as it took up to 2 nights to create some of the graphs represented in this TFG. Another important issue is that the program does not accept any type of document except the GEXF file created with specific programs like Table2Net. Another way of introducing data to the Gephi software is by setting up two different spreadsheet one for nodes and another one for edges, which is tedious in time and very prone to make mistakes. In the statistic view, when asking for text, the names of the different authors appeared all together being difficult to read.
- ∞ VOSviewer does not give the researcher the opportunity to make its own modification in the visualized graph as given in Gephi.
- ∞ Scopus has also some limitation when exporting big amounts of data, as it collapses several times. An alternative, the extraction of documents should be in sections. Another sensitive point to take in consideration is the selection of operators for the search. This step has a profound impact in the number of documents appearing in the research.
- ∞ In the classification of keywords for each represented cluster, we had a certain number of documents to extract from the dataset obtained from Scopus, however, all of those keywords were not exclusive to one cluster, as the same keyword appeared in various clusters due to its strong connection to many other groups. Therefore, some of the documents extracted for each cluster may share some other keywords in other clusters so are not exclusive of that specific cluster.

7. CONCLUSIONS

We have performed a bibliometric study on the applications of NMR molecular diffusion, covering an overall of 564 documents of 40 different publishers distributed in 17 subject areas. We have used two main databases such as Scopus and WebOfScience and several new programs such as Gephi, VOSviewer, CitNetExplorer, OpenRefine and Table2Net. We have revised the accessibility of the cited articles and reviews in the employed databases and developed density maps and several graphs based on co-authorship, bibliographic coupling, keyword co-occurrence and co-citation. The most active publishers were the American Chemistry Society, Elsevier, Royal Society of Chemistry and Wiley VCH. If we look at the countries with most publications, we have detected a relation between the origin countries of publishers within the most cited territories. United States, United Kingdom and Europe are the

most powerful in this field, and this might be explained by the development of those countries in terms of science. Asia is not as active as the rest of continents in this specific field, owing that only China and Japan appear in our study. Analyzing the index keywords, we have deduced that the most prominent topics covered are integrated in the groups of polymers, hydrodynamics, drug delivery systems and organometallic compounds. As the focus of scientists is steered into a specific field related to organic chemistry, the rest of keywords are referring to methods and instrumentation employed for the progress of this research field.

In addition, skills such as the practice of a second language, in this case English, have been achieved due to the consultation of many scientific and bibliographic databases and datasets. Finally, the critical judgment competence has been also worked along this TFG that have allowed me to analyze in a critical manner the documents, clusters, keywords, etc. that have been finally provided in this work.

8. REFERENCES

- ¹ Groos, O. V.; Pritchard, A. Documentation notes. *J. Doc.* **1969**, 25 (4), 344-349.
- ² Lund University Library Home Page. <https://www.ub.lu.se/en/publish/bibliometrics> (accessed February 22, 2020).
- ³ Scopus Home Page. <https://www.elsevier.com/solutions/scopus> (accessed February 22, 2020).
- ⁴ Van Eck, N. J.; Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, 84 (2), 523-538.
- ⁵ Cobo, M.J.; López-Herrera, A. G.; 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. *J. Informetr.* **2011**, 5 (1), 146-166.
- ⁶ Small, H. Tracking and predicting growth areas in science. *Scientometrics* **2006**, 68 (3), 595-610.
- ⁷ Hassan-Montero, Y.; Guerrero-Bote, V. P.; De-Moya-Anegón, F. Graphical interface of the SCIMAGO journal and country rank: an interactive approach to accessing bibliometric information. *EPI* **2014**, 23 (3), 272-278.
- ⁸ OpenRefine Home Page. <https://openrefine.org/> (accessed February 22,2020).
- ⁹ Blondel, V. D.; The Louvain method for community detection in large networks. Université Catholique de Louvain. **2011**, <https://perso.uclouvain.be/vincent.blondel/research/louvain.html> (accessed April 12, 2020).
- ¹⁰ The open Graph Viz Platform Home Page. <https://gephi.org/> (accessed February 22,2020).
- ¹¹ SciencesPo Médialab Tools Home Page. <http://tools.medialab.sciences-po.fr/> (accessed February 26, 2020).
- ¹² van Eck, N. J.; Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, 84 (2), 523-538.
- ¹³ Van Eck, N.J.; & Waltman, L. Visualizing bibliometric networks. *Measuring scholarly impact: Methods and practice*; Ding, Y., Rousseau, R., Wolfram, D., Eds.; Springer, 2014; pp 285-320.
- ¹⁴ Lambert, J. B.; Shurvell, H. F.; Lightner, D. A.; Cooks, R. G. Organic Structural Spectroscopy. *J. Chem. Educ.* **1998**.
- ¹⁵ Akitt, J. W.; Mann, B. E. *NMR and Chemistry: An introduction to modern NMR spectroscopy*; Fourth edition; University of Sheffield, 2000.
- ¹⁶ Chen, A.; Wu, D. H.; Johnson, C. S. Determination of Molecular Weight Distributions for Polymers by Diffusion-Ordered NMR. *J. Am. Chem. Soc.* **1995**; 117 (30), 7965-7970.
- ¹⁷ Jayawickrama, D. A.; Larive, C. K.; McCord, E. F.; Roe, D. C. Polymer additives mixture analysis using pulsed-field gradient NMR spectroscopy. *Magn. Reson. Chem.* **1998**, 36, 755-760.
- ¹⁸ Kapur, G. S.; Findeisen, M.; Berger, S. Analysis of hydrocarbon mixtures by diffusion-ordered NMR spectroscopy. *Fuel* **2000**, 79, 1347-1351.

- ¹⁹ Raya-Barón, A.; Oña-Burgos, P.; Fernández, I. Diffusion NMR spectroscopy applied to coordination and organometallic compounds. *Anal. Methods* **2019**, *11*, 125-191.
- ²⁰ Arrabal-Campos, F. M.; Aguilera-Sáez, L. M.; Fernández, I. A diffusion NMR method for the prediction of weight-averaged diffusion coefficients of different viscosity. *Anal. Methods* **2019**, *11*, 142-147.
- ²¹ a) Arrabal-Campos, F. M.; Oña-Burgos, P.; Fernández, I. Molecular Weight Prediction with No Dependence on Solvent Viscosity by Diffusion NMR Approach. *Polym. Chem.* **2016**, *7*, 4326-4329; b) Arrabal-Campos, F. M.; Álvarez, J. D.; García-Sancho, A.; Fernández, I. Diffusion NMR of polystyrene blends. Unprecedented use of a genetic algorithm in pulse field gradient spin echo (PGSE) NMR. *Soft Matter* **2017**, *13*, 1000-1004; c) Aguilera-Sáez, L. M.; Fernández, I. An algebraic reconstruction technique for diffusion NMR experiments. application to the molecular weight prediction. *Chem.* **2019**, *123*, 943-950.
- ²² Crutchfield, C. A.; Harris, D. J. Molecular mass estimation by PFG NMR spectroscopy. *J. Magn. Reson.* **2006**, *185*, 179-182.
- ²³ Esturau, N.; Espinosa, J. F. Optimization of diffusion-filtered NMR experiments for selective suppression of residual non-diffusing NMR spectra of organic compounds. *J. Org. Chem.* **2006**, *71*, 4103-4110.
- ²⁴ a) Price, W. S. Pulsed-field gradient nuclear magnetic resonance as a tool for studying translational diffusion: Part I. Basic theory. *J. Chem. Phys.* **1975**, *62*, 336; b) Price, W. S. Pulsed-field gradient nuclear magnetic resonance as a tool for studying translational diffusion: Part II. Experimental studies. *J. Chem. Phys.* **1975**, *62*, 197-237; c) Geil, B. Measurement of translational diffusion using ultrahigh magnetic field gradient NMR. *Concepts Magn. Reson.* **2005**, *10*, 197-237; d) Kumar, P. G. A.; Fernández, I. Pulsed Gradient Spin-Echo (PGSE) Diffusion and ¹H, ¹⁹F Heteronuclear Overhauser Spectroscopy in Organometallic Chemistry: Something Old and Something New. *Chem. Rev.* **2005**, *105*, 2977-2998; e) Pregosin, P. S. Application of Diffusion NMR to ion pairing in inorganic chemistry: a mini review. *Magn. Reson. Chem.* **2017**, *55*, 405-413.
- ²⁵ D. Sinnaeve, The Stejskal–Tanner Equation Generalized for Any Gradient Shape—An Overview of Most Pulse Sequences Measuring Diffusion. *Magn. Reson. Chem.* **2012**, *40A*, 39–65.
- ²⁶ Stejskal, E. O.; Tanner, J. E. Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient. *J. Chem. Phys.* **1973**, *57*, 223-232.
- ²⁷ Tanner, J. E. Use of the Stimulated Echo in NMR Diffusion Studies. *J. Chem. Phys.* **1970**, *52*, 2523-2526.
- ²⁸ Burstein, D. Stimulated echoes: Description, application, practical hints. *Concepts Magn. Reson.* **1996**, *8*, 269-278.
- ²⁹ Wu, D. H.; Chen, A. D.; Johnson, C. S. An Improved Diffusion-Ordered Spectroscopy Experiment Incorporating Bipolar Gradients. *J. Magn. Reson.* **1997**, *126*, 260-264.
- ³⁰ Gibbs, S. J.; Johnson Jr, C. S. A PFG NMR experiment for accurate diffusion and flow studies in the presence of eddy currents. *J. Magn. Reson.* **1997**, *126*, 265-270.
- ³¹ Khrapitchev, A. A.; Callaghan, P. T. Double PGSE NMR with Stimulated Echoes: Phase Cycles for the Selection of Desired Diffusion Components. *J. Magn. Reson.* **1997**, *126*, 271-276.
- ³² Martínez-Viviente, E.; Pregosin, P. S. Low Temperature ¹H-, ¹⁹F-, and ³¹P-PGSE Diffusion Measurements. Applications to the Study of Diffusion in Polymers. *Acta Chem. Scand.* **2003**, *86*, 2364.
- ³³ Jerschow, A.; Müller, N. Suppression of Convection Artifacts in Stimulated-Echo Diffusion Experiments. Double-Stimulated-Echo Diffusion NMR. *J. Magn. Reson.* **2003**, *175*, 372-375.

³⁴ Alcayde, A.; Montoya, F. G.; Baños, R.; Manzano-Agugliaro, F., A fast method for identifying worldwide scientific collaborations using the Scopus database. *Telemat. Inform.* **2018**, *35*, 168-185.

³⁵ Montoya F. G.; Baños R.; Alcayde A.; Montoya M. G.; Manzano-Agugliaro, F., Power Quality: Scientific Collaboration Networks and Research Trends. *Energies*. **2018**, *11*, 2067.

³⁶ Falagas, M. E.; Pitsouni, E. I.; Malietzis, G. A.; Pappas, G. Comparison of PubMed, Scopus, WebOfScience, and Google Scholar: strengths and weaknesses. *FASEB J.* **2007**, *22*, 338-342.

³⁷ Preparing data with OpenRefine Part II- Assign Unique Numerical Identifiers Home Page. <https://sites.temple.edu/tudsc/2016/12/13/prepa> (accessed March 5, 2020).

³⁸ Table2Net Home Page. <https://medialab.github.io/table2net/> (accessed March 5, 2020).

9. LIST OF ABBREVIATIONS

NMR: Nuclear Magnetic Resonance

TFG: Trabajo Fin de Grado

VoS: Visualization of Similarities

UTC: Université de Technologie Compiègne

IR Spectroscopy: Infrared Spectroscopy

UV-Vis Spectroscopy: Ultraviolet visible Spectroscopy

MS Spectroscopy: Mass Spectroscopy

DOSY: Diffusion-Ordered Spectroscopy

PGSE: Pulsed Gradient Spin Echo

ResNetBot: Research Network Bot

BMP: Bitmap

EMF: Enhanced Meta Format

EPS: Encapsulated PostScript

GIF: Graphics Interchange Format

JPG: Joint Photographic Engineering Group

PDF: Portable Document Format

PNG: Portable Network Graphics

SVG: Scalable Vector Graphics

SWF: Small Web Format

TIFF: Tagged Image File Format

ACS: American Chemical Society

RSC: Royal Society of Chemistry

W-VCH: Willey Verlag Chemie

API: Academic Press Inc

JW and Sons: John Wiley and sons

MDPI: Multidisciplinary Digital Publishing Institute

ANEXO



Figure A1. How to upload files to OpenRefine.

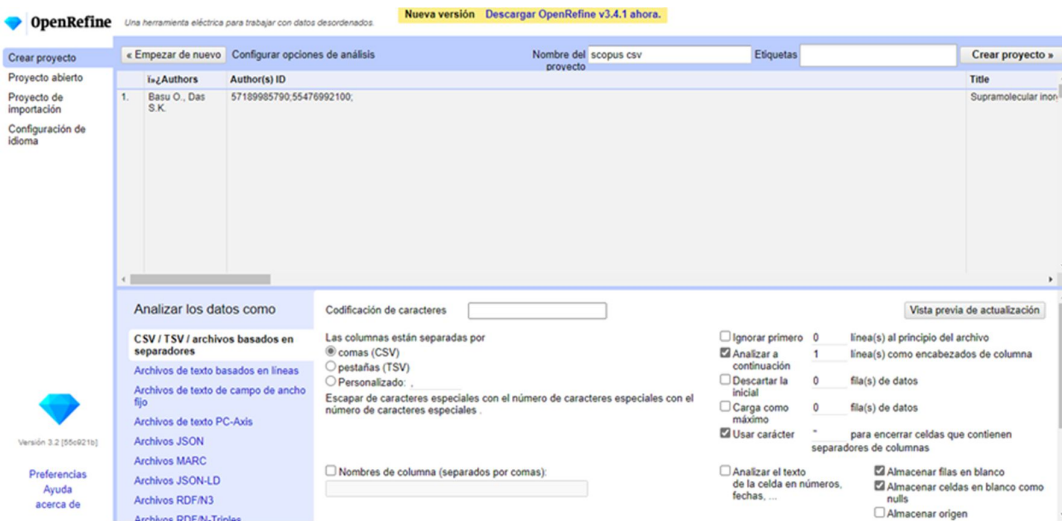


Figure A2. How to create a project in OpenRefine.

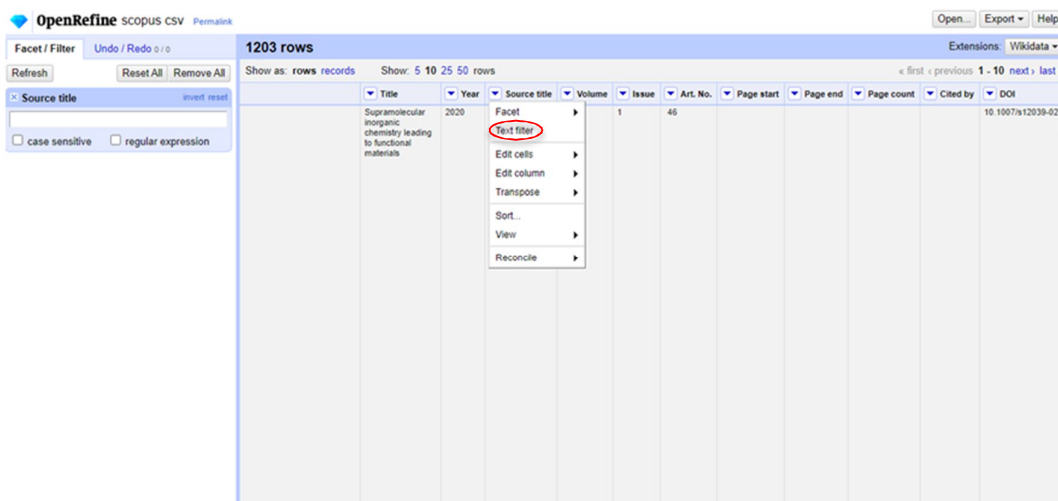


Figure A3. How to filter text in OpenRefine.

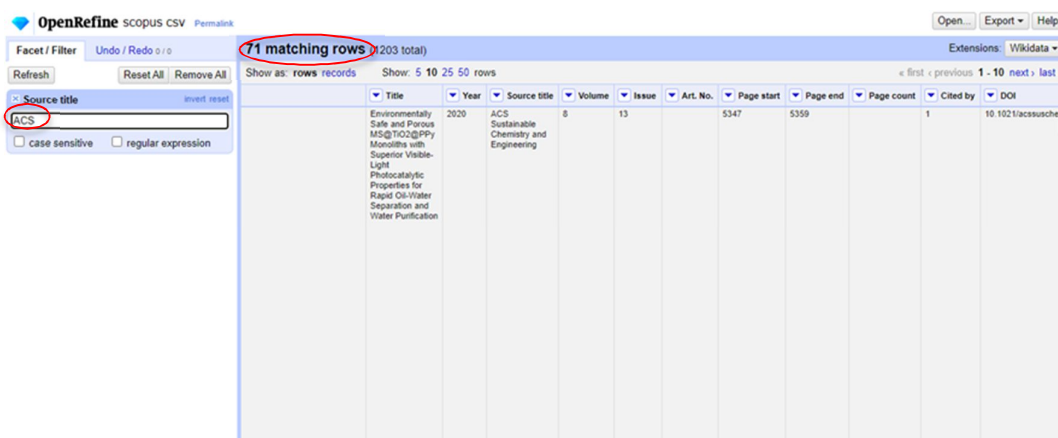


Figure A4. New data table by selecting the word "ACS".

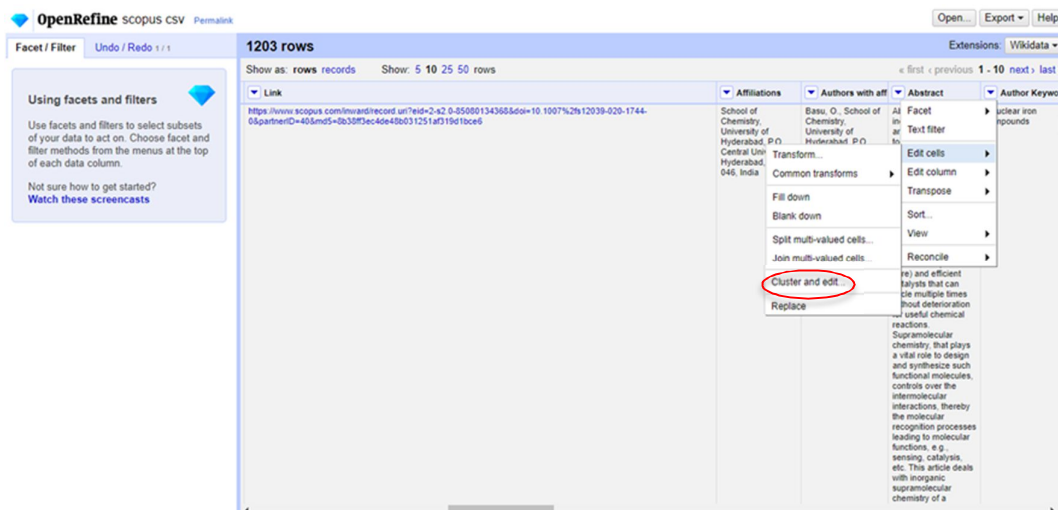


Figure A5. How to merge strings in OpenRefine.

Cluster & Edit column "Author Keywords 1"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more...

Method: **key collision** Keying Function: **fingerprint** 661 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	4	<ul style="list-style-type: none"> Bi-functional catalysts (2 rows) Bi-functional catalysts (1 rows) bifunctional catalysts (1 rows) 	<input type="checkbox"/>	Bi-functional catalysts
3	3	<ul style="list-style-type: none"> Electrocatalyst (1 rows) electro-catalyst (1 rows) electrocatalyst (1 rows) 	<input type="checkbox"/>	Electrocatalyst
2	3	<ul style="list-style-type: none"> Carbon nanotube (2 rows) Carbon Nanotube (1 rows) 	<input type="checkbox"/>	Carbon nanotube
2	7	<ul style="list-style-type: none"> heterogeneous catalysis (5 rows) Heterogeneous catalysis (2 rows) 	<input type="checkbox"/>	heterogeneous catalysis
2	2	<ul style="list-style-type: none"> Click chemistry (1 rows) click chemistry (1 rows) 	<input type="checkbox"/>	Click chemistry
2	2	<ul style="list-style-type: none"> Anti-bacterial property (1 rows) Antibacterial property (1 rows) 	<input type="checkbox"/>	Anti-bacterial property
2	4	<ul style="list-style-type: none"> electrocatalysis (3 rows) Electrocatalysis (1 rows) 	<input type="checkbox"/>	electrocatalysis

Buttons: Select All, Unselect All, Export Clusters, Merge Selected & Re-Cluster, Merge Selected & Close, Close

Figure A6. How to merge strings in OpenRefine.

OpenRefine scopus.csv Permains

Facet / Filter Undo / Redo 1 / 1

1203 rows Show as: rows records Show: 5 10 25 50 rows

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

Export project

- Tab-separated value
- Comma-separated value
- HTML table
- Excel (.xls)
- Excel 2007+ (.xlsx)
- ODF spreadsheet
- Custom tabular exporter...
- SQL Exporter...
- Templating...
- Upload edits to Wikidata
- Export to QuickStatements
- Export Wikidata schema

Background table columns: Link, Affiliations, At

Figure A7. How to export a refined document in OpenRefine.

Table 2 Net
+ Médiab Tools




Table 2 Net

Extract a network from a table. Set a column for nodes and a column for edges. It deals with multiple items per cell.

Load your CSV table

It has to be **comma-separated** and the first row must be dedicated to **column names**.

Seleccionar archivo Ningún archivo seleccionado

Note: you can drag and drop a file

[Tweet](#)
 We used:

[See also our other tools at Médiab Tools!](#)






 Developed by Mathieu Jacomy

Figure A8. How to upload the CSV file in Table2Net.

1. Type of Network

Normal (one type of node)



You will have to choose:

- Which column **X** will define the nodes
- Which column **Y** will define the links

You may extract different types of networks from a table. It depends on how you use columns to build the nodes and the edges.

- Normal:** if you want a single type of nodes, for instance *authors*. They will be linked when they share a value in another column, for instance *papers*.
- Dipartite:** if you want two types of nodes, for instance *authors* and *papers*, they will be linked when they appear in the same row of the table.
- Citation:** if you have a column containing references to another one, for instance *paper title* and *cited papers (title)*
- No link:** a single type of nodes, without link










Figure A9. Selection of the type of Network.

2. Nodes

① Which column defines the nodes?

Authors

Comma-separated ","

Sample of nodes extracted with these settings: (sample)

Example 1 Example 2 Example 3 Example 4

② Do you want nodes attributes?

Select one or several columns

You may transfer the content of some columns to the network as attributes of the nodes. This feature is only useful under certain circumstances, when the attribute columns actually qualify the node column. Else, it is possible (and probable) that multiple attributes correspond to a single node. If this happens, the multiple values will be concatenated with the | separator (pipe).

Warning: Adding metadata may cause a memory overload (a browser crash, not dangerous but you won't get any result)

3. Links

③ Which column defines the links?

Year

One expression per cell

Sample of items extracted with these settings: (sample)

2017 2019 2016 2019 2014

④ Do you want links attributes?

Select one or several columns

You may transfer the content of some columns to the network as attributes of the links. This feature is only useful under certain circumstances, when the attribute columns actually qualify the links column. In case of multiple values, they will be concatenated with the | separator (pipe).

Warning: Adding metadata may cause a memory overload (a browser crash, not dangerous but you won't get any result)

4. Additional settings

Optional: time series

No temporal data

Select only a column containing Integers.

Optional: edge weight

No weight

Links are naturally ranked by the number of rows matching in the table between the connected nodes. You may choose to weight the links according to it.

5. Build the network

Build and download the network (GEXF)

NB: this may take a while, please be patient.

After building the network, the download will trigger automatically. The network file is a GEXF, the Gephi file format.

Figure A10. Configuration options of Table2Net.

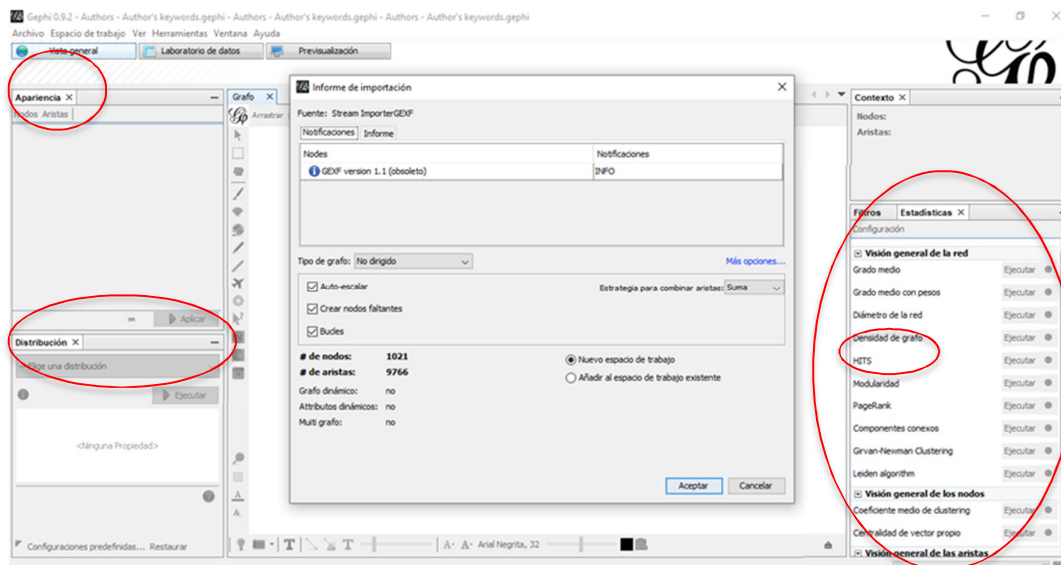


Figure A11. How to open a GEXF file in Gephi.

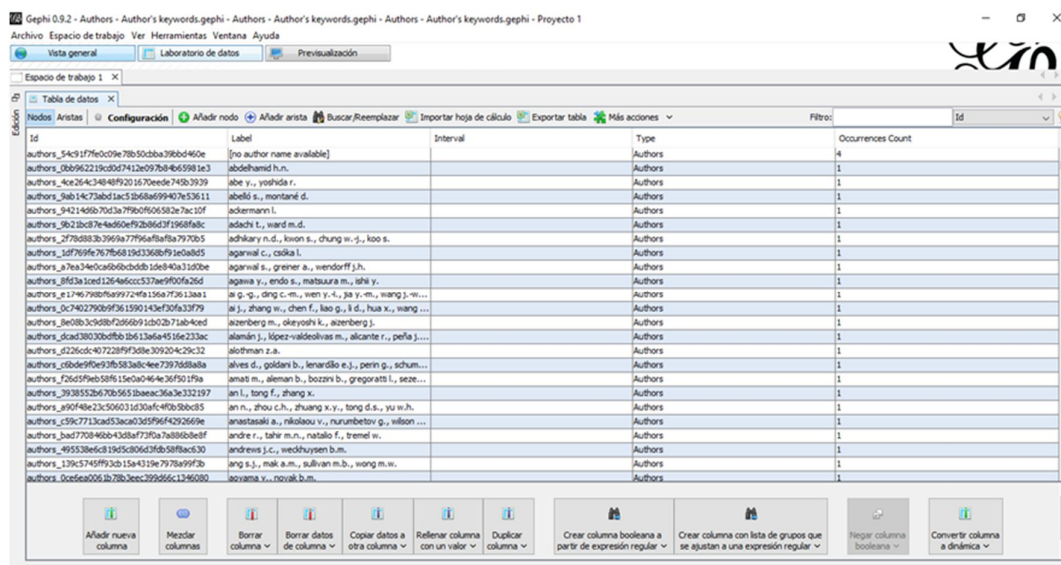


Figure A12. Data Laboratory in Gephi.

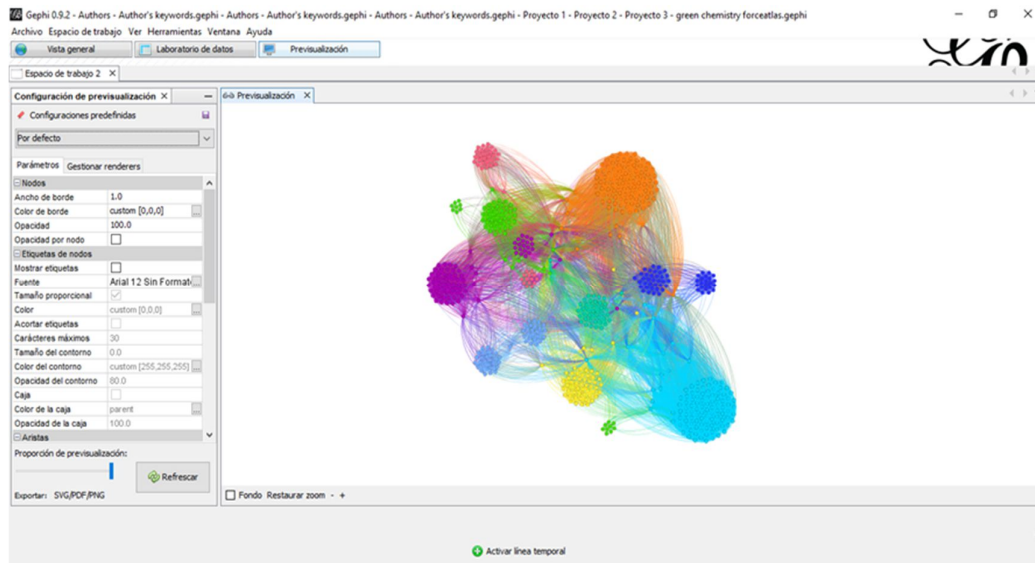


Figure A13. Previsualization in Gephi.

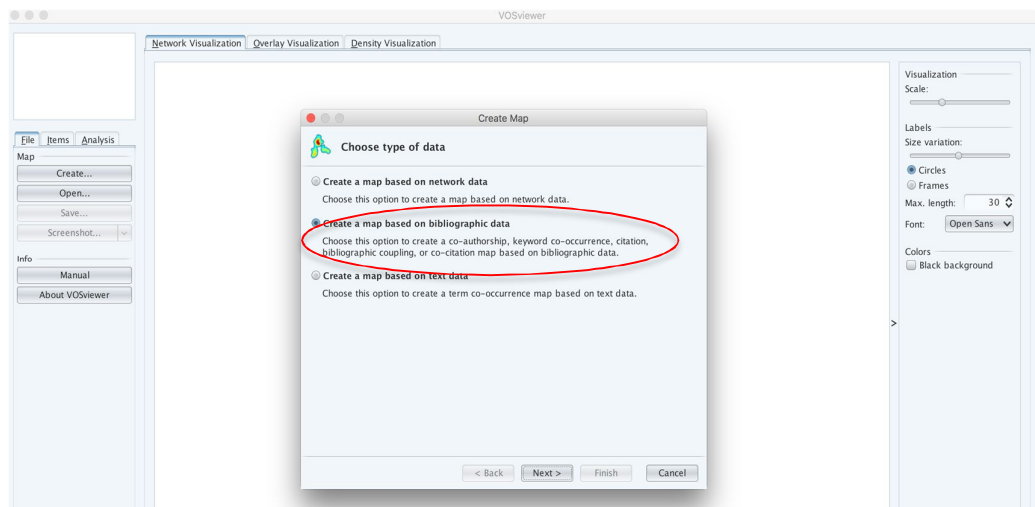


Figure A14. Creation of a map based on bibliographic data

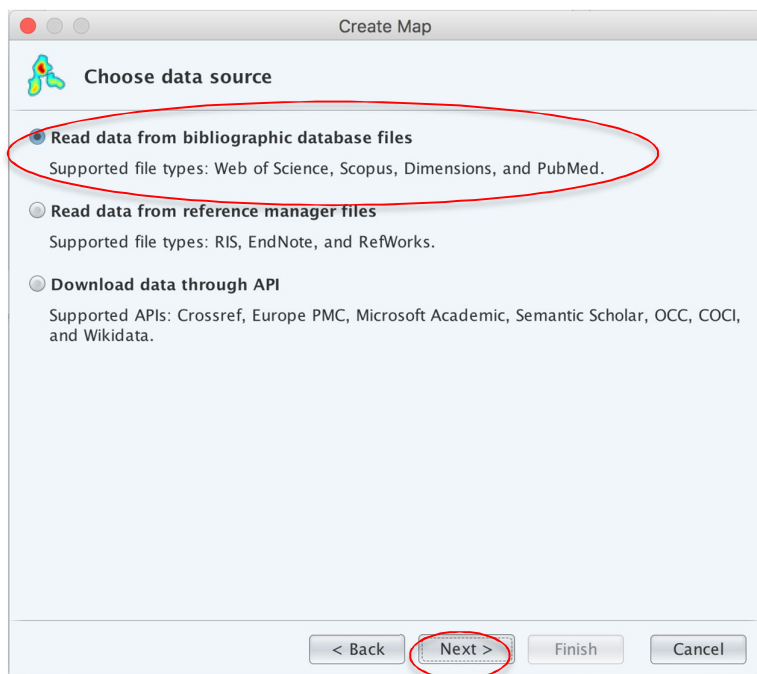


Figure A15. Selection the data from bibliographic database files.

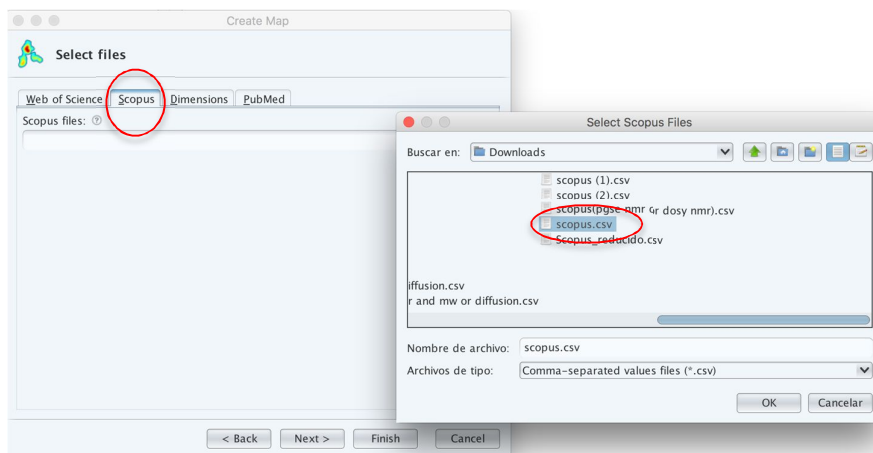


Figure A16. Selection of the .CSV file from Scopus database.

Create Map

Choose type of analysis and counting method

Type of analysis: Co-authorship
 Co-occurrence
 Citation
 Bibliographic coupling
 Co-citation

Unit of analysis: Authors
 Organizations
 Countries

Counting method: Full counting
 Fractional counting

VOSviewer thesaurus file (optional):

Ignore documents with a large number of authors
 Maximum number of authors per document: 25

< Back Next > Finish Cancel

Figure A17. Determination of the type and unit of analysis, counting methods, finally selection the maximum of authors per document.

Create Map

Choose thresholds

Minimum number of documents of an author: 1

Minimum number of citations of an author: 20

Of the 493 authors, 193 meet the thresholds.

< Back Next > Finish Cancel

Figure A18. Selection of the thresholds by defining the minimum number of documents and citations of an author.

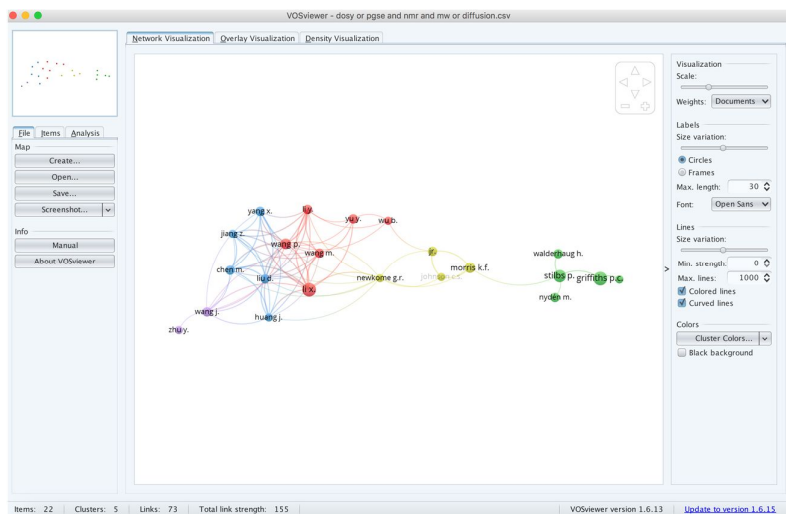


Figure A19. The primary visualized graph in VOSviewer.

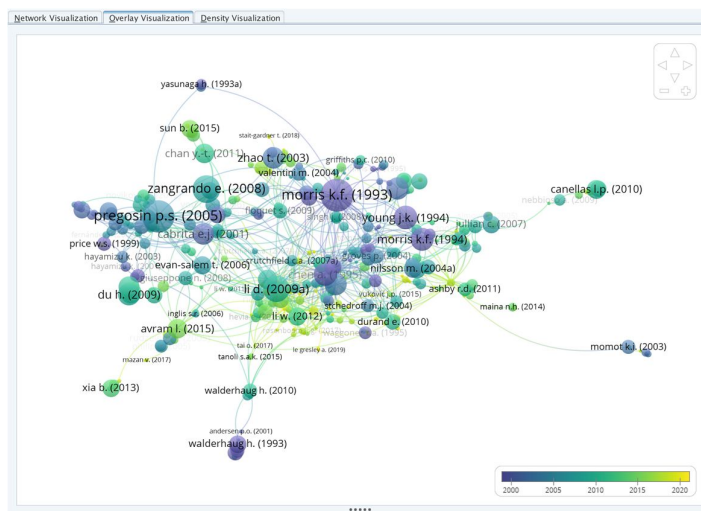


Figure A20a. The overlay visualized graph.

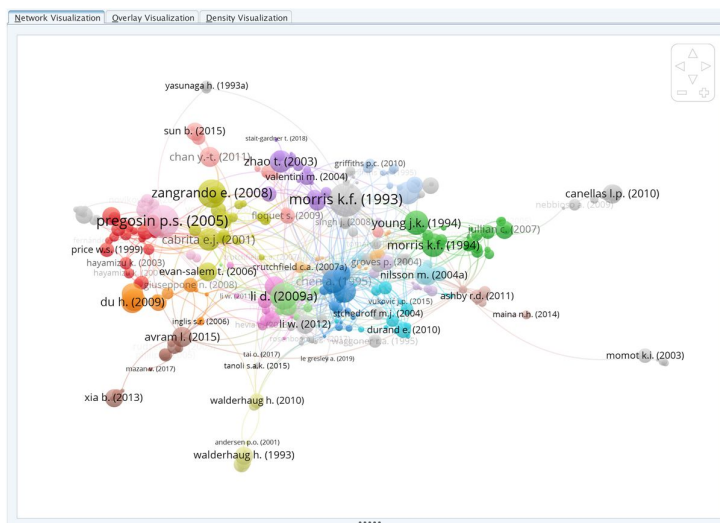


Figure A20b. The network visualized graph.

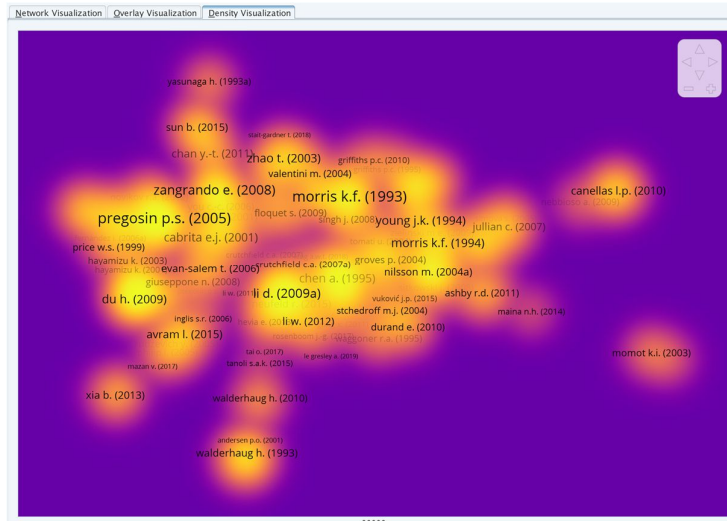


Figure A20c. The density visualized graph.



Figure A21. The aggrupation of items in various clusters.

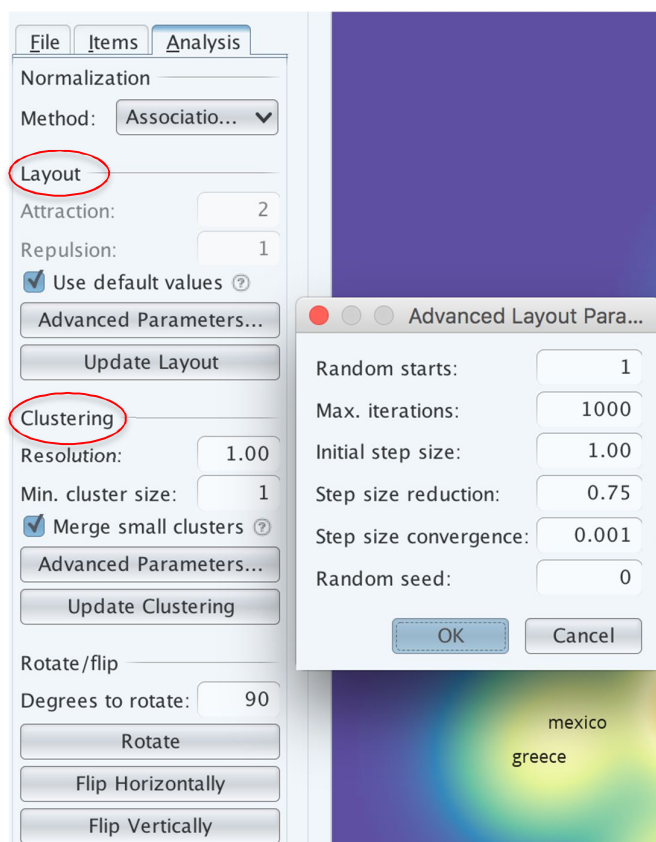


Figure A22. Parameters presented in the layout and clustering tools.

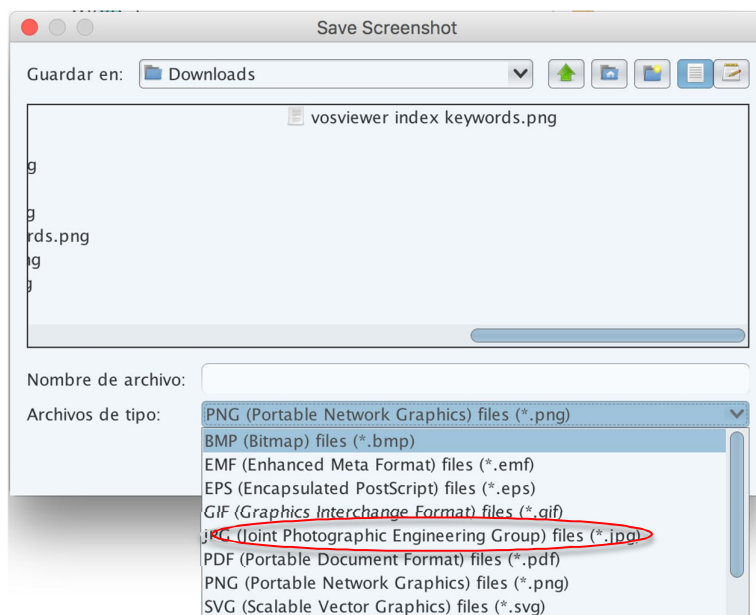


Figure A23. The exportation of the visualized graph.

Table A1. Clusters 1 to 4 based on index keywords.

Cluster 1	Links	Cluster 2	Links	Cluster 3	Links	Cluster 4	Links
Alpha helix	30	Adsorption	27	Block copolymers	28	Amides	40
Amino acids	59	Agglomeration	60	Chains	18	Bile acid	32
Binding affinity	29	Aqueous solution	77	Chromatography	39	Body fluids	44
Binding site	52	Aromatic compounds	29	Copolymerization	30	Catalysis	33
carbon	22	Asphaltenes	15	Crosslinking	36	Chlorine compounds	40
Carboxylic acids	32	Association reactions	55	Crystallization	14	Colloid	32
Chirality	39	Calorimetry	27	Glass transition	14	Critical micelle concentration	48
Circular dichroism	49	Chemical modification	22	Hydrolysis	27	Dimers	41
Complex formation	75	Esters	41	Lithium	26	Dynamic light scattering	32
Deuterium	23	Hydrophobicity	78	Molecular dynamics	102	Hydrophilicity	46
Fluid dynamics	40	Infrared spectroscopy	43	Nanoparticles	36	Kinetics	42
Glucose	21	Molecular interaction	84	Polyelectrolytes	20	Light scattering	52
hydrodynamics	76	Neutron scattering	31	Polyethylene oxides	46	Micelles	73
Hydrogen-ion concentration	46	Precipitation	25	Polymerization	71	Monomers	36
Metabolism	56	Sodium	51	Polymers	92	Porosity	14

Methodology	62	Surface property	51	Polystyrenes	29	Salts	42
Molecular recognition	46	Surfactant	42	Ring opening polymerization	26	Water	93
Peptides	30	Thermodynamic	44	Styrene	27		
Polysaccharide	24						
Proteins	24						
Protons	57						
Sodium dodecyl sulfate	34						
Stoichiometry	40						

Table A2. Clusters 5 to 8 based on index keywords.

Cluster 5	Links	Cluster 6	Links	Cluster 7	Links	Cluster 8	Links
Amines	28	Acetone	9	Anisotropy	19	Atomic force microscopy	37
Cation	26	Beta-cyclodextrins	28	Biological materials	26	Coordination reactions	21
Cell line, tumor	27	Binary mixtures	18	Carbohydrates	33	Crystallography	39
Copolymer	27	Covalent bond	21	Fluorine	26	Dimethyl sulfoxide	39
Cytology	40	Cyclodextrins	27	Gel penetration chromatography	22	Ionization	23
Degradation	26	Electrochemistry	17	Humic acid	29	Ligands	63

Dendrimers	31	Electrolytes	21	Hydrogen	64	Macrocyclic compounds	34
Dna	30	Hydrogen bonds	77	Isomers	23	Organometallic compounds	40
Drug delivery system	42	Ionic liquids	25	Molecular mechanics	27	Palladium	26
Encapsulation	31	Methanol	15	Organic acids	45	Pyridine derivative	37
Ethylene	40	Negative ions	15	Organic compounds	39	Pyridines	34
fluorescence	45	Positive ions	23	Organic polymers	33	Scanning electron microscopy	36
Hydrogels	26	Self-diffusion	16	Phenol derivative	34	Self-assembly	45
Macrogol	36	Viscosity	58	Semiconductor doping	19	Synthesis	83
Polyethylene glycols	52						