International Journal of Information Management

Information Management and Improvement of Citation Indices

Valentin Gomez-Jauregui [1,*], Cecilia Gomez-Jauregui [2], Cristina Manchado [3], Cesar Otero [4]

[1,*] *Corresponding author: EGICAD, School of Civil Engineering, University of Cantabria, 39005 Santander, Spain, Tel.: +34 942 206 757. E-mail: valen.gomez.jauregui@unican.es*
[2] *Panda Security. Dpt. Panda Research. Gran Vía. 48001 Bilbao. Spain, Email:cecigj@gmail.com*
[3] *EGICAD, School of Civil Engineering, University of Cantabria, 39005 Santander, Spain, Tel.: +34 942 206 757. E-mail: manchadoc@unican.es*
[4] *EGICAD, School of Civil Engineering, University of Cantabria, 39005 Santander, Spain, Tel.: +34 942 200 925. E-mail: oteroc@unican.es*

Abstract

Bibliometrics and citation analysis have become an important set of methods for library and information science, as well as an exceptional source of information and knowledge for many other areas. Their main sources are citation indices, which are bibliographic databases like Web of Science, Scopus, Google Scholar, etc. However, bibliographical databases lack perfection and standardization. There are several software tools that perform useful information management and bibliometric analysis importing data from them. A comparison has been carried out to identify which of them perform certain pre-processing tasks. Usually, they are not strong enough to detect all the duplications, mistakes, misspellings and variant names, leaving to the user the tedious and time-consuming task of correcting the data. Furthermore, some of them do not import datasets from different citation indices, but mainly from Web of Science (WoS).

A new software tool, called STICCI.eu (Software Tool for Improving and Converting Citation Indices - enhancing uniformity), which is freely available online, has been created to solve these problems. STICCI.eu is able to do conversions between bibliographical citation formats (WoS, Scopus, CSV, BibTex, RIS), correct the usual mistakes appearing in those databases, detect duplications, misspellings, etc., identify and transform the full or abbreviated titles of the journals, homogenize toponymical names of countries and relevant cities or regions and list the processed data in terms of the most cited authors, journals, references, etc.

Keywords

Information management; Bibliometrics; Citation indices; Data cleaning; Software

1. Introduction

1.1. Bibliometrics and citation analysis

Bibliometrics and citation analysis have become a very important set of methods for library and information science in the last four decades, as well as an exceptional source of information and knowledge for many other areas. However, they are not new fields, according to Weinberg, which is cited in Hood and Wilson (2001). He claims that the first Hebrew citation indexes date from about the 12th century; Sengupta (also cited by Hood and Wilson) states that the first bibliometric study was produced by Campbell in 1896. Bibliometrics (English equivalent of the term 'bibliometrie', coined by Paul Otlet in 1934), was defined by Pritchard (1969) as "the application of mathematics and statistical methods to books and other media of communication"; It is a set of methods to quantitatively analyze scientific and technological literature. It permits the exploration of the impact of any research field, the influence of a group of researchers or institutes, the impact of a certain publication or the quantitative research of academic outputs. Other closed and related concepts are scientometrics (concerned with the quantitative features and characteristics of science) and informetrics, which are "a recent extension of the traditional bibliometric analyses also to cover non-scholarly communities in which information is produced, communicated, and used" (Ingwersen & Christensen, 1997), or more briefly, "quantitative methods in library, documentation and information science" (Egghe & Rousseau, 1990).

Citation analysis deals with the examination of the documents cited by scholarly works . Its main application was originally information retrieval and analyzing its quality. However, for the last years it has also been used for bibliometrics, evolving to evaluating and mapping researches, measuring the production and dissemination of scientific knowledge, becoming progressively more significant for assigning funding or career development, and also for

establishing the journal impact factor. The main sources for citation analysis are citation indices, which are bibliographic databases that allows one to establish which later documents cite which earlier documents, which articles have been cited most frequently and who has cited them.

1.2. Citation indices and bibliographic databases

There are several citation indices, such as those published by Thomson Reuters' Institute for Scientific Information (ISI). Thomson Reuters' Web of Knowledge provides access to many bibliographic sources, such as MEDLINE and Web of Science (WoS), which collects many other citation indices. In total, they host a vast scholarly literature: 12,000 journals, 150,000 conference proceedings, 30,000 books and book chapters and 46 million records (Thomson Reuters, 2011).

Another important bibliographic database is SciVerse Scopus, published by Elsevier, which competes in completeness with WoS. Scopus claims to be the largest abstract and citation database of peer-reviewed research literature (Elsevier, 2011).

In comparison with these two citation indices by subscription, there are some other databases freely available: Google Scholar (with a vast information but with unreliable precision, as it will be discussed later), PubMed (specialized in life sciences and biomedical topics), CiteSeerX (the first automated citation indexing), Scirus (freely available search engine by Elsevier), getCITED (whose information is entered by members) or Microsoft Academic Search (academic search engine by Microsoft Research). Some other regional databases include SciELO, Dialnet, Latindex, etc. The fact of having multiple citation databases makes it necessary to compare them both from the scientometric and from the informetric points of view, by means of providing a set of measures for doing it systematically (Bar-Ilan, Levene, & Lin, 2007).

WoS and Scopus are the most reliable databases existing at the moment. In addition to their consistency, both citation indices offer a different selection of possibilities to export the records obtained by their respective search engines: plain text (.txt), tab-delimited (for Windows and Mac), comma separated values (CSV), web format (.html), BibTeX Bibliography Database (.bib), as well as for bibliographic management tools in Research Information Systems standardized format (.ris) like EndNote, Reference Manager, RefWorks, ProCite, etc. (Figure 1).

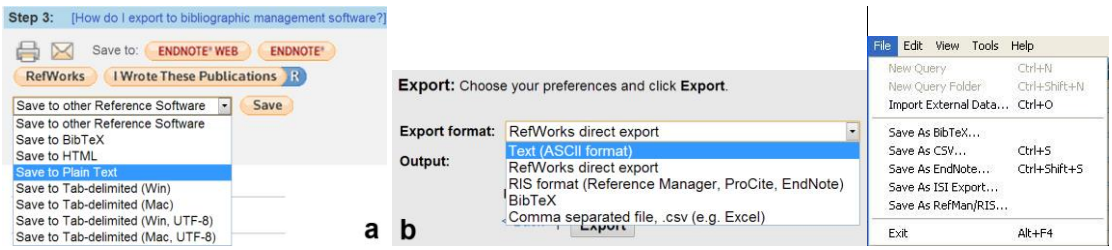


Fig. 1 Options for exporting citation datasets. a) From WoS. b) From Scopus. c) From Publish or Perish

Google Scholar (GS) and Microsoft Academic Search (MAS) cannot export their results directly, but only by means of Publish or Perish (Harzing 2007), a very useful citation analysis software program, but with some limitations for certain tasks. For instance, although Publish or Perish can export a dataset obtained from GS or MAS to several output formats (BibTeX, CSV, EndNote, ISI and RefMan/RIS), it cannot export the full citations included in each work (Figure 1.c).

1.3. Comparison between WoS, Scopus and GS

A comparison between WoS and Scopus shows that "Scopus includes a more expanded spectrum of journals (…), and its citation analysis is faster and includes more articles than the citation analysis of WoS" (Falagas, Pitsouni, Malietzis, & Pappas, 2008). Scopus claims a worldwide coverage with more than 50% of its documents coming from Europe, Latin America

and the Asian-Pacific Region (Elsevier, 2011) while WoS' contents are limited mainly to North America and Western European. Some other sources (Vieira & Gomes, 2009) confirm that Scopus provides the best coverage of social sciences literature, as well as for human-computer interaction literature, due to coverage of relevant ACM (Association for Computing Machinery) and IEEE (Institute of Electrical and Electronics Engineers) peer-reviewed conference proceedings. In addition, Scopus appears to have greater coverage of selected scientific areas, such as computer science, engineering, clinical medicine and biochemistry (Klavans & Boyack, 2007, Harzing 2010). Finally, Scopus displays full cited reference information (although many times not in the same order), unlike WoS, which only displays first author, year, journal title, volume, first page number and doi. Compared to both of them, GS is the best one at coverage, it is increasing rapidly and it is more successful at retrieving citations (Harzing 2010, Thornley et al. 2011).

There are several authors (Bornmann, Leydesdorff, Walch-Solimena, & Ettl, 2011; Vieira & Gomes, 2009) confirming that Scopus has more errors, misspellings, inconsistencies and name variants (e.g. Munchen, München, Munich), at least compared to WoS. According to Lopez-Piñero (1992), cited in Postigo Jiménez et al (2008), about 25% of the data obtained in the WoS were corrupted, while Postigo Jiménez et al (2008) states that, for instance, during their bibliometric studies this number was sensibly reduce to approx. 18%. Vieira & Gomes (2009), analyzing inconsistencies in addresses, dates, volumes and issues (not in authors, for instance), found a total of 72 errors in 1965 documents of WoS (4%) and 258 errors in 1979 documents of Scopus (13%). What's more, errors in citations can reach up to 50% depending on the journal, with a minimum rate of about 10% (Libmann, 2007). Related to the other main database, Thornley et al. (2011) state that "GS could be an impractical tool for author searching" and that it lacks accuracy in its date fields, which could be a severe limitation. Harzing (2007) also confirms that "some references contain mixed-up fields (…) because its

sources are inaccurate or difficult to parse automatically by Google's web crawler". The same author also confirms some other disadvantages of using GS: it includes some non-scholarly citations, not all scholarly journals are indexed in GS, its coverage might be uneven across different fields of study (e.g. the Natural and Health Sciences), it does not perform as well for older publications and GS automatic processing creates occasional nonsensical results (Harzing 2008).

In conclusion, even though the number of mistakes is being reduced along the years, there are still many inconsistencies and errors that must be corrected.

1.4. Pre-processing datasets

Thus, we should stress the importance of pre-processing the data imported from any bibliographic database, in order to avoid mistakes, misspellings and inconsistencies. By doing so, the results of any citation metrics or bibliometric analysis would be more accurate, realistic and valid. Moreover, the difference between doing it manually or with specialized tools can be significant in terms of rapidity, efficiency and precision, which are the main problems to be addressed in this work.

A key aspect of pre-processing the data is to clean it and to purge it. This task could only be categorized as part of the de-duplication process, which for each entity in a database either merges the identified duplicate records into one combined record, or removes some records from the database until it only contains a single record for each entity. Duplicate records must be detected and corrected because, for instance, they could incorrectly increase the number of appearances of one item in a dataset, even if they are meant to be the same one. For example, if a search is made for finding the works made by a certain author (e.g. "Carro Pérez, Consuelo") in two different databases (e.g. in Scopus with 15 records found and in WoS with 17 records), and 13 of her works appear in both datasets, it would be desirable to delete or

merge those 13 in the global combined dataset; the reason being that she would not be the author of 32 but of 19 publications and her h-index or g-index could be very different.

Another example, according to Thornley et al. (2011) would be the case of a certain work that could be found in several versions: draft, pre-print, post-print, as well as on different databases (WoS, Scopus, institutional repository, etc.) in the final published version. The process of detecting these duplicate records and considering them as being the same one is very important to obtain accurate results during the citation analysis. For example, if the freely available pre-print version of a research article has 15 citations and the post-print has only 3, it would not be accurate to limit the total impact of that work to only the three cites of its published version.

2. Existing Software for Bibliometric Analysis

There are several tools available for doing bibliometrics analysis. In the following, the most commonly used and most complete will be mentioned, as well as their strengths and weaknesses.

Different techniques and software tools have been proposed to carry out science mapping analysis running on bibliographic databases. Cobo et al. (2011) reviewed, analyzed, and compared some of these programs, taking into account aspects such as the bibliometric techniques available and the different kinds of analysis. According to them, the general workflow in a science mapping analysis has several steps: i) data retrieval, ii) pre-processing, iii) network extraction, iv) normalization, v) mapping, vi) analysis, vii) visualization and viii) interpretation. In their study, nine representative software tools specifically developed to analyze scientific domains by means of science mapping were taken into consideration: Bibexcel , CiteSpace II , CoPalRed, IN-SPIRE, Leydesdorff's software, Network Workbench Tool, Sci^2 Tool, VantagePoint and VOSViewer. Among the nine, only some of them were able to do

the pre-processing of the imported data, which could be split in four different methods: de-duplication (detecting duplicate and misspelling items), time slicing (dividing the data into different time subperiods, or time slices, to analyze the evolution of the research field under study), data reduction (selecting the most important data) and network reduction (selecting the most important nodes of the network of relationships between the units of analysis).

Among the nine software tools mentioned above, only four of them include de-duplication techniques (CoPalRed, Network Workbench Tool, Sci^2 Tool, VantagePoint). According to Cobo et al. (2011), CoPalRed can unify only keywords (lexical items, no authors, titles, sources, etc.) that represent the same concept. NWB removes duplicate records, detects and unifies duplicate items with different spelling. VantagePoint does the pre-processing and data cleaning by means of the Cleanup function, which attempts to identify the items that may be equivalent, performing fuzzy near matches on specific fields. Moreover, a list can be cleaned, applying a thesaurus.

One of the most complete programs, among the four, is the Science of Science Tool (Sci^2, version: 0.5 alpha) by Sci2 Team (2009), a modular toolset specifically designed for the study of science. It supports the temporal, geospatial, topical, and network analysis and visualization of datasets at the micro (individual), meso (local), and macro (global) levels. Data formats accepted by Sci^2 are “network formats, scientometrics formats and other formats” (Sci2 Team, 2009). Related to scientometrics, files from different types of databases can be imported: WoS, Scopus, Bibtex, Endnote and NSF. Its more relevant disadvantages are two: it is not possible to convert from one format to another, it just imports the data to be processed later in order to do the analysis; also, even though it uses some algorithms to detect duplicate nodes or records, there are a lot of limitations when performing that task. The algorithm produces three results: A 'Merge Table' that can be used to combine each set of duplicate nodes in a network into a single node; a report describing which nodes will be merged; and

another report describing which nodes will not be merged. The main problem is the difficulty of deciding and executing what records to keep as they are, which to combine or which format to keep while merging.

There are some other software tools that could be of interest for future researches (Chiang, 2009). HistCite (Thomson Reuters, 2012) only permits editing manually multiple records simultaneously to unify spelling variations in authors, addresses, or cited records. Publish or Perish (Harzing, 2007) is a software program that retrieves and analyzes academic citations from Google Scholar to obtain the raw citations and analyze them. DIVA (Morris, 2000) or Sitkis (Schildt, H.A., 2005), as well as many other tools created by Loet Leydesdorff (2010), are also able to import databases from WoS, and from Scopus in some occasions. However, all these tools leave for the user the tedious and time-consuming task of preparing the data, correcting repetitions, typos, mistakes and so on. Although they are good for analyzing the databases, they are not designed to deal with corrupt raw data.

Something similar happens to the CITAN**,** the CITation ANalysis package for R statistical computing environment. As it is stated, "CITAN implements experimental versions of the disambiguation search algorithms; they currently trigger too many "nuisance alarms" and cause the process of data cleaning to be time-consuming. Their improvement is left to the further research" (Gagolewski, 2011).

Of course, conventional data cleaning software tools also exist, but the fact that they are not oriented to bibliometrics purposes do not make them adequate for doing the required pre-processing of these kinds of databases.

One of the programs mentioned above, well known and widely used for co-citation analysis, is Sitkis (Schildt, H.A., 2005), which exports data from a certain dataset (only from WoS) into a Microsoft Access database. The program allows the imported data to be manipulated and

exported to different types of Ucinet-compatible networks and Microsoft Excel-compatible tables (Schildt & Mattsson, 2006). Unfortunately, the databases accepted by this tool are restricted to just one, WoS, not being compatible with some other important and significant citation indices. Resolving this issue by converting other databases to the WoS format style was one of the main motivations for developing the present work.

Related to conversions between format files of different bibliographic indexes, the authors have found only one set of tools (Scop2ISI, Scopus, Acc2ISI) that convert databases to other formats (e.g. WoS, Scopus, CSV file, Access, etc.), all of them developed by Loet Leydesdorff (2010). However, the format and accuracy of the output files are not perfect and, for instance, the tool Sitkis is not able to properly read a file converted by means of these tools from Scopus to WoS. The input format of the file obtained from Scopus is, in this case, a CSV file. Moreover, there is no cleaning or improvement of the quality of the data.

As a result of the analysis carried out by Cobo et al. (2011) and mentioned above, the same authors developed recently a new open-source software tool, SciMAT, which performs science mapping analysis within a longitudinal framework, carrying out all the steps of the science mapping workflow (Cobo et al. 2012). SciMAT (v1.1.01) implements a wide range of pre-processing tools such as detecting duplicate and misspelled items, time slicing, and data and network reduction.

Even though this software is useful and powerful, it lacks some performances that could be really interesting when carrying out the pre-processing. First of all, when doing a search for similarities among the different attributes or fields (author, title, source, etc.) it is not possible to visualize and select all the found similarities automatically at the same time; it is compulsory go through all the potential matches (candidate record pairs potentially referring to the same entity) one by one. This task can be very time-consuming and ineffective if the number of

items is very high. Furthermore, it is not possible to check occurrences in independent fields and at the same time having a look at the rest of the attributes of the record. For example, in Figure 2.a, when looking for similarities in the field Source independently, it is not possible to check what reference that source belongs to. The program gives the opportunity to change records as a full reference, watching the citation in its context, as can be seen in Figure 2.b; however, in this case it is not possible to change the fields independently, but only as a whole.

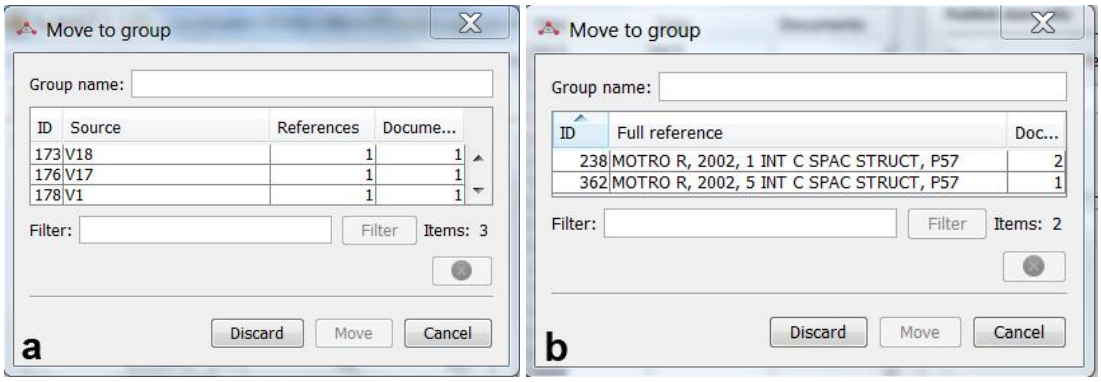


Fig. 2 Screens from SciMAT showing the possibilities to combine different records. a) For field Source. b) For full reference

Some other useful performances would be to deal with the sources, in case of journals, in a more complete and detailed manner, or to use a more appropriate string metric technique than the Levenshtein distance for measuring the difference between two sequences, as will be exposed in section 3.3. Another minor issue is that it does not properly import the characters with diacritical marks (accents, dieresis, tildes, etc.) from Scopus' RIS format files.

Finally, SciMAT is not able to convert files from a certain database type (Scopus, WoS, CSV) into another one; it just imports the data and process it internally. Output of the analysis, not the input data, is mainly done by means of reports in HTML or LATEX format, or graphically in PNG or SVG by means of a visualization module.

3. STICCI.eu: Features of the program

3.1. Motivations

As a result of all the issues exposed in previous sections, two main motivations are responsible for the creation of this free online software tool: widening the possibilities of data retrieval (by means of direct conversions between formats) and improving the pre-processing task (in order to perform a better and faster data cleaning and unification of records).

Firstly, the authors think that it is highly advisable being able to process the information from both WoS and Scopus citation indices. Thus, one of the main intentions of this work is to give it a wider scope and to improve the standardization of the data provided by both citation indices. As a secondary result, all the records hosted by Scopus will be able to be processed by all the programs that at the moment just do the data retrieval from WoS database. Likewise, it will also convert databases the other way around, from WoS to Scopus CSV, BibTex and RIS formats. As already said, several authors have developed many tools and software to organize, refine and process information only in WoS format (e.g. generating maps, visualization of international or institutional collaborations, analyzing word couplings, visualization of co-authorship networks, etc.), so the files generated by STICCI.eu could be easily processed by these programs.

Secondly, a very important and useful performance of this tool will be the possibility of correcting most of the errors, inconsistencies and misspellings appearing in the databases and homogenizing the information related to the same authors, titles of the documents or titles of the sources. As a consequence, the original data for the bibliometric analysis will not only be more complete but also more accurate and consistent. Thornley et al. (2011) recommend to import records into Excel for allowing manual addition, merging or deleting of works; their

work depends also on expert domain manually checking of search results; however, we think that this is not as effective as using a specialized tool for doing this task as well and as quick.

The result is a simple but useful program, called STICCI.eu (Software Tool for Improving and Converting Citation Indices - enhancing uniformity), able to perform not only the conversion from Scopus to WoS (and vice versa) but also the purge and standardization of the data incorporated in these two databases and others. It is important to note that the purpose of this tool is not to do bibliometric analysis, but to prepare datasets for being processed in such a way by other programs.

3.2. Conversion: Data retrieving and data exporting

The code of the program is divided in two main parts. The first one is a parser that converts the formats of each database into other formats (WoS, Scopus, CSV, BibTex, RIS); the second one corrects the mistakes and variants of the information, trying to standardize the data in order to have a consistent base for starting the analysis.

As has already been mentioned, the information of each database can be exported to several formats. WoS is able to export to html, plain text and tab delimited text with several variants (Figure 1.a). Scopus can generate files in plain text with ASCII format, RIS, BibTeX and CSV (Figure 1.b). Results from GS and MAS can be exported by the software Publish or Perish to several output formats: BibTeX, CSV, EndNote, ISI and RefMan/RIS (Figure 1.c). When designing STICCI.eu, it was decided to work with plain text format files (.txt), as well as BibTeX, RIS and CSV, in such a way that it was possible to process datasets generated by WoS, Scopus, GS and MAS.

When opening any of these txt archives, it can be easily noted that all the data contained on each dataset is structured differently, classifying the information under different fields and tags, such as title, authors, source of publication, affiliations, etc. Obviously, the names of the

fields of each database are different, even if they represent the same concept. Particularly, WoS and RIS formats use different tags for each field (AU for authors, SN for ISSN, etc.), while Scopus and BibTex include the whole set of words defining the field (e.g. AUTHOR KEYWORDS); Scopus even does not define anyone for the most important fields (authors, title and publication data). Although it is not straightforward to define the relationships between the fields of each database, the correspondence is understandable as presented on Table 1.

WoS	**Scopus**	**RIS**	**BibTex**	**Description**
PT		TY	type	Publication Type
AU	[Line 1]	AU / A1	author	Authors
TI	[Line 2]	TI / T1	title	DocumentTitle
ED	EDITORS	ED / A2	editor	Editors
SO	[Line 3-Item 2]	JF	journal	PublicationName
LA	LANGUAGE OF ORIGINAL DOCUMENT	N1	language	Language
DT	DOCUMENT TYPE	N1	document_type	DocumentType
SP	SPONSORS	N1	organization	Conference Sponsors
DE	AUTHOR KEYWORDS	KW	author_keywords	AuthorKeywords
AB	ABSTRACT	AB / N2	abstract	Abstract
C1	AFFILIATIONS	AD	affiliation	AuthorAddress
CR	REFERENCES	N1	references	CitedReferences
TC	[Line 3-7]	N1	note	Times Cited
PU	PUBLISHER	PB	publisher	Publisher
SN	ISSN	SN	issn	ISSN
RP	CORRESPONDENCE ADDRESS	N1	correspondence_address1	ReprintAddress
BN	ISBN	SN	isbn	ISBN
J9	ABBREVIATED SOURCE TITLE	JO	abbrev_source_title	29-Character SourceAbbreviation
PY	[Line 3- Item 1]	PY / Y1	year	YearPublished
VL	[Line 3- Item 3]	VL	volume	Volume
IS	[Line 3- Item 4]	IS	number	Issue
BP	[Line 3- Item 5]	SP	pages	Beginning Page
EP	[Line 3- Item 6]	EP	pages	Ending Page
DI	DOI	N1	doi	Digital ObjectIdentifier (DOI)
	[Line 4]	UR	url	URL

Table 1 Correspondence between the fields of WoS, Scopus, RIS and BibTex where only the most representative fields have been kept. Order corresponds to WoS database

Once this conversion step is processed, it is necessary to take into account the different formats of each exported file, as the use of tabulations, end-of-lines (EOL), separators (commas or semicolons), blank spaces, etc. differs in many citation indices in a noticeable manner. An example of an arbitrary record, as it is gathered in WoS and Scopus, is shown in Table 2, where it is possible to appreciate the important differences between both formats. It is also remarkable the difference in the amount of information gathered by each database, the order in which fields and citations are exposed, the use of diacritical marks, capital letters, etc.

In Scopus files, it is not unusual to find fields where the end-of-line (EOL) marker is incomplete, so two fields are presented together in the same line, as for instance:

LANGUAGE OF ORIGINAL DOCUMENT: EnglishABBREVIATED SOURCE TITLE: Int J Non Linear Mech

Another known issue in Scopus databases is that sometimes the field “CORRESPONDENCE ADDRESS” is included at the end of the last cited reference, instead of being in a separate line. STICCI.eu is able to correct this problem, because otherwise it would not be able to properly register the address of the corresponding author.

One of the most important divergences arises when trying to convert a citation, for instance, from Scopus to WoS. As can be seen in Table 2, citations in WoS use the abbreviated name of the journal, while Scopus writes the complete name. For this reason, STICCI.eu holds an updated version of all the abbreviated titles of the journals hosted by WoS in order to allow the conversion from full title to abbreviated title and vice versa, as will be exposed later. The user is the decision maker as to apply that conversion or to leave it as it is.

An interesting improvement of this software tool is that, during the parsing process, it splits the full citations into their different parts: cited authors, title, source, year, volume, issue, doi, etc. It allows the system to order these attributes depending on the format of the desired output file and, more importantly, it lets the user decide which will be the coincidences of

attributes (and not of the full citation) to take into account and the format and contents of these attributes once two or more records are merged.

SCOPUS	WEB OF SCIENCE
Gómez Jáuregui, V. *Habidite: industrialized modular dwellings [Habidite: viviendas modulares industrializadas]* *(2009) Informes de la Construccion, 61 (513), pp. 33-46.* *http://www.scopus.com/inward/record.url?eid=2-s2.0-67649255135&partnerID(...)* *AFFILIATIONS: Ingeniero de Caminos, Canales y Puertos. Trápaga, Vizcaya, Spain* *ABSTRACT: This paper is an introduction to (…)* *AUTHOR KEYWORDS: Construction; Industrialization; Integral; Modular; (…)* *REFERENCES: Ceballos-Lascuráin, H., (1973) La prefabricación y la vivienda en México, , Universidad Nacional Autónoma de México, Centro de Investigaciones Arquitectónicas;* *Aguiló Alonso, M., (1974) Prefabricación: Teoría y práctica, , Editores Técnicos Asociados, Barcelona;* *Collins, P., (2004) The Vision of a New Architecture, , Concrete. McGill-Queen's University Press, Montreal;* *Potter, E.T., (1890) Portable or Sectional Building, , U.S. Patent No. 425.250, 8 abril;* *Wisner, C.N., (1920) Improvements in and related to Concrete Buildings, , G.B. Patent No. 144.913, 24 junio;* *Witzel, J.R., (1920) Building Construction, , U.S. Patent No. 1.362.069, 14 diciembre;* *(…)* *CORRESPONDENCE ADDRESS: Gómez Jáuregui, V.; Ingeniero de Caminos, Canales y Puertos. Trápaga, Vizcaya, Spain; email: jauregui@gmail.com* *ISSN: 00200883* *DOI: 10.3989/ic.08.035* *LANGUAGE OF ORIGINAL DOCUMENT: Spanish* *ABBREVIATED SOURCE TITLE: Inf. Constr.* *DOCUMENT TYPE: Article* *SOURCE: Scopus*	*PT J* *AU Gomez-Jauregui, V* *AF Gomez-Jauregui, V.* *TI Habidite: industrialized modular dwellings* *SO INFORMES DE LA CONSTRUCCION* *LA Spanish* *DT Article* *DE Construction; Prefabrication; Industrialization; Modular; (…)* *AB This paper is an introduction to (…)* *EM valen.gomez.jauregui@gmail.com* *RI Gomez-Jauregui, Valentin/G-2696-2011* *CR ALONSO AM, 1974, PREFABRICACION TEORI* *ANSEDE M, 2007, CONTENEDORES HUMANOS, P34* *AUTOR S, 2007, CONSTRUCCION IND VIV, P118* *AVELLANEDA J, 2006, QUADERNS ARQUITECTUR, P84* *(…)* *NR 18* *TC 0* *Z9 0* *PU INST CIENCIAS CONSTRUCCION EDUARDO TORROJA* *PI MADRID* *PA SERRANO GALVACHE, 4, 28033 MADRID, SPAIN* *SN 0020-0883* *J9 INF CONSTR* *JI Inf. Constr.* *PY 2009* *VL 61* *IS 513* *BP 33* *EP 46* *DI 10.3989/ic.08.035* *PG 14* *WC Construction & Building Technology* *SC Construction & Building Technology* *GA 430GX* *UT WOS:000264978200004* *ER*

Table 2 The same publication gathered in two different databases: Left) Scopus format. Right) WoS format

3.3. Pre-processing: data cleaning, purge and standardization

One of the main features of this tool is the data cleaning by means of the similarities search. The data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from, in this case, a bibliographic database. The main tasks performed by this process are pattern recognition, duplicate detection, correction of typographical mistakes, standardization of name variants, fuzzy matching, etc.

The first step to implement is related to some known bugs appearing in the databases, especially in Scopus. It is not difficult to find some records where the initials of the middle names of the authors do not appear in capital letters. Most of the time, they are not actually middle names, but the typical super indices that refer to the affiliations of the citing authors. For instance, we can find in the authors field the following string: "Bradshaw, R.a , Campbell, D.b , Gargari, M.c , Mirrniran, A.d , Tripeny, P.e". In this case, a, b, c, d and e were used in the original paper as super indices to indicate affiliations (companies, institutes or universities where these researchers developed their activity). STICCI.eu detects these mistakes in the database and corrects them automatically.

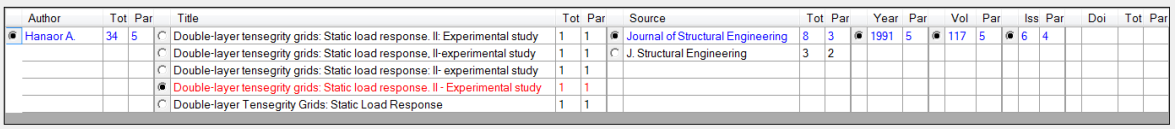
Multiple similarity functions have been developed to detect resemblances between different records that represent the same entity: Levenshtein, Affine Gap, Smith-Waterman, Jaro, Jaro-Winkler, Bigrams, Trigrams, Monge-Elkan, SoftTF-IDF, etc. In order to detect duplications, correct typographical mistakes and standardize the fields of the records, STICCI.eu uses the most appropriate procedure (Amón & Jiménez, 2010), based on the Smith-Waterman algorithm. It looks for all the records with a certain similarity in some fields: citing authors, cited authors, sources and/or full citations in WoS format (authors+title+source+year+volume). Some other tools use the Levenshtein distance between two strings, which is equal to the number of single-character edits required to change one string into the other, with the

allowable edit operations being insertion, deletion, or substitution of a single character. This technique could not be appropriate when dealing with strings of different lengths, thus, it is not used as the method for searching similarities. In contrast, STICCI.eu lets the user to select three parameters in order to customize that search:

- MS = Minimum Similarity (%): it defines the percentage of characters of two strings that must be equal (in type and position) to trigger the similarity flag.

- MLD = Maximum Length Difference (%): it defines the percentage of difference in length of two strings to trigger the similarity flag.

- NSL = Number of Starting Letters of a field concurring in different records: It limits the comparisons only between records whose fields share their first N number of letters (2 by default). Because the Smith–Waterman algorithm is fairly demanding on time (aligning two sequences of lengths m and n requires O(mn) time), this option reduces drastically the runtime for large datasets. If more accuracy is needed, this limit can be set to a lower amount of letters or it can be even set to zero in order to disable it.



Author	Tot	Par	Title	Tot	Par	Source	Tot	Par	Year	Par	Vol	Par	Iss	Par	Doi	Tot	Par
Hanaor A.	34	5	Double-layer tensegrity grids: Static load response. II: Experimental study	1	1	Journal of Structural Engineering	8	3	1991	5	117	5	6	4			
			Double-layer tensegrity grids: Static load response. II-experimental study	1	1	J. Structural Engineering	3	2									
			Double-layer tensegrity grids: Static load response. II- experimental study	1	1												
			Double-layer tensegrity grids: Static load response. II - Experimental study	1	1												
			Double-layer Tensegrity Grids: Static Load Response	1	1												

Fig. 3 Example of a group of similarities found for a certain dataset

If all or some of these fields of a record have similarities with any other record, this potential match will be shown to the user (Figure 3), who will specify, if desired (manually or automatically), which of the records refer to the same entity and which is the format to maintain when the merging is done. In order to facilitate this task, when STICCI.eu finds a

potential match between two or more records, it will specify how many times each of the different similar attribute values have been found for that certain search (partial scope, “Par” in Figure 3) and for the whole file of the database (total scope, “Tot” in Figure 3). If the user decides to apply an automatically generated matching, the format to maintain will be chosen following these criteria, in order of decision:

- the highest number of occurrences in the total scope
- the highest number of occurrences in the partial scope
- the maximum length of the string on that field
- the string with diacritical marks (accents, dieresis, tildes, etc.) or punctuation marks (colon, semicolon, dash, etc.)
- if no differences, the first occurrence.

Depending on this criteria, the system will alert the user about those similarities that are not evident enough, where the first two conditions do not give light to choose a certain option (because the total or partial number of occurrences is not evident or they are contradictory). In that case, the automatically selected field will be written in red, while in case the selection is made according to the first two criteria, the text will be shown in blue.

In the example of Figure 3, the field Title can be ambiguous, as there are five different occurrences, each one of them being found just once; thus, the chosen one (because it is the largest string) is written in red. The options chosen automatically in the rest of the fields are shown in blue because there are certain occurrences that appear more frequently in the partial or total scope. As a result, if the user decided to apply the automatic selection, the program would change all the five citations to: “Hanaor, A. (1991) Double-layer tensegrity grids: Static load response. II - Experimental study, Journal of Structural Engineering, V117 (6)”.

3.4. Standardized abbreviated and full titles of Journals

Another interesting characteristic of the program is that it gives the option of standardizing the names of the sources if they are indexed journals of public or customized databases. For instance, if in the field Source, a certain record has the name of a journal in an incorrect manner (e.g. “Int. J. of Solids and Structures”), it is possible to automatically substitute this attribute by the accepted name in that normalized list of journals. Thus, it would be changed to “INTERNATIONAL JOURNAL OF SOLIDS AND STRUCTURES” and would be marked in green and capital letters in order to easily highlight the match. Furthermore, as in some databases or for some purposes it is necessary to include the abbreviated title, it could also be changed to “INT J SOLIDS STRUCT” according to the needs of the user.

The program includes by default a vast list of journals belonging to the WoS repository, with more than 58.000 full and abbreviated titles of publications (Web of Science Help, 2011). Additionally, it supports the option for uploading other files (customized by the user or a segmented version, by subject categories, of that WoS list), in order to reduce the computational time required for searching into the referenced database of standardized journals.

3.5. Variants of toponymical names

Finally, STICCI.eu also offers the possibility of homogenizing variants of toponymical names (referring to countries and relevant cities or regions). Most of these inconsistencies are due to the different languages in which they are written or to orthographical symbols like accents or dieresis. For instance, if several records show the same city or country in different languages, like in the case of Munich, Munchen and München, or like in the case of Antwerp, Anvers and Amberes, the user will have the option of converting all of them to a unique one, in English, standardizing the addresses, in order to obtain a higher consistency on the final results of the

database. This option can be very useful when doing geographical analysis. Furthermore, the files that are used by the program to store the toponymical variants can be edited by the user, in order to include even more records or customize the names of the countries, cities and regions.

4. Workflow and characteristics of the program

4.1. Characteristics of the software

The final results of the features exposed above are gathered in a simple but effective software tool programmed in C# (Framework 2.0). This allows the tool to run on any version of the operating system Windows XP or later, and in 32 or 64-bit. There is no need of installation, it contains an executable file (.exe), a dynamic-link library (.dll) and a Microsoft Compiled HTML Help file (.chm). The whole packet is very light, occupying less than 3 Mb (without considering the datasets with the titles of the journals and the toponymical names that can be increased according to the user needs). It is freely available in www.sticci.eu (Gomez-Jauregui & Gomez-Jauregui, 2012).

4.2. Workflow and graphical interface

The final user of this tool will interact with the application by means of a friendly graphical interface. It works in only four sequential steps, which makes it easy and intuitive. Moreover, when passing the cursor over any button, without clicking it, a descriptive note or "hover box" with information about the functionality of that item appears automatically (tooltip). In addition, a help guide is available for further information that the user may need.

When launching the program, in Step 1, it just requires the user to select a file to work with. The system is able to read, in this version, four different types of text files (WoS, Scopus, RIS

and BibTex). After loading it, the program automatically identifies which of those is the original format of the file.

When accepting the loaded file, the user will find in Step 2 another simple interface (Figure 4) that will permit several operations to be selected that the user could accomplish, in a simple and intuitive manner. The different options and performances made by the software (explained above) could be selected or not depending on the preferences of the user: Search similarities by Citing author, Cited author, Journal title, Full citation or Toponymical variants.

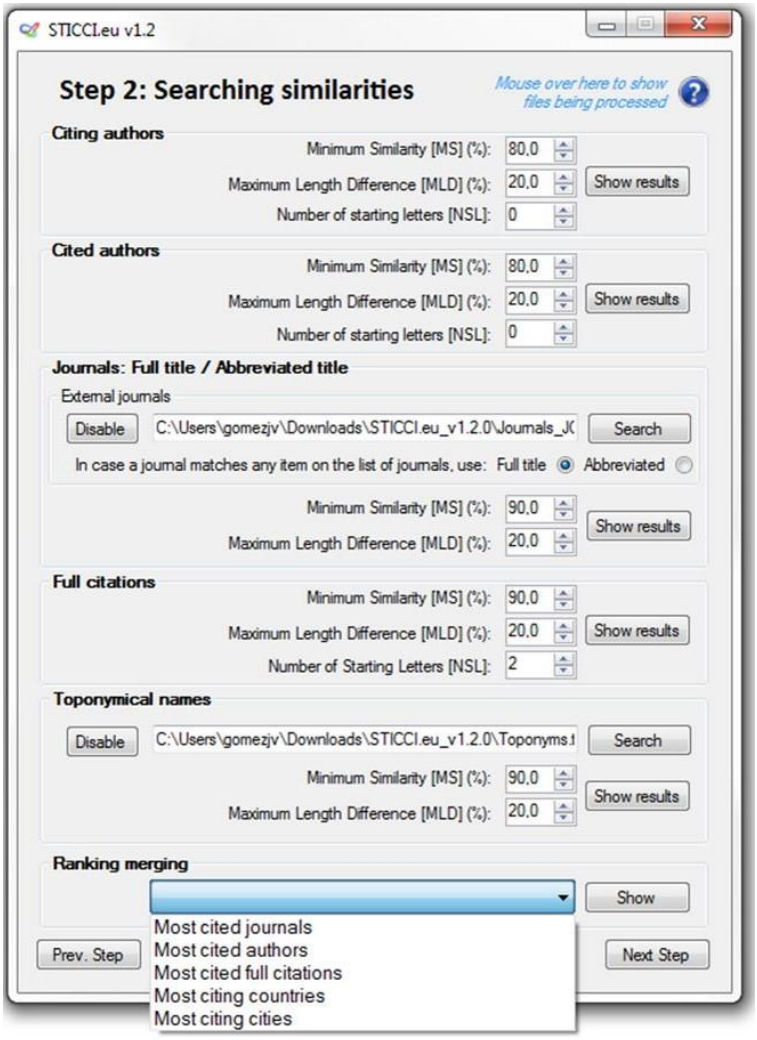


Fig. 4 Window for the Step 2 with all the options for searching similarities

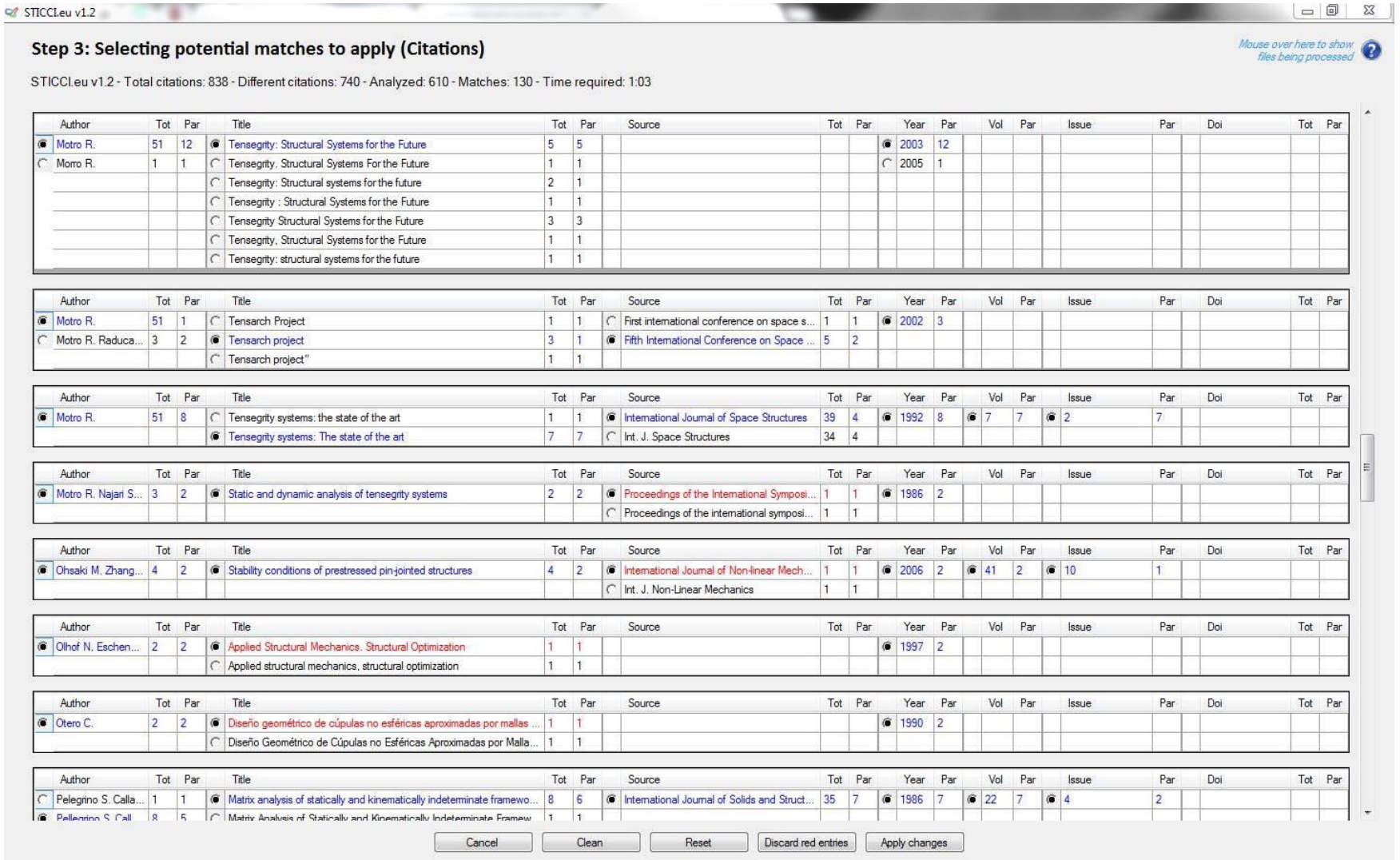


Fig. 5 Window for the Step 3 with all the similarities found and options of merging

Each one of those searches will lead to Step 3 (Figure 5), where the user will be able to define which similarities must be combined and in which style. In the particular case of the search of similarities among full citations, the selection could be independent for each different field (author, title, source, year, volume, issue and doi). As explained above, in this window, visual tools, by means of colors, will help the user to locate the most feasible similarities (in blue), and the more ambiguous (in red), depending on the partial and total counters of each potential match. Automatic selection of only the most plausible matches can be done by clicking the button “Discard red entries”.

After applying all the changes done automatically by the system and refined by the user, the interface of Step 2 is displayed again, where it is possible to do as many additional searches and combinations as desired. A powerful tool is the so-called “Ranking Merging”, where the user can see the ranking (listed by relevance or in alphabetical order) of the most common attributes (cited authors, journals, full citations, etc.) and merge those who can easily be detected as being the same one.

Once all the searches and selections are done and the user approves all the purging and merging applied during the previous process, in the final window, Step 4, it is possible to decide the output format of the clean and standardized dataset: WoS, Scopus, CSV, BibTex or RIS. When exporting to a Scopus file, even if the original file was also Scopus, order of the full citations would be changed if necessary to always follow the same structure: authors, year, title, source, volume, issue, page numbers and doi.

Finally, in addition to the output file, already clean, purged and corrected, the program also generates a log file with an excerpt of the corrections made, the merged records and the de-duplicated records.

4.3. Runtime and efficiency in searching

As an example of the performance of the program, a simple case has been carried out, related to a searching of similarities between the complete citations of a certain dataset with around 800 citations from 57 records. The dataset corresponds to the exporting of all the records where the words “Tensegrity” and “Grid” were found in the field “Article Title, Abstract, Keywords” in Scopus on the 12th July 2013. The following graphics show the times required for doing the analysis and the number of similarities found, depending on MLD (Maximum Length Difference), MS (Minimum Similarity) and NSL (Number of Starting Letters). Calculations were done on an Dell Inspiron 1750 with 2,13 GHz processor and 4GB RAM.

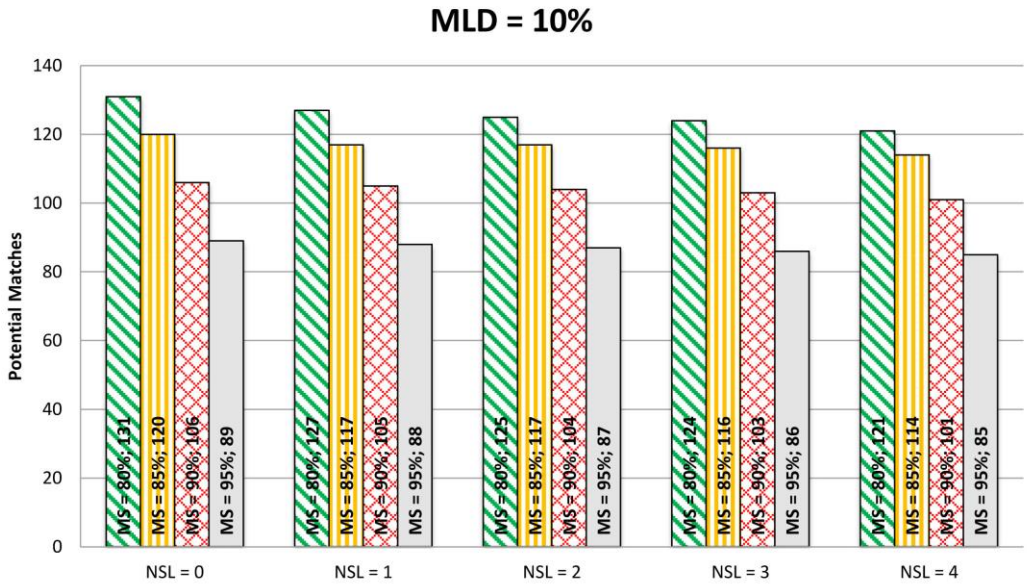


Fig. 6 Graph of potential matches for a fix value of MLD (10%) and different sets of NSL and MS

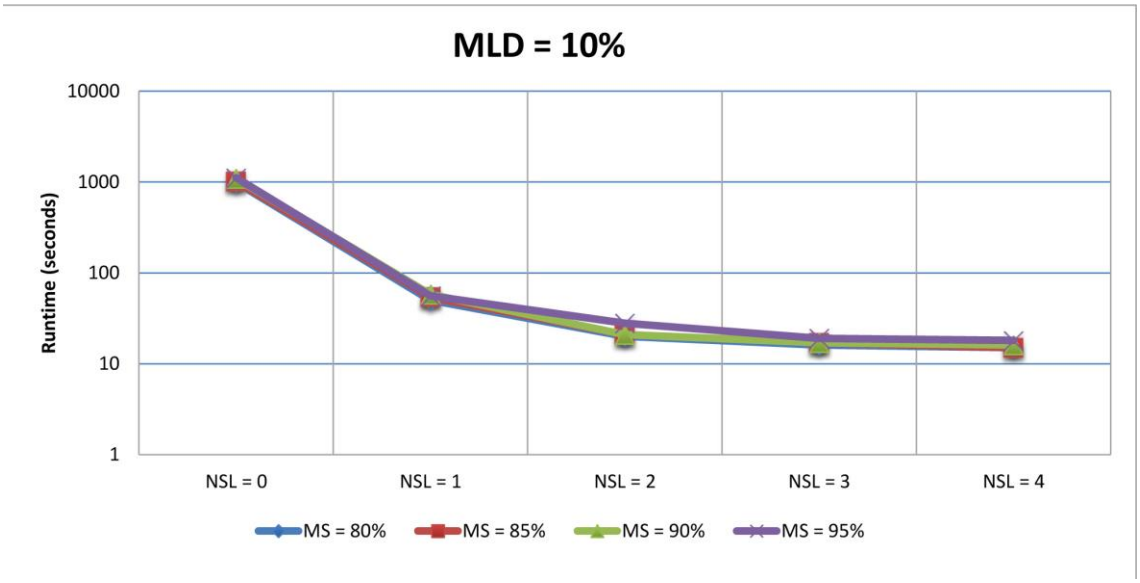


Fig. 7 Graph of Runtimes (in seconds, logarithmic scale) for a fix value of MLD (10%) and different sets of NSL and MS

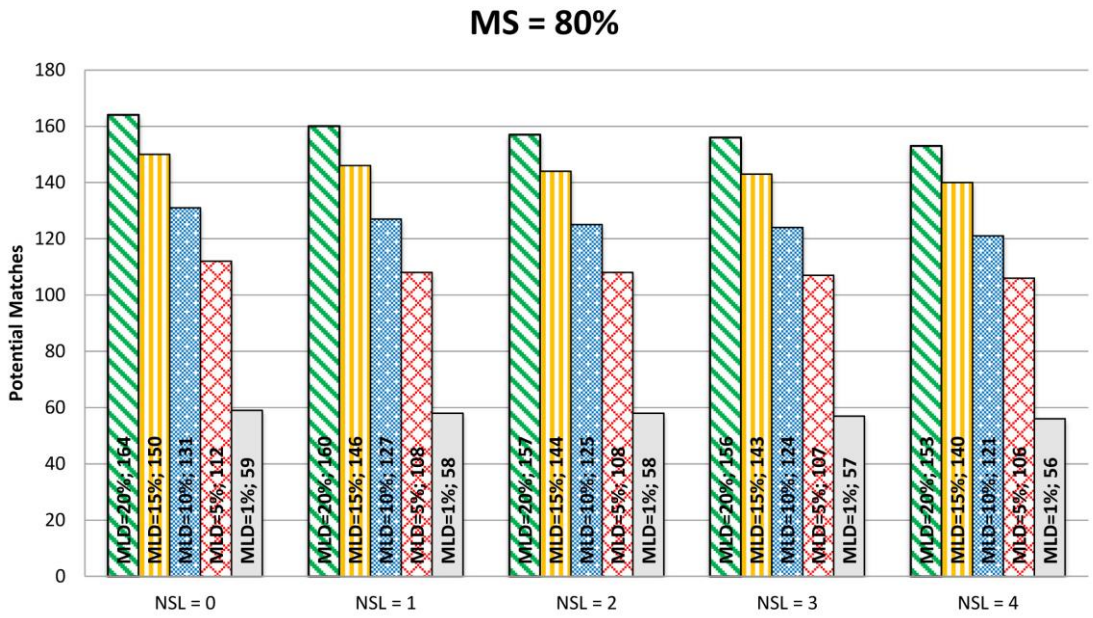


Fig. 8 Graph of Potential Matches for a fix value of MS (80%) and different sets of NSL and MLD

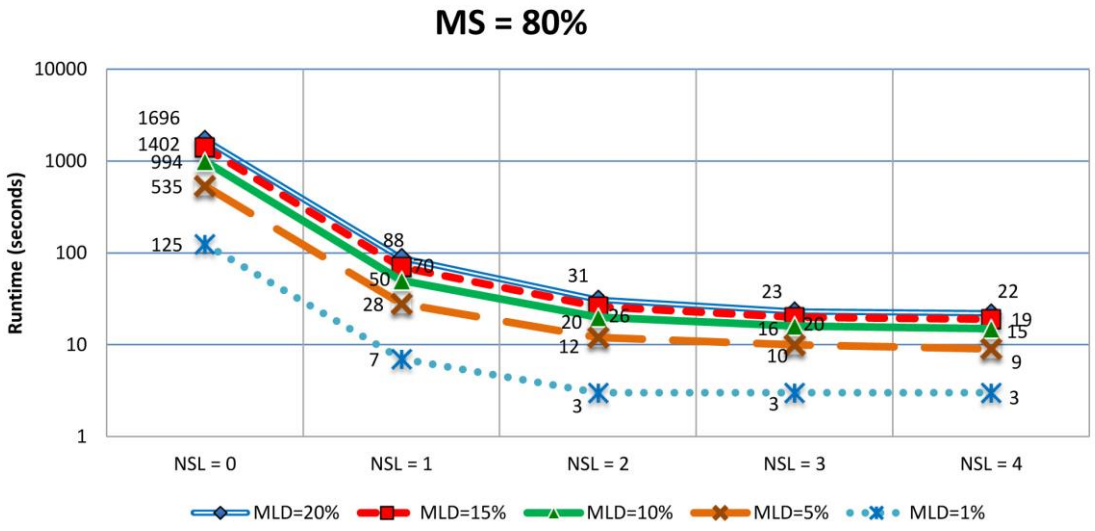


Fig. 9 Graph of Runtimes (in seconds, logarithmic scale) for a fix value of MS (80%) and different sets of NSL and MLD

Two mains conclusions can be easily deducted from these graphs. Related to computational time, Figure 9 shows that there is a certain difference in runtimes depending on both MLD and NSL factors, while Figure 7 illustrates that the MS does not affect it. In the same figure, when activating the option NSL for comparing the first N characters of the string, the times are comparatively much shorter when N is below 2. Furthermore, in terms of matching efficiency, all the factors matter when taking into consideration the potential mergings, but NSL does it in a much smaller scale, because the number of similarities found depending on this factor are not very different (Figure 6 and Figure 8).

As a consequence of both conclusions, the activating of NSL options is highly recommended, as it drastically reduces the runtimes while the risk of omitting potential matches is minor; however, decreasing the MS factor will find more potential matches without consuming more processing time.

5. Case study with a Scopus dataset

In order to exemplify the performances of STICCI.eu, a case study has been carried out with the same dataset used in the previous section (around 800 citations from 57 records where the words “Tensegrity” and “Grid” were found in the field “Article Title, Abstract, Keywords” in Scopus on the 12th July 2013). Two analyses are exposed in this paper: comparison of journal rankings and comparison of full citations. Citing authors and citing journals are excluded from this analysis due to its triviality (only 57 records, easily studied without additional software but just a datasheet).

For this experiment, three different software tools have been taken into consideration:

- Case 1 (SCIMAT): Importing the dataset with SciMat, from the RIS file.
- Case 2 (SCI2): Importing the dataset with Sci2, from the CSV file.
- Case 3 (STICCI.eu): Importing the dataset with STICCI.eu, from the txt format file.

The pre-processing with STICCI.eu (Case 3) will be done by applying all the pertinent possibilities of the program, that is, searching sequentially for matches related to citing authors, cited authors, cited journals and finally full citations, in which coincidences found in previous searches will be taken into account every time (toponymical names will not be studied in this example).

For the Citing Authors search, factors will be set to their default values (MS=80%, MLD=20%, NSL=0), which will find only two potential matches, both of them true matches (the two records in the pair correspond to the same entity). After applying these changes, the next search is carried out related to Cited Authors. In this case, with values of MS=80%, MLD=20%, NSL=0, the application finds 29 coincidences, from which only one of them is a false match (the two records of the pair refer to two different entities), easily detected due to their colors in

the results panel, and another one can be refined changing settings to MS=85%. Three more matches can be found by setting MLD to 75%.

After this, the Journals search is applied (MS=90%, MLD=20%), which finds that 45 journals are similar among our dataset and 63 journals are gathered in the list of JCR publications. Among them, seven were false matches and two of them were easily refined after changing settings to MS=95%. In this case, the option of converting these entries into full titles will be chosen. By doing that, a higher grade of standardizing will be applied to our dataset, which will make easier to find true matches in the search of Full Citations.

Finally, the search for Full Citations will be more straightforward because the Cited Authors and Sources (journals) have already been processed. Using the searching parameters MS=80%, MLD=20%, NSL=0, 157 matches are found (if NSL=1 the software founds only 4 matches less but the search is much faster). There are only six inaccurate matches (3,8% of false matches). After a second iteration with parameters MS=90%, MLD=20%, NSL=1, 19 more matches are found, all of them easily correcting the previous inaccurate matches. Finally, a quick manual merging is performed by means of the "Ranking Merging" tool. This is very useful for this purpose because it lets the user list the citations by relevance or in alphabetical order.

5.1. Comparison of Full Citations

A useful analysis is to know which are the most cited documents related to a certain subject, in order to locate the most influent works. The differences found between the three software tools are worth remarking upon. For the three of them, two cases are considered: A) Detection of matches among full citations after *parsing* the imported file. B) Detection of matches among full citations after *pre-processing* the imported file with the data cleaning tools provided by each software. The number of total citations differs among these tools due to the different kind of parsing executed by each one of them.

In case 1A (tool 1, case A), before doing the pre-processing, the SciMAT application finds a total of 808 citations, from which 762 are different (records appearing only once in that dataset) and there are 46 repetitions (records appearing twice or more), which means 5,7% of the total (see Case 1A in Table 3). The most cited document has only four cites (shown in Table 4, case 1A). This software allows doing a search for similarities among references, determining the maximum Levenshtein distance. Setting this value to 10, and after a long time consuming process (SciMAT shows all the matches one by one, while they are being found, which must be merged manually), the program discovers 57 potential matches, from which 55 are true matches (3,5% of false matches) that are set in 55 groups, while the rest of the other matches are false and must be discarded manually by the user. Unfortunately, there is no way of listing the original repetitions of references and the 55 reference groups together, so an additional step must be done by the user in order to obtain the combination of the most recurrent citations (shown in Table 4, case 1B). As can be seen, there is a certain improvement close to 16,5% on the repeated cites compared to the original dataset.

Citations	**Case 1A SciMAT**	**Case 1B SciMAT**	**Case 2A Sci2**	**Case 2B Sci2**	**Case 3A Sticci**	**Case 3B Sticci**
Total citations	808	808	807	807	798	798
Different citations	762	629	753	606	741	481
Repeated citations	46	179	54	201	57	317
Percentage of repeated citations	5,7%	22,2%	6,7%	24,9%	7,1%	39,7%
Improvement after pre-processing	16,5%		18,2%		32,6%	

Table 3 Comparison of the total, different and repeated citations before and after applying the different software tools

Sci2 gives better results than the precedent program. Analysis of the dataset with this software tool finds originally 807 citations too, from which 753 are different and 54 are repeated, that is a 6,7% (case 2A of Table 3), just a little better than in case 1. By using the tool “Detect Duplicate Nodes”, the program checks if any nodes in the network have labels that are very

similar to each other. Setting the similarity factor to 0,8 (80%) and the number of shared letters to zero, it finds 115 matches, 13 of them being inaccurate (11% are false matches). The impossibility of individually discarding the false matches forces the user to change the input values. After several trials, setting the similarity factor to 0,9 (90%), it finds 82 more duplicates, with only one of them being a false match (1,2% are false matches). It means an improvement of 18,2% on repeated citations after pre-processing, significantly higher than in case 1 (case 2B of Table 3). The system also detects some other potential matches (below similarity factor of 90%) that will not be considered and merged. After combining the 82 coincidences, the ranking of cites for this case 2B is shown in Table 4. As can be seen, results are better than in case 1, finding references with up to nine cites, over the maximum of seven cites found manually with SciMAT.

REF	**Full Citation**	**Case 1A SciMAT**	**Case 1B SciMAT**	**Case 2A Sci2**	**Case 2B Sci2**	**Case 3A Sticci**	**Case 3B Sticci**
Ref. 01	Motro R.;2003;Tensegrity: Structural Systems for the Future;;;;;;Kogan Page Science, London	2	6	3	8	2	20
Ref. 02	Pellegrino S. Calladine C.R.;1986;Matrix analysis of statically and kinematically indeterminate frameworks;Int. J. Solids Struct.;22;409-428;;;	1	7	2	6	4	11
Ref. 03	Quirant J. Kazi-Aoual M.N. Motro R.;2003;Designing tensegrity systems: The case of a double layer grid;ENGINEERING STRUCTURES;25;1121-1130;10.1016/S0141-0296(03)00021-X;;	2	4	2	7	3	10
Ref. 04	Motro R.;1992;Tensegrity systems: The state of the art;Int J Space Struct;7;75-82;;;	4	5	4	9	4	9
Ref. 05	Fuller R.B.;1975;Synergetics: Explorations in the Geometry of Thinking;;;;;;Collier Maccnillan Publishers. London	1	3	1	5	1	8
Ref. 06	Hanaor A.;1993;Double layer tensegrity grids as deployable structures;Int J Space Struct;8;135-143;;;	2	3	2	4	2	7
Ref. 07	Tibert A.G. Pellegrino S.;2003;Review of form-finding methods for tensegrity structures;Int J Space Struct;18;209-223;;;	1	3	1	7	1	7

Ref. 08	Kebiche K. Kazi-Aoual M.N. Motro R.;1999;Geometrical non-linear analysis of tensegrity systems;ENGINEERING STRUCTURES;21;864-876;;;	1	4	1	5	2	7
Ref. 09	Snelson K.;1973;Tensegrity Mast;;;;;;Bolinas Californie, Shelter Publications	1	3	1	4	1	7
Ref. 10	Motro R.;2002;Tensarch project;Fifth International Conference on Space Structures;;;;; University of Surrey, Guildford	1	2	1	2	1	7
Ref. 11	Fuller R.B.;1973;The Dymaxion World of Buckminster Fuller;;;;;;Anchoor Books, New York	1	3	2	6	1	7
Ref. 12	Schek H.J.;1974;The force density method for form finding and computation of general networks;COMPUTER METHODS IN APPLIED MECHANICS AND ENGINEERING;3;115-134;;;	1	2	1	3	1	6
Ref. 13	Emmerich D.G.;1988;Structures Tendues et Autotendantes;;;;;;Ecole d'architecture de Paris la Villette, Paris	1	2	2	2	1	6
Ref. 14	Hanaor A.;1992;Aspects of design of double-layer tensegrity domes;Int J Space Struct;7;101-113;;;	3	4	3	5	3	6
Ref. 15	Hanaor A.;1991;Double-layer tensegrity grids: Static load response. II - Experimental study;JOURNAL OF STRUCTURAL ENGINEERING-ASCE;117;1675-1684;;;	1	3	1	3	1	6
Ref. 16	Argyris J.H. Scharpf D.W.;1972;Large deflection analysis of prestressed networks;JOURNAL OF THE STRUCTURAL DIVISION-ASCE;106;633-654;;; ASCE	2	2	2	4	2	6
Ref. 17	Motro R.;1990;Tensegrity systems and geodesic domes;Int J Space Struct;5;341-351;;; England	3	3	3	6	3	6
Ref. 18	Hanaor A.;1993;Developments in tensegrity systems: An overview;Proceedings of the 4th Conference on Space Structures;;987-997;;;H. Nooshin. University of Surrey	1	2	1	2	1	5
Ref. 19	Pugh A.;1976;An introduction to tensegrity;;;;;;University of California Press, Berkeley	1	2	2	3	1	5
Ref. 20	Zhang J.Y. Ohsaki M.;2006;Adaptive force density method for form-finding problem of tensegrity structures;Int. J. Solids Struct.;43;5658-5673;10.1016/j.ijsolstr.2005.10.011;;	1	2	1	6	1	5

Table 4 Ranking of Full Citations, by number of appearances, of the different cases 1A to 3B (ordered from most repeated to less repeated in case 3B of STICCI.eu ranking)

In case 3, performed with STICCI.eu as shownabove, it is remarkable that, even without pre-processing, just by means of doing an efficient parsing, it is possible to improve the results of case 1 and 2 for the most cited documents (shown in Table 4, case 3A). The percentage of repeated citations before doing the pre-processing (7,1%) is better than with the other tools (case 3A of Table 3). Furthermore, after using STICCI.eu, fewer different citations are found (481), which means that 39,7% of the citations have been recognized as real matches. Thus, the searching of similarities is very effective as it is able to combine similar citations that refer to the same work. After the pre-processing with this application, the table of most cited references differs considerably, providing more citations to the most relevant works (up to 20), as can be seen in Table 4 (Case 3B), with an improvement on 32,6% on the recognition of repeated citations (case 3B of Table 3). As stated above, the ratio of false matches is only 3,8%.

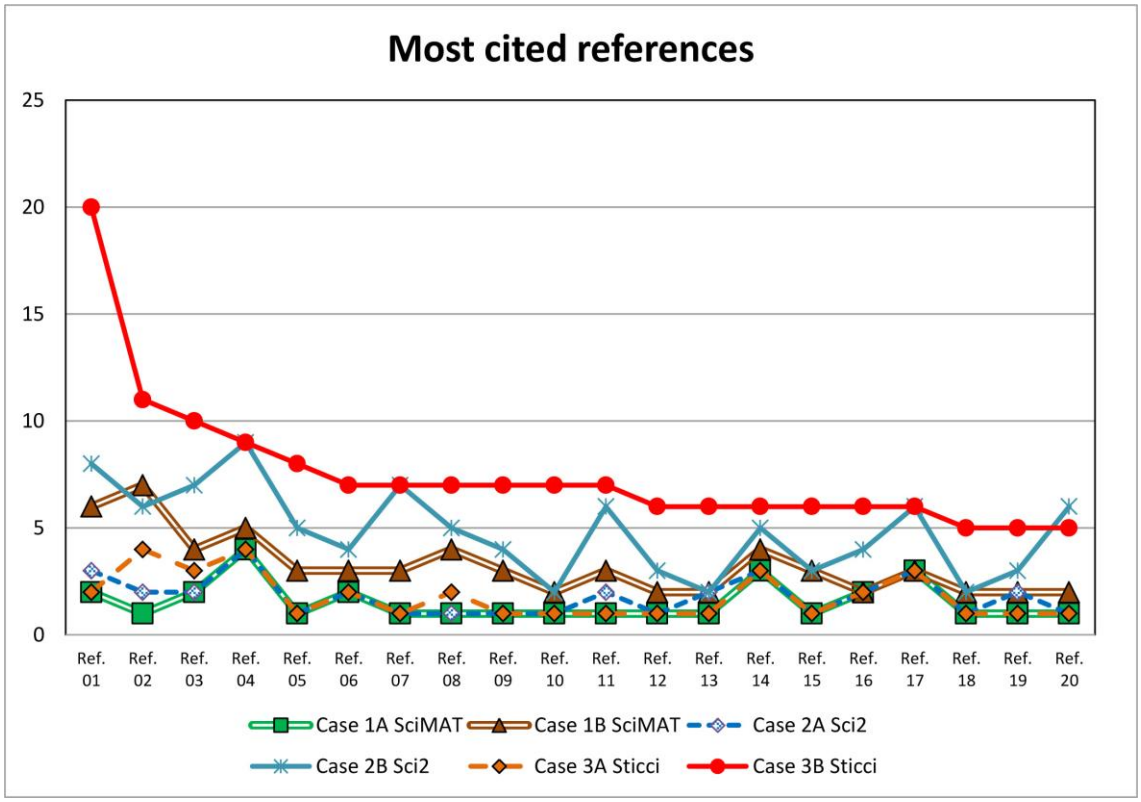


Fig. 10 Comparison in the ranking of the most cited documents before (case A) and after (case B) doing the pre-processing with the different software tools. Ref.01 to Ref.20 correspond to Full Citations of table 4.

In Figure 10, it is possible to compare the results exposed in this section in an illustrative graphic, showing the comparison of the ranking of references among the different applications and proving the better behavior of STICCI.eu above the other applications. Note that this figure, as well as the following ones, exposes data by means of a line graph for clarity, although the references of papers and journals are categorical variables.

5.2. Comparison of Cited Journals

Sometimes is necessary to find out which are the most cited journals related to a certain area of knowledge, which gives a clear sight about the most influent publications that are worthwhile consulting. Thus, it is necessary to gather information about the sources of the references and citations appearing in the bibliometric datasets, by means of a correct parsing and an efficient merging of those attributes.

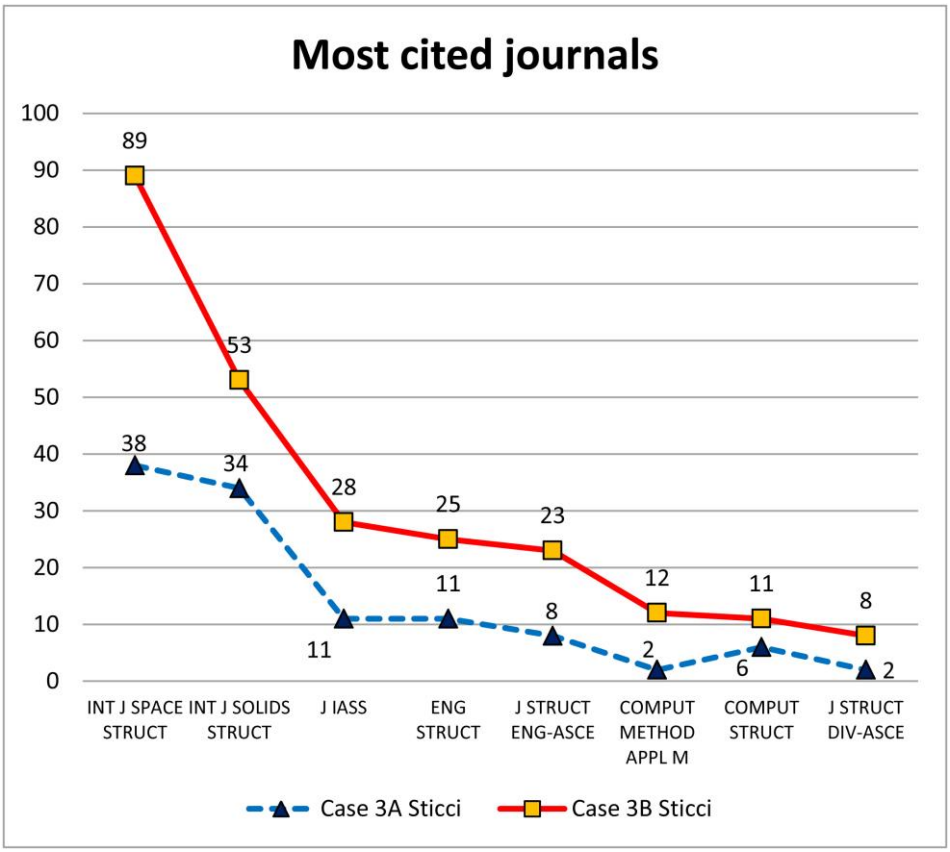


Fig. 11 Comparison in the ranking of most cited journals in the Scopus dataset before (case A) and after (case B) being processed with STICCI.eu

STICCI.eu allows the user to analyze the most cited journals even if they are provided by Scopus. So, in this case study, a comparison can be done between the ranking of the most cited journals before and after being processed by the application. As can be understood from Figure 11, improvement is huge, because even if at the beginning it could be interpreted that the first two journals of the ranking had a similar influence on this area, STICCI.eu helps to recognize that actually one journal is cited nearly twice as often as the other.

Software Sci2 performs a so-called "Merge Document Sources", but unfortunately it can only be applied to ISI files (exported from WoS), so it is not available for this case study. The same situation happens to SciMAT, which is only able to manage references coming from a WoS dataset. For this reason, this part of the case study will also be applied to another dataset obtained from WoS: 34 records where the words "Tensegrity" and "Grid" were found in the field "Topic" on the 4th June 2013).

For Sci2, the algorithm "Merge Document Sources" does not exist on the menu, but rather is run automatically when the ISI database is loaded into the tool. After loading the mentioned WoS file, no merging table is created to merge the identical journals. Besides, the command "Extract Document Source Citation Network" is not available for unknown reasons, so it is not possible to obtain processed data about the sources of the citations.

SciMAT originally finds a total of 656 sources, from which 378 are different and there are 278 repeated sources, i.e. 42,4%. However, this application provides a tool to search for similarities among sources of the references coming from a WoS dataset determining the maximum Levenshtein distance. By setting this value to one, a total of 10 possible matches are detected, from which only three of them are true matches (70% of false matches), which have to be validated manually one by one. Despite this, there are no remarkable changes in the final

journal ranking, as can be seen in Figure 12, where both lines Case 1A and 1B, representing before and after the preprocessing, are almost the same one.

Finally, the same exercise performed with STICCI.eu originally finds a total of 912 sources, from which 382 are different and there are 530 repeated sources, 58,11%, i.e. improving more than 15% the results of SciMAT. This is due to the fact that STICCI.eu does a better parsing and finds more sources and more duplicates among them. The journal search is applied with MS=90% and MLD=20%, which finds that 16 journals are similar among our dataset and 105 journals are gathered in the list of JCR publications. In this search, there were only eight false matches (7,5%). In this case, the option of converting these entries into abbreviated titles will be chosen because that is the format used in WoS for citations. After doing a quick manual merging by means of the "Ranking Merging" tool, the final results are shown in Figure 12.

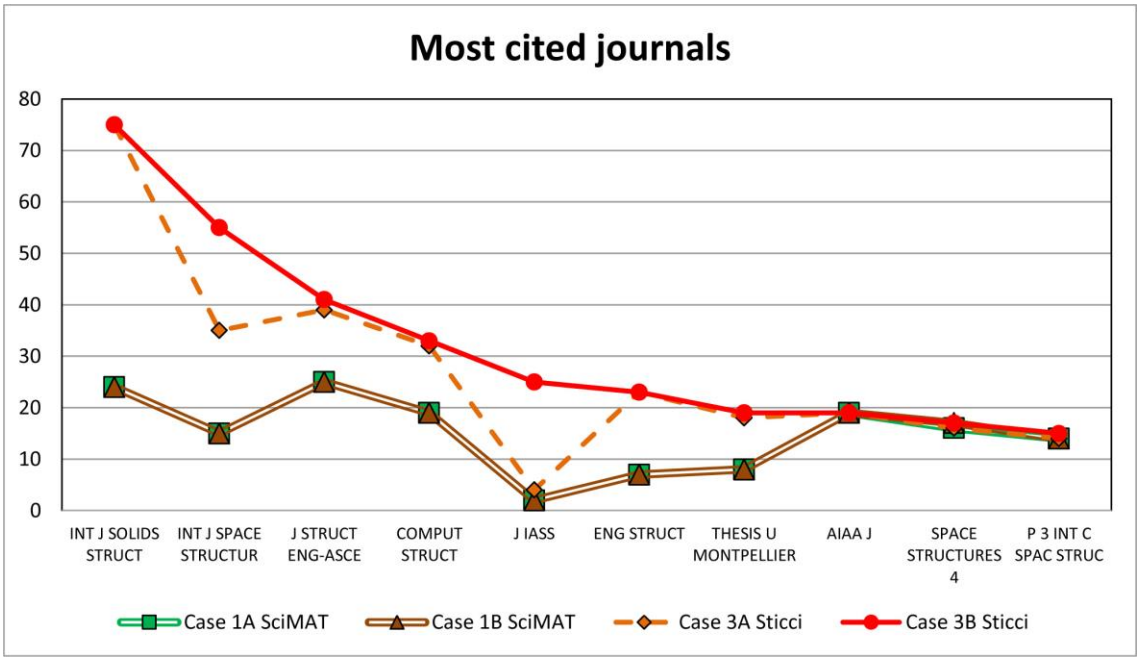


Figure 12. Comparison in the ranking of most cited journals in WoS dataset before (case A) and after (case B) being processed with SciMAT and STICCI.eu

It is remarkable that there is a much better performance of STICCI.eu upon SciMAT in both steps, parsing (case A) and merging (case B). Figure 12 shows that STICCI.eu obtains

continuous better results than SciMAT and there is almost always an improvement on the number of cited journals after the preprocessing.

6. Further developments

In order to show the advantages of the use of this tool, several case studies are being carried out at the moment. The aim is to prove the objective improvement of a certain dataset after being processed by STICCI.eu. Additionally, while the software is being used by different users, they are requested to fill a short questionnaire for validating STICCI.eu. The document itself is an adapted version of the User Interface Usability Evaluation with Web-Based Questionnaires (Perlman, 2011) and it is possible to fill it out online in the program webpage (www.sticci.eu).

The current version of the software, STICCI.eu v.1.2, is the first one released to the general public. In the future, the main objective will be to focus in the inclusion of some other existing and powerful databases (Google Scholar, PubMed, etc.)

Furthermore, it is always possible to increase the computational efficiency of the searching algorithm, as well as for the generation of the purged and standardized database.

Another important issue to be kept in mind is that the results will be written not only in the Scopus, WoS, RIS, BibTex text files and CSV, but also directly in a Microsoft Access file that could be processed without intermediate steps by other bibliometric tools like Sitkis.

The intention of the authors is to provide STICCI.eu for some other operative systems. Thus, the next step will be to develop it using an interpreted programming language, in such a way that it can be used with any platform (Windows, MacOS, Linux, etc.).

For these and other purposes, any suggestion, comment, report or collaboration made by the readers and users of this program will be gratefully welcome.

7. Conclusions

It has been clearly stated that the bibliographical databases lack perfection and standardization. There are several software tools that perform useful bibliometric analysis importing data from those corrupted databases. Some of them perform certain pre-processing tasks, but many times they are not strong enough to detect all the duplications, mistakes, misspellings and variant names. Besides, some of them are not able to import datasets from different citation indices, so they are limited to the information gathered by only some of them (mainly WoS).

The authors have created a new software tool, called STICCI.eu (Software Tool for Improving and Converting Citation Indices - enhancing uniformity), which is freely available online, to solve these problems in an effective and simple manner. Results obtained by this program have proven to enhance the efficiency in these tasks compared to other software tools. The summary of its features and advantages is the following:

- Converting between bibliographical citation formats (WoS, Scopus, CSV, BibTex and RIS)
- Correcting the usual mistakes appearing in those databases (especially Scopus, but also WoS)
- Performing a really deep and profound pre-processing task, taking into consideration the default automatic criteria or the selection of the user, combining and merging the fields that are the same record but with minor differences.
- Making the task of locating the similarities and the changes to be applied easier, before the merging is done (in a graphical window), and also after the conversion is done (by means of a log file with an excerpt of the corrections made, the merged records and the de-duplicated records).

- Using a procedure, based on the Smith-Waterman algorithm, for measuring the difference between two sequences, more appropriate than the Levenshtein distance, used by other applications, which is not as reliable for a string with 10 characters as for another one with 50.

- Detecting and transforming the complete or abbreviated titles of the journals, exporting all of them in one or the other format in a unified manner.

- Homogenizing toponymical variant of names, referring to countries and relevant cities or regions, although with the possibility for the user of adding more records in a customized manner.

- Listing in a straightforward manner the most cited authors, journals, full references, affiliations, geographical locations, etc., in a handy format that allows the user to do an accurate ranking, merging records that represent the same item.

References

Amón, I., & Jiménez, C. (2010). Funciones de Similitud sobre Cadenas de Texto: Una Comparación Basada en la Naturaleza de los Datos. Presented at the International Conference on Information Resources Managament (CONF-IRM2010), Montego Bay, Jamaica.

Bar-Ilan, J., Levene, M., & Lin, A. (2007). Some measures for comparing citation databases. *Journal of Informetrics*, *1*(1), 26–34. doi:10.1016/j.joi.2006.08.001

Bornmann, L., Leydesdorff, L., Walch-Solimena, C., & Ettl, C. (2011). Mapping excellence in the geography of science: An approach based on Scopus data. *Journal of Informetrics*, *5*(4), 537–546. doi:10.1016/j.joi.2011.05.005

Chiang, K. S. (2009). Tools, software - Translational Librarianship Consortium - Confluence. Retrieved December 27, 2012, from https://confluence.cornell.edu/display/TLC/Tools%2C+software

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, *62*(7), 1382–1402. doi:10.1002/asi.21525

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, *63*(8), 1609–1630. doi:10.1002/asi.22688

Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics : quantitative methods in library, documentation and information science*. Elsevier Science Publishers.

Elsevier. (2011). Content Coverage Guide of SciVerse Scopus. Retrieved October 7, 2012, from http://www.info.sciverse.com/UserFiles/sciverse_scopus_content_coverage_0.pdf

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *Faseb Journal*, *22*(2), 338–342. doi:10.1096/fj.07-9492LSF

Gagolewski, M. (2011). Bibliometric impact assessment with R and the CITAN package. *Journal of Informetrics*, *5*(4), 678–692. doi:10.1016/j.joi.2011.06.006

Gomez-Jauregui, V., & Gomez-Jauregui, C. (2012). STICCI.eu. Retrieved January 4, 2013, from http://www.sticci.eu/

Harzing, A. W. (2007). Publish or Perish. Retrieved January 3, 2013, from http://www.harzing.com/pop.htm#training

Harzing, A. W. (2008). Google Scholar - a new data source for citation analysis. Retrieved December 15, 2013, from http://www.harzing.com/pop_gs.htm

Harzing, A.W. (2010) Citation analysis across disciplines: The Impact of different data sources and citation metrics. Retrieved December 15, 2013, from http://www.harzing.com/data_metrics_comparison.htm

Hood, W., & Wilson, C. (2001). The Literature of Bibliometrics, Scientometrics, and Informetrics. *Scientometrics*, *52*(2), 291–314. doi:10.1023/A:1017919924342

Ingwersen, P., & Christensen, F. H. (1997). Data set isolation for bibliometric online analyses of research publications: Fundamental methodological issues. *Journal of the American Society for Information Science*, *48*(3), 205–217. doi:10.1002/(SICI)1097-4571(199703)48:3<205::AID-ASI3>3.0.CO;2-0

Klavans, R., & Boyack, K. W. (2007). *Is there a convergent structure of science? A comparison of maps using the ISI and Scopus databases*. (D. TorresSalinas & H. F. Moed, Eds.).

Leydesdorff, L. (2010). Software and data of Loet Leydesdorff. Retrieved December 25, 2012, from http://www.leydesdorff.net/software.htm

Libmann, F. (2007). Web of Science, Scopus, and Classical Online: Philosophies of Searching. *Online*, *31*(3), 36–40.

Morris, S. (2000). DIVA software / FrontPage. Retrieved December 25, 2012, from http://conceptsymbols.pbworks.com/w/page/16330153/FrontPage

Perlman, G. (2011). User Interface Usability Evaluation with Web-Based Questionnaires. Retrieved December 8, 2012, from http://hcibib.org/perlman/question.html

Postigo Jimenez, M. V., Díaz Casero, J. C., & Hernández Mogollón, R. (2008). Revisión de la literatura en fracaso empresarial: aproximación bibliométrica. In *Estableciendo puentes en*

una economía global. Presented at the Asociación Europea de Dirección y Economía de Empresa. Congreso Nacional, Salamanca.

Pritchard, A. (1969). Statistical Bibliography Or Bibliometrics. *Journal of Documentation*, *25*(4), 348–349.

Schildt, H. A., & Mattsson, J. T. (2006). A dense network sub-grouping algorithm for co-citation analysis and its implementation in the software tool Sitkis. *Scientometrics*, *67*(1), 143–163. doi:10.1556/Scient.67.2006.1.9

Schildt, H.A. (2005). SITKIS - A software tool for bibliometric analysis. Retrieved September 29, 2012, from http://users.tkk.fi/hschildt/sitkis/news.html

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

Thomson Reuters. (2011). Web of Science factsheet. Retrieved October 7, 2012, from http://thomsonreuters.com/content/science/pdf/Web_of_Science_factsheet.pdf

Thomson Reuters. (2012). HistCite. Retrieved January 3, 2013, from http://thomsonreuters.com/products_services/science/science_products/a-z/histcite/#tab1

Thornley, C. V., McLoughlin S. J., Johnson, A. C. & Smeaton, A. F. (2011). A bibliometric study of Video Retrieval Evaluation Benchmarking (TRECVid): A methodological analysis. *Journal of Information Science*. December 2011, vol. 37(6): pp. 577-593., first published on November 4, 2011

Vieira, E. S., & Gomes, J. A. N. F. (2009). A comparison of Scopus and Web of Science for a typical university. *Scientometrics*, *81*(2), 587–600. doi:10.1007/s11192-009-2178-0

Web of Science Help. (2011). Web of Science Journal Title Abbreviations. Retrieved November 27, 2012, from http://images.webofknowledge.com/WOK45/help/WOS/0-9_abrvjt.html