Further steps in integrating the platforms of WoS and Scopus: Historiography with HistCite™ and main-path analysis

Leydesdorff, L.; Thor, A.; Bornmann, L.

Published in:
El Profesional de la Información

DOI:
10.3145/epi.2017.jul.10

Citation for published version (APA):
FURTHER STEPS IN INTEGRATING THE PLATFORMS OF WoS AND SCOPUS: HISTORIOGRAPHY WITH HISTCITE™ AND MAIN-PATH ANALYSIS

Avances para integrar las plataformas WoS y Scopus: historiografía con HistCite™ y análisis de camino principal

Loet Leydesdorff, Andreas Thor and Lutz Bornmann

Loet Leydesdorff (Ph.D. Sociology, M.A. Philosophy, and M.Sc. Biochemistry) is professor emeritus at the Amsterdam School of Communications Research (ASCoR) of the University of Amsterdam. He is associate faculty at the Science and Technology Policy Research Unit (SPRU) of the University of Sussex, visiting professor of the Institute of Scientific and Technical Information of China (ISTIC) in Beijing, guest professor at Zhejiang University in Hangzhou, and visiting fellow at the School of Management, Birkbeck, University of London. He has published extensively in systems theory, social network analysis, scientometrics, and the sociology of innovation. With Henry Etzkowitz, he initiated a series of workshops, conferences, and special issues about the Triple Helix of University-Industry-Government Relations. Thomson Reuters/Clarivate lists him as a highly-cited author since 2014.

http://www.leydesdorff.net/list.htm
http://www.leydesdorff.net/th2/index.htm
http://highlycited.com
http://orcid.org/0000-0002-7835-3098

Andreas Thor is full professor at the Leipzig University of Telecommunications (HfTL), a private university owned by Deutsche Telekom. Before joining HfTL in 2013, he was an acting professor at the University of Passau and a visiting research scientist at the University of Maryland. In 2008, he received a PhD in Computer Science from the University of Leipzig where he worked as a researcher at the database group for several years. Andreas’ research interests are in the area of scalable data management. More specifically, he has been working on approaches for entity resolution, ontology alignment, and optimizing cloud technologies. He is regularly applying advanced data management techniques to research problems in the life sciences and bibliometrics.

http://orcid.org/0000-0003-2575-2893

Lutz Bornmann works as a sociologist of science at the Division for Science and Innovation Studies in the Administrative Headquarters of the Max Planck Society in Munich (Germany). His research interests include research evaluation, peer review, bibliometrics, and altmetrics. Since 2004, he has published more than 250 papers in journals covered by Clarivate Analytics Web of Science with a total of more than 4,000 citations. His publication list and citation metrics are available at:

http://www.researcherid.com/rid/A-3926-2008 Clarivate Analytics lists him among the most-highly cited researchers worldwide over the last ten years (since the first release of this service in 2014):
http://highlycited.com
http://orcid.org/0000-0003-0810-7091

University of Amsterdam, Amsterdam School of Communication Research (ASCoR)
PO Box 15793, 1001 NG Amsterdam, The Netherlands
loet@leydesdorff.net

University of Applied Sciences for Telecommunications Leipzig
Gustav-Freytag-Str. 43-45, 04277 Leipzig, Germany
thor@hft-leipzig.de

Administrative Headquarters of the Max Planck Society, Division for Science and Innovation Studies,
Hofgartenstr. 8, 80539 Munich, Germany
bornmann@gv.mpg.de

Manuscript received on 01-07-2017
Approved on 24-07-2017
Further steps in integrating the platforms of WoS and Scopus: Historiography with HistCite™ and main-path analysis

Abstract
The program HistCite™ enables an analyst to identify significant works on a given topic using the citation links between them diachronically. However, using Scopus data for drawing historiograms with HistCite™ has hitherto been a problem. In the new version of the program CREXplorer, one can translate citation data from Scopus to WoS formats (or vice versa) and then import the data into HistCite™. In this brief communication, we demonstrate these options using the papers of Eugene Garfield (1925-2017) in Scopus for main-path analysis. The two historiograms are considerably different: unlike the WoS set, the network connections between the time lines are sparse in the representation of the Scopus data; the secondary documents (e.g., editorials in Current Contents) not processed in Scopus, but included in WoS enrich the representation. Furthermore, HistCite™ has an option to export the citation network as a Pajek file that can be read by most network analysis and visualization programs. Garfield’s texts are centered by him as a personality and entrepreneur in different domains more than intellectually, although there are a number of recurring themes. Our own main paths are shaped along a line along which components indicate longer-term projects.

Keywords
Scopus; HistCite™; CREXplorer; Historiography; Eugene Garfield; Main path.

1. Introduction
Following up on his initial work with Irving Sher in the 1960s (Garfield; Sher; Torpie, 1964), Eugene Garfield joined forces with Alexander Pudovkin in the early 2000s to further develop a program for the historiography of science, namely HistCite™ (Garfield; Pudovkin; Istimin, 2003). HistCite™ reads records retrieved from the Web of Science (WoS) and generates a historiogram —that is, a visual representation of the historiography in terms of citation relations— based on the cited references among the papers in the sample. The program also allows for exporting the citation data in the network format of Pajek (Leydesdorff; Bornmann; Comins; Marx; Thor, 2016). The Pajek format has become a kind of currency among programs for network analysis and visualizations. In social network analysis, Hummon & Doreian (1989); Carley, Hummon, & Harty (1993) have developed main-path analysis with a similar objective (Liu; Lu, 2016; Batagelj, 2003; Batagelj; Doreian; Ferligoj; Kejzar, 2014; Lucio-Arias; Leydesdorff, 2008).

Hitherto, data downloaded from Scopus could not easily be imported into HistCite™, Pajek, and other network analysis programs because of incompatibilities in the format. Scopus has provided an alternative to WoS since 2006. Like WoS, Scopus contains citation information, but based on a larger set of journals (Leydesdorff; De-Moya-Anegón; De-Nooy, 2016). The standardization of cited references (but also the address information) in Scopus is different from that of WoS. In the most recent version of CREXplorer, however, the translation of Scopus’ citation data into the WoS format and vice versa is flawless. In this brief communication, we show the possibilities for translating data across platforms using Garfield’s œuvre as data; we compare the results with those based on our own publications.

The standardization of cited references (but also the address information) in Scopus is different from that of WoS.


El profesional de la información, 2017, julio-agosto, v. 26, n. 4. eISSN: 1699-2407 663
2. Eugene Garfield

Eugene Garfield (1925-2017) is best known for two major inventions: the Science Citation Index (Garfield, 1955; cf. Cronin; Atkins, 2000; Wouters, 2000) —later organized with similar citation indices into WoS— and the (2-year) journal impact factor (JIF) (Garfield, 1972; Garfield; Sher, 1963) —later extended to other journal indicators such as the 5-year JIF and the immediacy index (Price, 1978). In addition to these entrepreneurial initiatives, Garfield took a keen interest in developing the quantitative study of the sciences around the journal Scientometrics (Garfield, 1979; Price, 1978; cf. Glänzel; Schubert; Schlemmer, 2007) and by further developing tools for the mapping of the sciences (Small; Garfield, 1985) and their historical evolution (Elkan et al., 1978; Leydesdorff, 2010). He also served as President of the American Society for Information Science and Technology (1999-2000).

Garfield’s œuvre has been analyzed in a number of studies using HistCite™ and other routines (bornmann; Haunschild; Leydesdorff, in preparation; Leydesdorff, 2010; Wouters, 1999). Using data retrieved from WoS, Bornmann et al. (in preparation) note that 77.8% (n = 1,063) of Garfield’s 1,558 papers included in WoS were editorials in Current Contents. Another 148 papers were published in The Scientist, a professional magazine for scientists founded by Garfield in 1986. Garfield’s scholarly œuvre included in WoS contains the remaining 257 documents. Only 86 of these documents contain cited references. Documents which do not contain cited references can still be part of a citation network in their quality of being cited. Using Scopus, one retrieves approximately the same number of documents authored by Garfield as his scholarly contributions (n = 249).

An additional 2,541 documents are indicated as “secondary documents” these documents are cited in journals covered by Scopus, but not processed by Scopus as source documents of the database.

3. Import and export of data into and from CRExplorer

Data was downloaded from Scopus in the week of June 10, 2017. Table 1 provides summary statistics.

After exporting the retrieval (including the bibliographic references) from Scopus in csv-format, the Scopus output can be read into CRExplorer. By default, CRExplorer converts only those records which include cited references, but this can be changed in the Settings of the program (at File > Settings > Import/Export > Advanced Export Options > Include Publications without CRs in export). When CRExplorer writes the records into the WoS format, it modifies the header as follows: “FN Thomson Reuters Web of Science modified by CRExplorer”. The resulting file can be read into HistCite™ after changing this header into “FN ISI Export Format”.

CRExplorer was developed for disambiguation of the cited references (Thor; Marx; Leydesdorff; Bornmann, 2016a, 2016b). The program detects variants of the same cited reference, clusters them, and merges their occurrences. The clustering and merging of the data is especially important for Scopus data, because these are more heterogeneous than WoS data. The cited references contain more information in Scopus than WoS: all authors are listed, and the titles of the referenced publications are provided. However, these specifics increase the number of variants among the cited references. Furthermore, Scopus data sometimes contain fragmented cited references data which cannot be completely parsed into the bibliographic categories (for example when authors or volume numbers are missing in cited references).

The disambiguation of cited references is important for error control in citation analysis. Citation scores can thus be made more reliable. However, the usage of CRExplorer in this project is different: citation scores are attributed to units of analysis (e.g., authors, departments, or journals). This leads to attribute statistics and rankings. SPSS, for example, can be used for this type of analysis. Our usage in this study leads to network statistics using relations as units of analysis. From a graph-analytical perspective, disambiguation is less important because the relations among texts in a graph are unique. Using the disambiguation in CRExplorer, one risks generating cycles in the citation flows, as we shall see below.

Table 1. Summary statistics of the data used in the analysis.

<table>
<thead>
<tr>
<th>Author</th>
<th>Scopus N of documents (secondary documents)</th>
<th>WoS N of documents</th>
<th>Cited references after disambiguation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Garfield</td>
<td>249 (2,541)</td>
<td>1,558 (257 in œuvre)</td>
<td>1,317 952 20,900 15,195</td>
</tr>
<tr>
<td>Leydesdorff</td>
<td>349 (1,062)</td>
<td>314</td>
<td>11,657 5,308 6,112 5,901</td>
</tr>
<tr>
<td>Bornmann</td>
<td>255 (438)</td>
<td>242</td>
<td>6,909 3,917 4,195 4,096</td>
</tr>
</tbody>
</table>
Further steps in integrating the platforms of WoS and Scopus: Historiography with HistCite™ and main-path analysis

4. Use of data from CRExplorer in HistCite™

By default, HistCite™ generates a historiogram of the 30 papers with the highest citation scores within the set (the so-called “Local Citation Scores” or LCS). Within the drawing screen, one can also select “full” under Size for exporting the resulting graph as .png-file. A .png-file can be edited, for example, using InkScape, a freeware program that allows for embellishing the output. InkScape is available at https://inkscape.org/en

In Figure 1, for example, the central nodes were labeled (by us) on the basis of the legends to the HistCite™ output. Figure 1 shows the historiogram of Garfield’s work on the basis of Scopus data. Although the number of cited references is reduced by disambiguation from 1,317 to 952, the resulting historiogram is precisely the same as the one based on all cited references without disambiguation. Figure 2 shows the corresponding historiogram on the basis of the much larger set of WoS data.

The two historiograms are considerably different. Both originate from Garfield’s (1955) initial publication in Science proposing the Science Citation Index. Garfield’s (1972) article in Science introducing the JIF is indicated in both historiograms as well. Unlike the WoS set, however, the networked connections between the time lines are sparse in the representation of the Scopus data. The secondary documents (e.g.,
editorials in *Current Contents*) not processed in *Scopus*, but included in *WoS* enrich the representation.

5. Export from *HistCite™* to *Pajek*: main-path analysis

*HistCite™* graphs can be exported in the *Pajek* format. By default the histiogram shown on the screen, is exported. However, one can untick the limit in the left-top corner of the pane and ask for the full set (Figure 3). Even in the case of moderately large sets such as the ones under study above (in section 3), *HistCite™* fails to generate a graph, but one nevertheless can export the full set in the network format.

Within *Pajek* (or a similar program) one can pursue network analysis and visualization. Using the largest component of networked papers (*n* = 61) within the *Scopus* set (*n* = 249), one can, for example, reconstruct the so-called “main path.” Main-path analysis calculates the extent to which a particular citation is needed for linking articles as a transversal weight divided by the total number of paths between a citing and cited document in the citation network (*De-Nooy et al.*, pp. 281 ff.). Since the analysis of the main path was originally proposed by *Hummon & Doreian* (1989); *Carley, Hummon, & Harty* (1993), a large number of variants of main-path analysis have been elaborated (*Batagelj et al.*, 2014; *Liu; Lu*, 2012). For example, one can also count transversal weights of nodes (instead of links).

Figure 4, for example shows the standard (global) main paths for this set after weighing the citations using the SPC (search path counts) algorithm. We used *VOSviewer* for the visualization of the *Pajek* output. The differences with other methods of main-path analysis are often marginal,

---

Figure 3. Unlimited export of *HistCite™* results into *Pajek* format

---


Figure 4. The main path in the largest component (*N* = 61) of Garfield’s œuvre based on *Scopus* data; two components distinguished (*Q* = 0.279); *VOSviewer* used for the layout and visualization.
Further steps in integrating the platforms of WoS and Scopus: Historiography with HistCite™ and main-path analysis

and discussing them would lead us beyond the scope of this communication.

In accordance with the strongly decentralized patterns in this set, the main path is indicated by only six papers (among the 249 in total). Two components are distinguished when using a community-finding algorithm ($Q = 0.28$; Blondel et al., 2008): the early part is oriented to the shaping and legitimation of citation analysis, and the later part (since 1998) is reflexive on the use of citation analysis in research and research evaluations. In summary, Garfield’s texts are centered by him as a personality and entrepreneur more than intellectually, although there are a number of recurring themes. The largest component in the citation network contains only 61 of the 249 (24.5%) documents (without disambiguation of the cited references, the largest component contains 60 documents.)

As a further exploration, Figures 5 and 6 show similarly constructed main paths among the publications of two of the authors of this paper: Bornmann and Leydesdorff, respectively. Both main paths show a flow of related and overlapping research lines evolving over time. The largest components contain 92.4% and 95.07% of the sets, respectively, as against 24.5% for Garfield. The linear sequencing within both sets makes it possible to distinguish components indi-

Figure 5. 25 documents on the main path among 233 documents (co-)authored by Bornmann and forming the largest component among 255 documents retrieved from Scopus ($Q = 0.69$).

Figure 6. 34 documents on the main path among 326 documents (co-)authored by Leydesdorff and forming the largest component among 349 documents retrieved from Scopus ($Q = 0.71$).
Citing periods with a modularity of 0.69 and 0.71, respectively, as against 0.28 for the star-shaped pattern in the case of Garfield’s œuvre.

In Figure 6, documents #84 and #140 are marked with a hash, indicating that these references represent cycles which were shrunk (in Pajek; Batagelj et al., 2014). These cycles are generated by the disambiguation in CRExplorer. Without disambiguation, however, the citation network is acyclic, the largest component in the set is 326, and the main path contains 33 (instead of 34) documents. In other words, the two functions of CRExplorer (disambiguation and translation between Scopus and WoS) serve different purposes and should not be confused. The overall purpose of CRExplorer is data analysis which is supported by import/export formats as well as data cleaning and transformation routines (e.g., disambiguation, filtering/deleting, etc.).

CRExplorer was developed for disambiguation of the cited references, detecting variants of the same cited reference, clustering them, and merging their occurrences.

6. Limitations

One major limitation of using HistCite™ for generating network data is that the exported graphs are non-valued; each link is counted as one. In citation analysis, one may wish to work with valued networks. The same data in WoS format — either taken directly from WoS or processed via CRExplorer retrieved from Scopus — can alternatively be used as input to CitNetw.exe for generating a valued graph. Available at: http://www.leydesdorff.net/software/citnetw

In the current version of CRExplorer, the translation of address information from Scopus into the WoS format is not yet fully reliable. We expect to improve this translation further in a next version of CRExplorer.

7. Conclusion

We have demonstrated how the new option in CRExplorer enables users to cross the divide between Scopus and WoS. The possibility to use both WoS and Scopus data is relevant for users who have access to only one of the databases, but may also be important for studying the reliability of results, the sensitivity to data choices (Leydesdorff; De-Moya-Anegón; De-Noye, 2016), and the exchange and reproducibility of results among researchers (e.g., Leydesdorff; Wagner; Bornmann, 2016; Vilchez-Román, 2017).

Using the data for three researchers, we have shown that one of them (Garfield) entertained a star-shaped citation network. The main path was short: from one cluster including important papers in the early stage of the author’s career to another containing more reflexive and programmatic publications. Bornmann and Leydesdorff’s main paths tend to be shaped along a line leading to a much higher modularity (Q ~ 0.7 instead of 0.3). The main path organizes the majority of the papers along a line connecting longer-term research projects that can be visualized as components. However, Garfield’s attention is distributed given his roles as an entrepreneur, a scholar, a journalist and editor, and a science-policy maker.

In our opinion, the relative ease with which the network characteristics can be retrieved makes it interesting to add this second-order perspective of network analysis in terms of links to the analysis and ranking of nodes in terms of attributes such as citation scores or the h-index values. The unit of analysis is then the communication as different from the communicators (Luhmann, 1996).

By default, HistCite™ generates a histogram of the 30 papers with the highest citation scores within a set (the so-called “Local Citation Scores” or LCS)

Notes

1. HistCite™ is freely available upon registration at http://ip-science.thomsonreuters.com/thanks/HistCite
2. Pajek is a network analysis and visualization program, freely available for non-commercial usage at http://mrvar.fdv.uni-lj.si/pajek
3. CRExplorer is freely available at http://www.crexplorer.net
4. CitNetExplorer is software allowing for similar reconstructions on the basis of citation data. The program is freely available at http://www.citnetexplorer.nl
6. Secondary documents in Scopus are cited, but not processed as source documents (“citing”).
7. An alternative to HistCite™ is provided by CitNetExplorer at http://www.citnetexplorer.nl
8. SPC is one among the algorithms which can be used for main path analysis (Batagelj et al., 2014, pp. 76 ff.). Three algorithms are included in Pajek (Batagelj, 2003): (i) Search Path Counts (SPC) accounts for the highest traversal weight of the arcs connecting a source with a sink; (ii) the Search Path Link Count (SPLC) accounts for the number of all possible search paths through the network emanating from an origin; (iii) the Search Path Node Pair (SPNC) accounts for all connected vertex pairs along the paths (Hummon; Doreian, 1989, pp. 50-51).
9. VOSviewer is a network visualization program available at http://www.vosviewer.com
Further steps in integrating the platforms of WoS and Scopus: Historiography with HistCite™ and main-path analysis

Since the analysis of the main path was originally proposed by Hummon & Doreian (1989); Carley, Hummon, & Harty (1993), a large number of variants of main-path analysis have been elaborated.

8. References


Bornmann, Lutz; Haunschild, Robin; Leydesdorff, Loet (in preparation). Reference Publication Year Spectroscopy (RPYS) of Eugene Garfield’s publications. https://doi.org/10.7717/peerj-cs.32


Leydesdorff, Loet (2004b). “Top-down decomposition of the Journal Citation Reports of the Social Sciences Citation Index: Graph- and factor-analytical approaches”. Scientometrics, v. 60, n. 2, pp. 159-180. https://goo.gl/RKAM2f

Loet Leydesdorff, Andreas Thor, and Lutz Bornmann

https://goo.gl/7V3xdq


https://doi.org/10.1016/j.joi.2016.02.005


https://doi.org/10.1007/s11192-016-2082-3


https://www.crodoc.es

Cronología de la Documentación Española