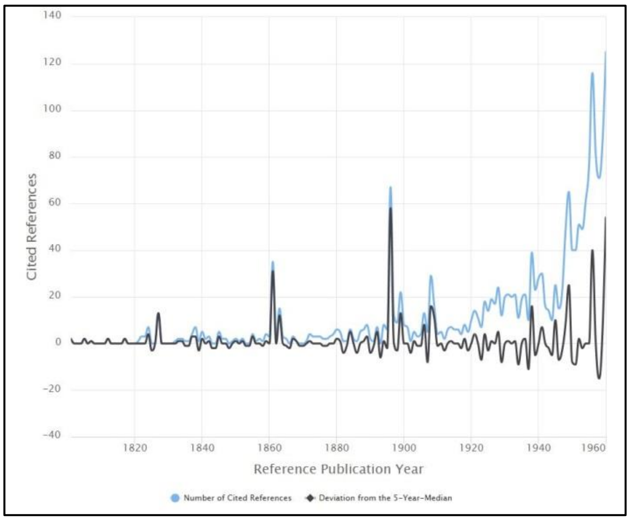
CitedReferencesExplorer (CRExplorer) Manual

(for Version 1.7.7 from June, 30 2017)



Software development: Andreas Thor*,

Content development: Lutz Bornmann** and Werner Marx***

with further support of

Robin Haunschild***, Loet Leydesdorff****, and Rüdiger Mutz*****

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

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

***Max Planck Institute for Solid State
Research
Information Service
Heisenbergstrasse 1,
70506 Stuttgart, Germany.
Email: w.marx@fkf.mpg.de
Email: R.Haunschild@fkf.mpg.de

****Amsterdam School of
Communication Research (ASCoR)
University of Amsterdam
P.O. Box 15793
1001 NG Amsterdam, The Netherlands
Email: loet@leydesdorff.net

*****ETH Zürich
Mühlegasse 21
8001 Zurich, Switzerland
Email: mutz@gess.ethz.ch

Table of contents

1 Introduction

The program CitedReferencesExplorer (CRExplorer) can be used to analyse the cited references (CRs) data in a publication set retrieved from Web of Science (WoS, Clarivate Analytics) or Scopus (Elsevier) as well as to produce the results of the analyses in a graphical or table format for inclusion in a paper or presentation (Thor, Marx, Leydesdorff, & Bornmann, 2016a, 2016b). The program can be retrieved from www.crexplorer.net. It is written in the Java programming language and, thus, runs on most hardware and operating system platforms. The program can be used free of charge. There are two options to run the program: (1) CRExplorer can be launched directly from www.crexplorer.net using Java Web Start Launcher; (2) a executable JAR file can be downloaded from www.crexplorer.net. This manual (and possible newer versions) can also be downloaded from www.crexplorer.net.

The program was primarily developed to identify those publications in fields, of topics, or by researchers which have been most frequently referenced. It is especially suitable to study the historical roots of fields, topics, or researchers by Reference Publication Year Spectroscopy (RPYS; (e.g. Barth, Marx, Bornmann, & Mutz, 2014; Marx, Bornmann, Barth, & Leydesdorff, 2014). RPYS was introduced by Marx et al. (2014) and “is based on the analysis of the frequency with which references are cited in the publications of a specific research field in terms of the publication years of these CRs. The origins show up in the form of more or less pronounced peaks mostly caused by individual publications that are cited particularly frequently” (p. 751).

CRExplorer reads, analyses, and edits the CRs of publications which are previously retrieved from WoS or Scopus. In order to analyse the CRs, the user can consult (1) a graph for identifying most frequently cited reference publication years (RPYs) and (2) a table of CRs which account for specific RPYs. Field-normalization in impact measurement is ensured by the first step of the analysis of CRs: the selection of the publication set on which citation impact is measured.

CRExplorer includes a disambiguation feature which clusters and merges variants of the same CR. This means that the program can also be used as a tool for preparing CR data for other programs, e.g. VOSviewer (van Eck & Waltman, 2010), metaknowledge (McLevey & McIlroy-Young, 2017), or RPYS i/o (Comins & Leydesdorff, 2016). For this purpose, the data are exported in WoS or Scopus format and imported in other programs for further processing. Furthermore, the data can be transferred from one format into another: Imports from Scopus can be exported as WoS files and imports from WoS can be exported as Scopus files (Thor et al., 2016b).

2 Example – analysing the discovery of the "greenhouse effect"

For demonstrating the potential of CRExplorer, Figure 1 shows the citation classics concerning the discovery of the "greenhouse effect", a basic component of climate change. As dataset, we downloaded from the WoS 3,244 publications containing the term "greenhouse effect" in the title or in the abstract or as a keyword (date of searching: 27.10.2016). These papers contain 81,126 CRs to publications which have appeared over 379 years. Figure 1 – as produced by CRExplorer – shows three distinct peaks during the 19th century and a few further peaks during the first half of the 20th century.

The first three pronounced peaks go back to the following publications: Fourier's (1827) paper, entitled "Mémoire sur les températures du globe terrestre et des espaces planétaires", can be seen as the first decisive publication. Fourier found that the earth is warmer than expected. He attributed this to the phenomenon that the earth's atmosphere is transparent for solar radiation but not for the infrared radiation from the ground. Thus, he discovered the (natural) greenhouse effect. Tyndall's (1861) study, entitled "On the absorption and radiation of heat by gases and vapours, and on the physical connexion of radiation, absorption, and conduction", proved that the earth's atmosphere has a greenhouse effect. He concluded that water vapour is the principal gas controlling air temperature. Arrhenius (1896), entitled "On the influence of carbonic acid in the air upon the temperature of the ground", is the first study with a calculation of how changes in the levels of carbon dioxide in the atmosphere can alter the surface temperature through the greenhouse effect.

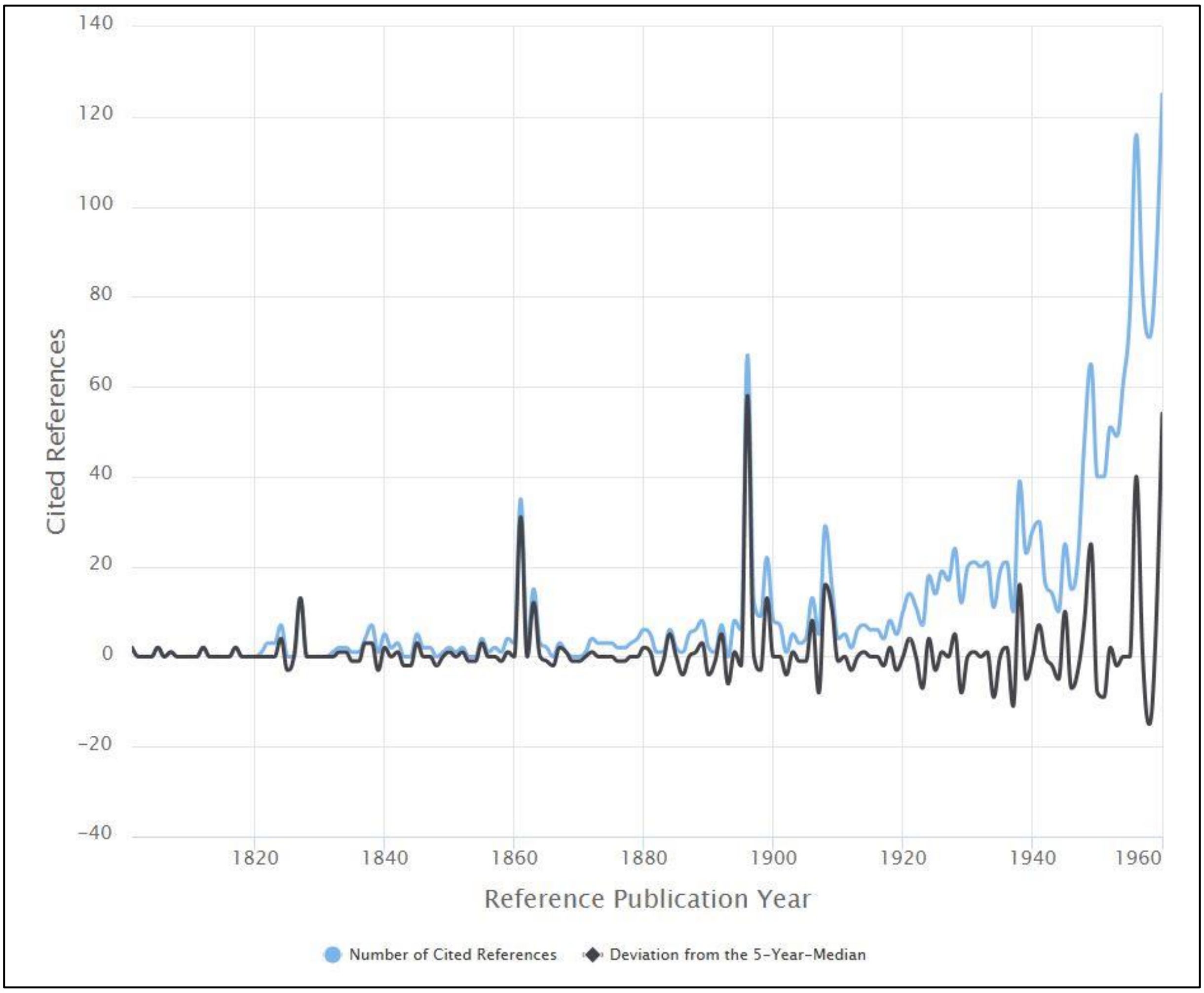


Figure 1. Citation classics concerning the discovery of the "greenhouse effect" and appearing as peaks in the spectrogram provided by CRExplorer

The subsequently following peaks can be assigned to the works of Chamberlin (1899), Arrhenius (1908), Callendar (1938, 1949), and Plass (1956): Chamberlin proposed the possibility that changes in climate could result from changes in the concentration of atmospheric carbon dioxide – thereby supporting the theory of Arrhenius. The 1908 book by Arrhenius was directed at a general audience. Callendar (1938) developed the first complete theory of climatic change and stated in 1938 that carbon dioxide caused the warming trend of the preceding decades (see also Callendar, 1949). He presented evidence that both temperature and the CO_2 level in the atmosphere had been rising over the past half-century. Plass (1956) calculated the transmission of radiation through the earth's atmosphere and predicted that doubling the CO_2 level would bring a 3-4 °C rise. He was the first to use a computer for climate modeling.

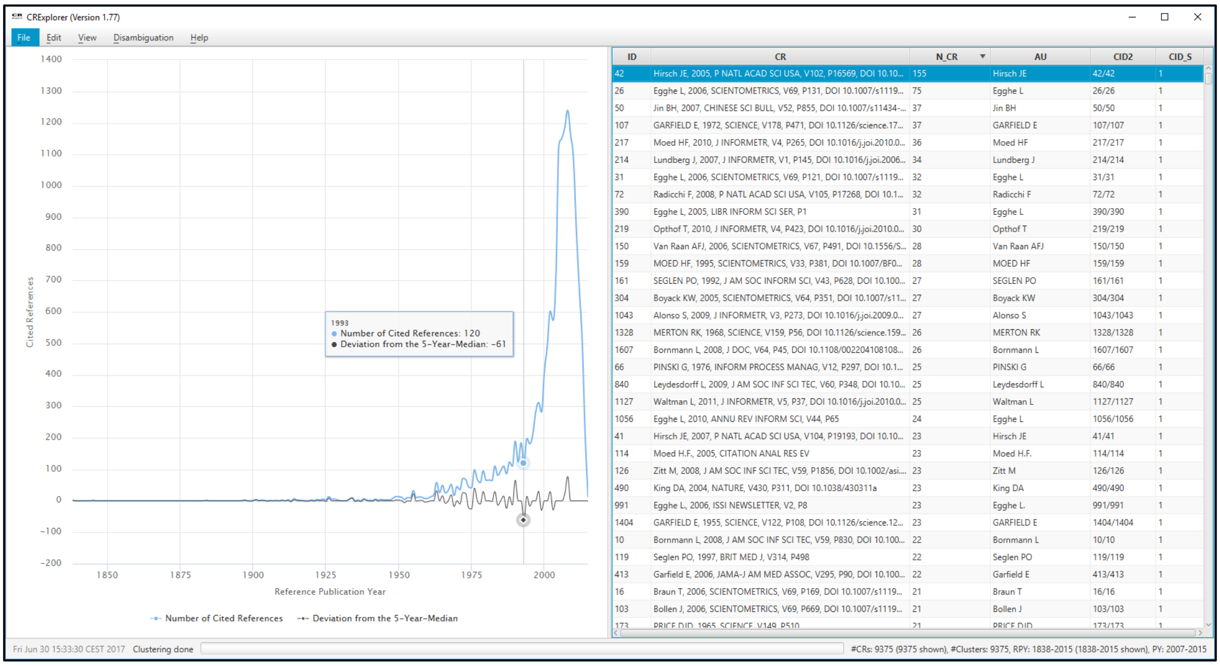
The small study about the discovery of the "greenhouse effect" demonstrates that the analysis of the CRs using CRExplorer is able to identify the citation classics in the climate change literature. These classics deal with the possibility that climatic change results from changes in the concentration of atmospheric carbon dioxide. The RPYS reveals that the discovery of the earth's greenhouse effect and the role of carbon dioxide and water vapor as greenhouse gases are no recent findings, but dates back to the beginning of the nineteenth century.

3 User interface

The screen of CRExplorer has two parts (see Figure 2): **data visualization** on the left side of the screen and **table view** on the right side. The **status bar** informs continously about the number of (shown) CRs, the number of clusters, and the range of RPYs (cited publications) and PYs (citing publications).

3.1 Data visualization

The figure shows the number of CRs per RPY as **JFreeChart** or **HighCharts**. One of these chart types can be choosen by the user (see section 4.1). Furthermore, the figure visualizes the deviation of the number of CRs in each year (*Y*) from the median for the number of CRs in the *X* previous, the current, and the *X* following years (*X* can be set under Settings, see section 4.1.6). The default is the deviation from the 5-Year-Median (*Y* – 2; *Y* – 1; *Y* ; *Y* + 1; *Y* + 2). This deviation from the *X*-year median provides a curve smoother than the one in terms of absolute numbers. If there is no CR in a specific RPY, the number of CRs of this RPY is set to zero for the calculation of the deviation from the median. Using the *X*-year median deviation curve, peaks in the data can be identified more easily than with the number of CRs curve, since each year is compared with its adjacent years.


CRExplorer (Version 1.77)
File
Edit
View
Disambiguation
Help
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
-100
-200
Cited References
1850
1875
1900
1925
1950
1975
2000
Reference Publication Year
1993
● Number of Cited References: 120
● Deviation from the 5-Year-Median: -61
Number of Cited References
Deviation from the 5-Year-Median

ID	CR	N_CR	AU	CID2	CID_S
42	Hirsch JE, 2005, P NATL ACAD SCI USA, V102, P16569, DOI 10.10...	155	Hirsch JE	42/42	1
26	Egghe L, 2006, SCIENTOMETRICS, V69, P131, DOI 10.1007/s1119...	75	Egghe L	26/26	1
50	Jin BH, 2007, CHINESE SCI BULL, V52, P855, DOI 10.1007/s11434-...	37	Jin BH	50/50	1
107	GARFIELD E, 1972, SCIENCE, V178, P471, DOI 10.1126/science.17...	37	GARFIELD E	107/107	1
217	Moed HF, 2010, J INFORMETR, V4, P265, DOI 10.1016/j.joi.2010.0...	36	Moed HF	217/217	1
214	Lundberg J, 2007, J INFORMETR, V1, P145, DOI 10.1016/j.joi.2006...	34	Lundberg J	214/214	1
31	Egghe L, 2006, SCIENTOMETRICS, V69, P121, DOI 10.1007/s1119...	32	Egghe L	31/31	1
72	Radicchi F, 2008, P NATL ACAD SCI USA, V105, P17268, DOI 10.1...	32	Radicchi F	72/72	1
390	Egghe L, 2005, LIBR INFORM SCI SER, P1	31	Egghe L	390/390	1
219	Opthof T, 2010, J INFORMETR, V4, P423, DOI 10.1016/j.joi.2010.0...	30	Opthof T	219/219	1
150	Van Raan AFJ, 2006, SCIENTOMETRICS, V67, P491, DOI 10.1556/S...	28	Van Raan AFJ	150/150	1
159	MOED HF, 1995, SCIENTOMETRICS, V33, P381, DOI 10.1007/BF0...	28	MOED HF	159/159	1
161	SEGLEN PO, 1992, J AM SOC INFORM SCI, V43, P628, DOI 10.100...	27	SEGLEN PO	161/161	1
304	Boyack KW, 2005, SCIENTOMETRICS, V64, P351, DOI 10.1007/s11...	27	Boyack KW	304/304	1
1043	Alonso S, 2009, J INFORMETR, V3, P273, DOI 10.1016/j.joi.2009.0...	27	Alonso S	1043/1043	1
1328	MERTON RK, 1968, SCIENCE, V159, P56, DOI 10.1126/science.159...	26	MERTON RK	1328/1328	1
1607	Bornmann L, 2008, J DOC, V64, P45, DOI 10.1108/002204108108...	26	Bornmann L	1607/1607	1
66	PINSKI G, 1976, INFORM PROCESS MANAG, V12, P297, DOI 10.1...	25	PINSKI G	66/66	1
840	Leydesdorff L, 2009, J AM SOC INF SCI TEC, V60, P348, DOI 10.10...	25	Leydesdorff L	840/840	1
1127	Waltman L, 2011, J INFORMETR, V5, P37, DOI 10.1016/j.joi.2010.0...	25	Waltman L	1127/1127	1
1056	Egghe L, 2010, ANNU REV INFORM SCI, V44, P65	24	Egghe L	1056/1056	1
41	Hirsch JE, 2007, P NATL ACAD SCI USA, V104, P19193, DOI 10.10...	23	Hirsch JE	41/41	1
114	Moed H.F., 2005, CITATION ANAL RES EV	23	Moed H.F.	114/114	1
126	Zitt M, 2008, J AM SOC INF SCI TEC, V59, P1856, DOI 10.1002/asi....	23	Zitt M	126/126	1
490	King DA, 2004, NATURE, V430, P311, DOI 10.1038/430311a	23	King DA	490/490	1
991	Egghe L., 2006, ISSI NEWSLETTER, V2, P8	23	Egghe L.	991/991	1
1404	GARFIELD E, 1955, SCIENCE, V122, P108, DOI 10.1126/science.12...	23	GARFIELD E	1404/1404	1
10	Bornmann L, 2008, J AM SOC INF SCI TEC, V59, P830, DOI 10.100...	22	Bornmann L	10/10	1
119	Seglen PO, 1997, BRIT MED J, V314, P498	22	Seglen PO	119/119	1
413	Garfield E, 2006, JAMA-J AM MED ASSOC, V295, P90, DOI 10.100...	22	Garfield E	413/413	1
16	Braun T, 2006, SCIENTOMETRICS, V69, P169, DOI 10.1007/s1119...	21	Braun T	16/16	1
103	Bollen J, 2006, SCIENTOMETRICS, V69, P669, DOI 10.1007/s1119...	21	Bollen J	103/103	1
173	PRICE DJD, 1965, SCIENCE, V149, P510	21	PRICE DJD	173/173	1

Fri Jun 30 15:33:30 CEST 2017 Clustering done

#CRs: 9375 (9375 shown), #Clusters: 9375, RPY: 1838-2015 (1838-2015 shown), PY: 2007-2015

Figure 2. Screenshot of CRExplorer

Hovering the mouse cursor over a data point on one of the curves in the graph opens a small pop-up window showing the corresponding RPY, the sum of the CRs in this year (**Number of Cited References**), and the **Deviation from the *X*-Year Median**. Thus, the years of peaks in the curves and their impact can be identified easily.

By clicking on **Number of Cited References** or **Deviation from the *X*-Year Median**, respectively, the corresponding curve in the graph appears or disappears.

Using the mouse (and pressing simultaneously the control button), one can mark an area on the graph and restrict the visualized graph to the marked area. The user can recall the initial graph (and thus dissolve any changes) by right-clicking on the graph or by clicking on the button **Reset zoom**. This can also be accomplished via the **View** menue, see section 4.3.

3.2 Table view

The table lists all CRs included in the analysis: The table initially shows the CRs as found in the WoS or Scopus datasets. The data in the columns of the table can be sorted in ascending or descending order. It is also possible to sort by multiple columns: For example, if one wants to sort by colum *X* and then by column *Y*, click firstly on column *Y* and then on column *X*.

To show all bibliographic details of a specific CR, select it in the table, and press the space bar.

Four areas of columns can be distinguished:

1) Cited References.

ID. Every CR receives a unique identification number (column **ID**).

CR (Cited Reference). Original CR from the imported dataset.

RPY (Reference Publication Year**)**. Publication year of the CR.

N_CR (Number of Cited References). The number of times the CR has been cited.

AU (Author). First author.

AU_A: All author names if available in the input file, e.g. from Scopus.

AU_L (Last Name). Last name of the first author.

AU_F (First Name Initial). Initial of the first name of the first author.

J (Source). Showing journal title, volume, issue, and first page in case of journal papers. Other information is shown in case of other document types.

J_N (Source Title). Showing mostly the journal title or the abbreviated book title. Other information is shown in case of other document types.

J_S (Title Short). The column contains the first letters of the words in **Source title** if there is more than one word. If there is only a single word, this word appears here.

VOL (Volume). Volume of the CR.

PAG (Page). First page of the CR.

DOI. DOI of the CR.

2) Indicators.

PERC_YR (Percent in Year): The proportion of the number of times a CR has been cited among the number of all CRs in the same RPY.

PERC_ALL (Percent over all Years): The proportion of the number of times a CR has been cited among the number of CRs over all RPYs.

N_PYEARS (Number of Citing Years): Several CRs are cited in more than one publication with different publication years. **N_PYEARS** shows the number of different publication years in which the CR has been cited.

PERC_PYEARS (Percentage of Citing Years): The publication set which is analysed with CRExplorer includes as a rule publications from different publication years. **PERC_PYEARS** shows the percentage of years (citing years), in which the CR was cited at least once.

N_TOP50, N_TOP75, N_TOP90 (Top 50%, 75%, 90% Cited Reference): These indicators can be used to identify those CRs which have been cited more frequently in the citing years than other CRs in the dataset. In order to identify these CRs, thresholds are computed which identify the top 50%, top 75%, and top 90% in one citing year. In the first step of the computation, the citations in one citing year are sorted in ascending order. In the second step, the thresholds for the top 50%, 75%, and 90% are determined in a given year. In the third step, those CRs are identified which are above the three thresholds. In the fourth step, the numbers of citing years are counted in which the CRs are above the thresholds. These numbers yield **N_TOP50**, **N_TOP75**, and **N_TOP90**.

3) Clustering.

ClusterID (CID2). Each cluster of the standardization procedure (see section **Fehler! Verweisquelle konnte nicht gefunden werden.**) is uniquely identified by its ClusterID, i.e., all CRs of a cluster are marked with a corresponding ClusterID. Thus, the results of the similarity computation can be inspected using the column ClusterID.

ClusterSize (CID_S). The number of CRs in each cluster.

4) Searching.

Score from Search Process (SEARCH_SCORE). The column contains the value 1 for CRs including the string used by the user for searching and the value 0 otherwise (see section 4.3).

3.3 Working with data visualization and table view together

Clicking on a data point in the graph, the CRs data in the table (on the right side of the screen) is sorted by **Reference Publication Year** and **Number of Cited References / Percent in Year**, respectively (in descending order). Furthermore, the first CR with the highest percentage in the particular year is marked. Since the data is sorted by the **Number of Cited References / Percent in Year**, respectively, one can inspect the most important CRs which are responsible for a peak.

4 Options panel

4.1 File

4.1.1 Open

The program uses an internal file format “*.cre”, which can be used as a “working format” and for the exchange of working files. The file contains all data including changes by the user. Using Microsoft Windows, one can double click on any *.cre file and thus run CRExplorer automatically.

4.1.2 Import

Web of Science. CRExplorer opens one or several datasets from WoS. The datasets are downloaded using the option **Save to Other File Formats**. As **Record Content** select **Full Record and Cited References** and as **File Format** select **Other Reference Software**. The records have to be searched in the WoS Core Collection in order to be able to save full records including the CRs.

Scopus. CRExplorer opens one or several datasets from Scopus. The file format **CSV** (including citations, abstracts and references) are chosen for downloading records.

4.1.3 Save

The program saves the dataset in the internal file format “*.cre”.

4.1.4 Save as

The program saves the dataset in the internal file format “*.cre” and asks for a file name.

4.1.5 Export

Web of Science. The dataset is exported in the WoS format as specified in the import section.

Scopus. The dataset is exported in the Scopus format as specified in the import section.

Note that the export can only contain data from the import, when one transfers from Scopus to WoS. WoS files, for example, only consider the first authors while Scopus files include all authors in the CR field. If one transfers from WoS to Scopus, not all information can be provided, since this information is not available in downloads from WoS.

CSV (Graph). The graph data are exported for further processing in programs such as Excel, R, Stata, or GnuPlot. We provide Stata (plotrpys) and R (BibPlots) commands which produce monocrome and colored graphs. Both can be found in SSC Archive (in the case of Stata) and CRAN (Comprehensive R Archive Network, in the case of R), respectively.

CSV (Cited References). A table with CRs data is saved as a csv-file.

CSV (Citing publications). A table with data on the citing publications is saved as a csv-file.

CSV (Cited References + Citing publications). A table with both CRs data and data on the citing publications is saved as a csv-file.

4.1.6 Settings

Table.

In the sections **Cited References**, **Indicators**, **Clustering**, and **Searching** the columns can be selected which should be displayed in the table. It is also possible to select/deselect all columns. These functions enable to restrict the columns to those which are needed for a specific analysis. In the section **Value Settings**, the **Number of Digits** for all numerical columns can be modified.

Furthermore, the **N_PCT Range** can be adjusted. It might be a problem in computing **N_TOP50**, **N_TOP75**, and **N_TOP90** that the citation counts in a citing year are inflated by zeros (and/or similar values). Thus, we included the option in the program to extend the number of citing years which are considered in calculating **N_TOP50**, **N_TOP75**, and **N_TOP90**. If only the citing year itself should be considered in the analysis, the **NPCT Range** is set to 0. If it is set to 1, the thresholds for the top 50%, 75%, and 90% are computed on the base of the citations from the preceding (*t*-1) and succeeding (*t*+1) citing years. This doubles the underlying dataset in the first and last citing year (since year *t*-1 or *t*+1, respectively, do not exist) and triples it in all other years.

Chart.

Chart Layout.

The user can select the lines which should be displayed: **Number of Cited References**, **Deviation from the Median** or both. The deviation of the number of CRs in each year *Y* from the median for the number of CRs in the previous, the current, and the following years can be set. The default is 2: $Y - 2$; $Y - 1$; Y ; $Y + 1$; $Y + 2$. Thus, the user can change the number to any other number and can thus work with medians calculated based on different time windows. Furthermore, **Stroke Size** and **Shape size** for the lines in the chart as well as **Label Font Size** and **Tick Font Size** for the axes can be set.

Chart Engines.

Two different chart types (**JFreeChart** and **HighCharts**) can be selected. Both types have the same functionality: the user can zoom into the graph and click on a peak whereby the underlying CRs are sorted and marked correspondingly in the table. However, both types are different in look and feel. **JFreeChart** is a static visualization, similar to graph types in Excel or GnuPlot; **HighCharts** is a dynamic, web-based (“modern”) visualization. The underlying data of any graph can be downloaded with **File** – **Export** – **Graph CSV** and can be visualized in another software (e.g. Excel, R, or GnuPlot). Both chart engines are available using the web-start, but only JFreeChart is available using the downloaded JAR version of CRExplorer.

The **JFreeChart** can be saved in various formats by right-clicking on the graph. This is not possible with **HighCharts**. If the user wants to include the graph from **HighCharts** into another program (e.g. Microsoft Word), he or she should use programs such as the Snipping Tool (which is available in Microsoft Windows) or KSnapshot (which is available in KDE) to cut the graph.

If there are any problems with changing of chart types, try **View** – **Reset Chart** (see section 4.3.9) or restart the CRExplorer.

Import/Export.

Restrict Import of Cited References.

Maximum Number of CRs. Setting the number to zero means that there is no import limit, but the processing is limited by the available memory on the computer. It is the default to import a maximum of 100,000 CRs. This is the number which can be processed by most computers.

Maximum Number of Publikations. The number of citing papers can be restricted. Setting the number to zero means that there is no import limit, but the processing is limited by the available memory on the computer.

Minimum / Maximum Publication Year. The imported range of RPYs can be restricted. The default is zero for minimum and maximum which means that the RPY is not relevant for the import. This function can also be used to define the minimum import range (e.g. Minimum=1900 and Maximum=0 mean that all CRs with RPY>=1900 are imported).

Advanced Export Options.

Include Publications without CRs in export. If this option is set, the export to WoS, Scopus, or CSV include all citing papers that have been imported. Otherwise (and this is the default) the export do not include the citing papers without CRs.

4.1.7 Exit

Leave CRExplorer.

4.2 Edit

The user may often wish to restrict the CR analysis to a certain time period. Very early and most recent years are frequently not very helpful for the identification of the most frequently cited publications in the history. Although CRExplorer allows for the selection of periods in the graph, this selection does not lead to changes in the dataset. However, there are several ways of removing data from a dataset.

4.2.1 Remove selected Cited References

Rows in the table can be marked and deleted using the menu item.

4.2.2 Remove selected Cited References w/o Year

All CRs are removed without a year in the column RPY.

4.2.3 Remove by Reference Publication Year.

The user can remove the data for specific RPYs.

4.2.4 Remove by Number of Cited References

All CRs with a number of reference counts (column **N_CR**) within the specified range are deleted. This kind of restriction is helpful in identifying publications from early RPYs with a substantial impact (and to suppress the noise of less cited publications). Furthermore, in RPYs with many sparsely cited publications the publications with substantial impact can then be easier identified.

4.2.5 Remove by Percent in Year

CRs can be removed by using thresholds for the column **PERC_YR**. Thus, it is possible to remove lowly CRs whereby "lowly" is defined in terms of the citation distribution in the RPYs.

4.2.6 Retain Cited References by ID

One can select the citation environment of a specific CR (or of two or even more CRs) in the form of all co-cited CRs and analyse these CRs (e.g. for applying RPYS-CO, see Marx, Haunschild, Thor, & Bornmann, 2017). The specific CRs can be prominent and seminal works which are used as a kind of marker or tracer references for a specific topic in a field. We assume that papers which cite the selected CRs are potential candidates for citing also many other CRs relevant in a specific historical context. This method takes advantage of the fact that concurrently cited (co-cited) papers are more or less closely related to each other (Small, 1977).

By selecting **Retain Cited References by ID**, the CRs are restricted based on the IDs specified by the user (e.g. 20, 25, 40). Thus, only CRs are retained which are co-cited with the IDs (CRs) specified.

4.2.7 Retain Publications citing Selected Cited References

The CRs are restricted based on the CRs marked by the user in the table. Thus, only CRs are retained which are co-cited with the CRs specified.

4.2.8 Retain Publications within Citing Publication Year

The CRs are restricted based on the publication years of the citing publications. The user can specify the range of (citing) publication years. Thus, only CRs are retained which are cited or co-cited with the CRs in citing publications from the specified years.

4.2.9 Copy Selected Cited References

The CR (CRs) which is (are) selected in the table is (are) copied to the clipboard for use in other programs (e.g. spreadsheet or word processing programs).

4.3 View

4.3.1 Info

The message box provides some basic information on the dataset: **Number of Cited References** and **Number of Cited References (shown)**. The user has the option to select data temporarily. Thus, the complete number of CRs and the temporarily selected number are shown. Furthermore, the user can cluster and merge CR variants, and the number of calculated clusters is shown. The message box also contains the **Range of Cited References Years**, the **Number of different Cited References Years**, the **Number of Publications** (including publications without CRs), the **Number of Citing Publications** (excluding publications without CRs), the **Range of Citing Publication Years**, and the **Number of different Citing Publication Years**.

4.3.2 Cited Reference (Details)

An info box including all information from the columns appears for a selected CR.

4.3.3 Citing Publications

The publications are shown which cite the CR (CRs) selected by the user.

4.3.4 Show Cited References w/o Years

Scopus or WoS data sets contain CRs without RPY. As a rule, these CRs are not considered in the analyses of CRExplorer, but can be shown optionally in the table.

4.3.5 Filter by Reference Publication Year

The graph is temporarily restricted to the minimum and maximum years included.

4.3.6 Show Cited References of selected cluster(s) only

Restrict the CRs to only those which are in the same cluster(s) as the selected CR(s).

4.3.7 Search Cited References

CRs including the search string are sorted to the first positions in the tabular list of CRs. The column **Search_Score** in the table contains the value 1 for CRs including the search string and the value 0 otherwise. The column **Search_Score** itself can be used for sorting the data.

4.3.8 Show all Cited References (currently X of Y)

All filters set by the user for filtering CRs are deactivated. If *X* in the menue item is smaller than *Y*, a filter is active.

4.3.9 Reset Chart

The **JFreeChart** or **HighCharts** axes are reset to their maximum range.

4.4 Disambiguation

The user has the possibility to detect variants of the same CR, cluster them, and merge their occurrences (number of CRs). The automatic clustering of variants is restricted to variants within the same publication year, but not across publication years. Thus, a reference to a first edition of a book and a reference to a later edition are not clustered by this routine.

However, the user can make further adjustments to the author names, the journal or book title and other bibliographic information of CRs (across publication years). The clustering uses the table with the list of CRs as input file. Since the clustering (and merging) is a complex process which needs a lot of computer resources, the user is adviced to cluster only those data which is of interest.

The clustering and merging of the data is especially important for the Scopus data, since the CR data is in Scopus more heterogeneous than in WoS. Scopus data contains more information than WoS data (all authors and the titles of the referenced publication) which increases the probability of variants of the same CR. Furthermore, Scopus data may contain fragmented CR data which cannot be completely parsed into the bibliographic categories of CRExplorer (e.g. authors, titles, or volume numbers). The heterogeneous data of Scopus can be inspected best by sorting the list of CRs under the column **Authors**. A possible way of dealing with the heterogeneous CRs is to try their clustering and merging with complete CRs (if fragmented CRs are variants of complete CRs).

4.4.1 Cluster equivalent Cited References

In the first step of eliminating variants of CRs, the variants of the same cited publication are identified. Two attributes are used for a first similarity computation: **Last name** of the first author and **Source title**. Based on this data, CRExplorer determines the pair-wise similarity of variants of CRs. The program computes the Levenshtein similarity (as provided by the SimMetrics library https://github.com/Simmetrics/simmetrics) of both attributes (see also Wasi & Flaaen, 2015). The Levenshtein similarity of two strings s_1 and s_2 is defined as sim(s_1, s_2) = 1 - LD(s_1, s_2)/max(|s_1|, | s_2|). Here |s| denotes the length of a string s and LD (s_1, s_2) is the Levenshtein distance which is defined as follows: The Levenshtein distance between two strings s_1 and s_2 is the minimal number of single-character edit operations (i.e., insertion, deletion, or substitution) required to transform string s_1 into s_2. The Levenshtein distance is 0 for equal strings (no edit operations necessary) and equals max(|s_1|, |s_2|) for totally different strings (substitute the first min(|s_1|, |s_2|) characters and insert / delete the remaining characters). Therefore, for any two strings the Levenshtein similarity is between 0 to 1 where 0 corresponds to “totally different” and 1 to “identical”.

For two CRs o_1 and o_2, CRExplorer computes the Levenshtein similarity of the first authors’ last names as well as the similarity of the source titles. The two CRs o_1 and o_2 are considered as “matching” if the weighted average (ratio 2:1) of the two similarity values is equal to or greater than the threshold of 0.75 (this can be changed by the user, see below). The combination of multiple similarity values that are based on different attributes typically achieves a better match quality compared to a single similarity of the entire CR strings. First, it restricts the similarity computation to relevant (and available) attribute values. Second, the combination allows for an appropriate weighting of attributes independent from their actual string length.

CRExplorer performs a clustering based on the matching results, i.e., the list of the matching CR pairs. Two CRs o_1 and o_2 are assigned the same cluster, if the pair (o_1, o_2) appears in the matching result or if there is a list of other CRs t_1, ..., t_n so that (o_1,t_1), (t_1,t_2), (t_2,t_3), ..., (t_n-1,t_n), and (t_n, o_2) are all among the matching pairs. Each cluster is uniquely identified by its **ClusterID**, i.e., all CRs of a cluster are marked with a corresponding **ClusterID**. The results of the similarity computation can be inspected using the column **ClusterID** in the table. The number of CRs in each cluster is provided by the column labeled **ClusterSize**.

4.4.2 Merge clustered Cited References

In the second step of eliminating variants of CRs (subsequent to the clustering of variants), the **ClusterID** is used for the aggregation of the variants, i.e., the values of the corresponding lines in the table are summed up per cluster.

The clustering and merging procedure can be helpful in aggregating variants of the same CR. However, this procedure itself is prone to error. For example, if there are several CRs from the same authors and published in the same journal in one year, these CRs are clustered, although they refer to different publications. This aggregation error affects mainly journal papers.

For this reason, we strongly recommend that a user of the procedure controls the results from the clustering procedure and corrects wrong matches. For the manual correction we have implemented some features in the program which support the user in post-processing the clustered results. Corresponding control buttons appear above the table, when the user starts the clustering process. The features can be used separately or in combination to change the cluster results before the merging of the CR variants is started. The effects of the features can be inspected by the values in **ClusterID** which has two components: The first value in the column (before the slash) shows the cluster numbers which result from the initial clustering process which was done automatically. The second value (after the slash) marks sub-clusters which change after using the features. Thus, the user should inspect the second value of the **ClusterID** to assess the results of the chosen post-processing.

The features implemented in the program are the following:

Slide control: For matching similar CRs, Levenshtein is initially used as similarity function with a threshold of 0.75. However, the user can change this afterwards by using the slide control which accepts values between 0.5 (shown as 50) and 1 (shown as 100). If the slider is moved in the direction of 50, less similar CRs are matched; by moving the slider in the direction of 100, the matching process becomes increasingly restricted.

Volume, **Page**, and **DOI**: The user can select **Volume**, **Page**, and **DOI** in order to differentiate the clusters further. These selections affect the whole dataset and not only CRs which are marked by the user. Note that the Levenshtein approach is not applied to **Volume**, **Page**, and **DOI**; a precise match is required for these attributes.

Manual generation of sub-clusters: The tool offers three different ways for the manual changing of sub-clusters. They are named as **Same**, **Different**, **Extract**, and **Undo**. Most of the problems with the automated clustering procedure occur with false positives: The algorithm matches CRs, although they should be kept separate. For the manual separation of clusters, the user can apply **Different** or **Extract**. **Different** assigns different sub-cluster IDs to those CRs which are marked by the user manually (using the mouse click). **Extract** puts the marked CRs in a separated sub-cluster. **Same** gives marked CRs the same sub-cluster-ID. Manual changes based on **Same**, **Different**, and **Extract** can be rolled back using the **Undo** button.

4.5 Help

4.5.1 Online Manual

The manual of CRExplorer is opened.

4.5.2 Info

Some information is given by the program.

5 A practically oriented short guide to use CRExplorer

The following hints and rules of thumb may be helpful for the use of CRExplorer.

5.1 Establishing a publication set

The publication set to be analysed may comprise the publications of specific authors, journals, research fields or any other publication corpora of interest. Based on our experience hitherto, we recommend that the size of the relevant publication set should not be much less than 100 papers for a meaningful CR analysis. On the other hand, a typical research field normally comprises much larger publication sets. Here, the size of the publication set used for the CR analysis is limited by more practical considerations; memory requirements increase with the number of CRs.

The sample under study does not need to comprise every relevant publication (e.g. related to a specific research field) and should not contain too many irrelevant papers at the same time. As a rule, some missing publications do not change the overall picture derived from CRExplorer analysis and the location of the peaks in the spectrogram (the graph produced by CRExplorer) will hardly be affected. On the other hand, the presence of too many irrelevant papers in the set increases the noise in the spectrogram and reduces the height and distinctness of the peaks.

After uploading the WoS or Scopus records into CRExplorer, one can analyse the complete range of RPYs, for example, in order to reconstruct the evolution of a research field as reflected by the references cited (by the members of the corresponding scientific community). Alternatively, one can focus on early references in order to investigate the origins and historical roots of a research field or one may wish to analyse recent RPYs (e.g. the last decade only) to reveal recently published highly-cited papers. The historical roots have been investigated in most of the studies published so far (see the publication list at www.crexplorer.net).

5.2 Clustering of CRs

For a first overview, one is advied to sort the table of references by RPY in temporal order and concurrently by the reference counts. For example, this procedure helps to identify the earliest CRs and the number of references in more recent years. We suggest marking the following columns via the table settings: **Cited Reference**, **Reference Publication Year**, **Number of Cited References**, **Percent in Year**, **Percent over all Years**, **ClusterID**, and **ClusterSize**. The other columns can be ignored for the moment.

Note that the reference counts of all CRs within a specific RPY are mutually comparable since they can be considered as field-normalized given the common search string: all citing papers belong to the same set, i.e., papers from a single research field, topic or author. Thus, the CRs generally originate from the same citation culture.

In the next step, the equivalent references are clustered. This clean-up procedure (the so-called “disambiguation”) is needed because there are many incomplete and misspelled references among the CRs (in particular with regard to the source name, volume, and page numbers). The automatic clustering procedure of CRExplorer does not work absolutely correctly. For example, the program cannot differentiate between papers published by the same author in the same journal and year. In other words, different publications are identified by the program as variants of the same publication and are clustered. Using volume and page numbers for clustering reference variants routinely leads to satisfactory results in most of the cases (see the options above the table of CRs in CRExplorer after activating the clustering function).

However, using volume and page numbers may be problematic for papers where volume numbers are missing or where page numbers within the range of pages are cited (rather than the starting pages). The use of the DOIs (in addition to volume and page numbers) to cluster reference variants normally results in detecting fewer variants and incomplete clustering, because DOIs are not available in many cases or are not properly assigned to CRs. Therefore, CRExplorer offers the possibility of cleaning-up the data manually. However, the manual cleaning-up of a dataset is only feasable in the case of relatively low numbers of CRs in the sample. This is normally the case for references which are published earlier than 1900 (and sometimes also for references published before 1950).

As an example for reference variants in the data, we show in Table 1 a list of CRs from our analysis of the discovery of the “greenhouse effect” presented in section 2. There are some reference variants of the Arrhenius (1896) paper which can be clustered and merged.

One reference variant (ID 333) cites the journal (*Philosophical Magazine*) with incorrect volume number, two others (ID 4998 and 5002) cite the corresponding Swedish papers, further two variants (ID 2553 and 4999) do not cite the starting page number, three variants (ID 5587, 50801 and 70088) cite journal title variants, and two CRs (ID 61256 and 80494) cite the title of the paper rather than the journal title.

Table 1. Reference variants of the Arrhenius (1896) paper with the number of occurrences.

ID	Reference variant	N
12156	Arrhenius S., 1896, PHILOS MAG 5, V41, P237	50
333	Arrhenius S, 1896, PHILOS MAG 5, V5, P237	22
4998	Arrhenius S., 1896, NORDISK TIDSKRIFT, V14, P121	4
5002	ARRHENIUS S, 1896, BIHANG TILL KUNGL SV, V22, P1	3
2553	ARRHENIUS S, 1896, PHILOS MAG, V41, P267	2
4999	ARRHENIUS S, 1896, PHILOS MAG, V41, P274	2
5587	Arrhenius S., 1896, LONDON EDINBURGH DUB, V41, P237	2
50801	Arrhenius Svante, 1896, LONDON EDINBURGH DUB, V5, P237	2
61256	ARRHENIUS S, 1896, ON INFLUENCE CARBONI	1
70088	ARRHENIUS S, 1996, PHIL MAG J SCI APR, P237	1
80494	ARRHENIUS S, 1896, INFLUENCE CARBONIC A	1

5.3 Manual Cleaning

If manual cleaning-up is applied, we suggest ordering the table items by the number of references per cluster (**ClusterSize**). In a first step, the items of larger clusters (which usually comprise the majority of CRs) should be checked and cleaned-up. If the dataset contains a manageable number of clusters and the user needs a (more or less) completely cleaned-up dataset, clusters with a smaller cluster size (or even with cluster size one) should also be investigated. The items with cluster size one can best be checked after ordering the references alphabetically (second ordering criterion in the program after cluster size). If the referenced authors in the dataset are cited more or less correctly, the variants of the CR to be checked appear one after another. In order to cope with a large number of CRs when using CRExplorer, a substantial cut may be necessary to master the flood of references extracted from the publication set. In the case of very large reference sets it is helpful to exclude all references with reference count one. These references are usually the majority of the CR items within a given publication set, but are only a small fraction of the total of CR counts. These references should be excluded before one checks for reference variants and inspects the spectrogram. However, the best strategy for manual cleaning depends on the publication set and the intended analysis.

5.4 Inspection of the spectrogram

In many publication sets, the references to be analysed have been published over a long period with quite different publication and citation cultures: the average number of references per RPY increases substantially in the course of time. We may distinguish between the period of “little science” (prior to 1950) and that of “big science” (since 1950) (Marx, 2011). In particular, the reference counts before the RPY 1900 are comparatively low. Whereas the average (and maximum) reference count (**Number of Cited References**) increases with the passage of time, the share of reference counts accounting for a specific reference in a single year (**Percent in Year**) tends to decrease. This is the result of the continuously increasing number of papers and CRs, respectively, leading to increasingly less pronounced peaks in the spectrogram.

The spectrogram may not exhibit distinct peaks unless the range of RPYs is limited by excluding the more recent period (**Edit** – **Remove by Reference Publication Year**). If the analysis aims to detect influential early works, it is reasonable to remove all references with RPYs later than either 1950 or 1900. With regard to the inspection and interpretation of the spectrogram, it might also be helpful to select two (or more) consecutive RPY periods (e.g.

1800-1900 and 1901-1950) rather than one single period. Thus, one would analyse the references and reveal the reference peaks using two or more separate spectrograms. This simplifies the analysis and interpretation.

After the clustering process, both the spectrogram and the table with the CRs can be further adjusted and revised by selecting a minimum reference count (**Edit** – **Remove by Number of Cited References**). Removing the many references with reference count 1 (or in the case of large data sets: 2-3) makes the spectrogram more pronounced and the table of references better manageable. During the inspection of the spectrogram the question typically arises, which specific peaks should be considered as distinct reference peaks for further analysis and discussion. This decision is rather arbitrary and depends on the specific data set and the maximum number of top references to be discussed. A minimum reference count of 10, for example, has proved to be reasonable for investigating referenced papers published prior to 1900 (e.g. if the analysis aims to detect influential early works).

For the identification of the peaks and the corresponding top references, both the overall number of CRs (red curve in the spectrogram of the **JFreeChart**) and the (absolute) deviation from the median (blue curve of the **JFreeChart**) can be considered. Normally, both curves deliver the same amount of information and can be used alternatively or concurrently. There may be cases for which one or the other curve might be better suited.

If one would like to analyse the recent evolution of a research field and focus on the more recent decades of the RPY, the spectrogram is less informative. The peaks are less pronounced, because each reference, although highly cited, comprises an increasingly smaller share of the reference counts of a RPY. This kind of analysis can best be performed via the table of references ordered concurrently by the RPY (**Reference Publication Year**) and the reference counts (**Number of Cited References**) (with the most CRs at the top).

5.5 The RPYS-CO approach

CRExplorer allows the restriction of the CRs to only those which are co-cited with at least one selected CR using the menu items **Edit** – **Retain Publications citing Selected Cited References**. This restriction takes advantage of the fact that concurrently cited (co-cited) papers are more or less closely related to each other. One can select the citation environment of a specific reference (or of two or even more references) in the form of all co-cited references and analyse these references applying RPYS-CO. The specific reference should be a prominent and seminal work which is used as a kind of marker or tracer reference for a specific topic in a field. We assume that papers which cite the selected reference(s) are potential candidates for citing also many other references relevant in a specific historical context.

For example, if we are interested to refine the analysis of the discovery of the "greenhouse effect" with regard to the earliest roots, we could use the seminal paper by Arrhenius (1896), the most pronounced pre-1900 peak in Figure 1. Svante Arrhenius was the first scientist who calculated how changes in the levels of carbon dioxide in the atmosphere could alter the surface temperature through the greenhouse effect. He predicted that emissions of carbon dioxide from the burning of fossil fuels were large enough to cause global warming. His paper can be seen as a cornerstone in the evolution of climate change research.

By analysing the co-citations of Arrhenius (1896) as a marker reference, we investigate the discovery of the greenhouse effect and the specific role of carbon dioxide. This research topic marks the historical roots and origins of the current climate change research. As a result of applying RPYS-CO to the history of climate change research in a recent study (Marx et al., 2017) the decisive works of the French mathematician and physicist Joseph Fourier become more clearly visible. We get to know that his 1827 paper is a reproduction of his 1824 work, which is much more pronounced in the RPYS-CO spectrogram.

5.6 Conclusions

In summary, the strategy for using CRExplorer strongly depends both on the size of the publication and reference set to be analysed and on the specific focus of the analysis (early or more recent works). The strategy has to be adapted to the specific goal of the analysis. CRExplorer can be applied for three main objectives: (1) the detection of the knowledge basis (i.e. the origins and historical roots) of research topics, (2) the investigation of influential works published more recently, and (3) the disambiguation of CRs data.

References

Arrhenius, S. (1896). On the influence of carbonic acid in the air upon the temperature of the ground. *Philosophical Magazine and Journal of Science Series, 5*(41), 237-276.

Arrhenius, S. (1908). *Worlds in the making: the evolution of the universe*. New York and London: Harper and Brothers Publishers.

Barth, A., Marx, W., Bornmann, L., & Mutz, R. (2014). On the origins and the historical roots of the Higgs boson research from a bibliometric perspective. *The European Physical Journal Plus, 129*(6), 1-13. doi: 10.1140/epjp/i2014-14111-6.

Callendar, G. S. (1938). The artificial production of carbon dioxide and its influence on temperature. *Quarterly Journal of the Royal Meteorological Society, 64*, 223-237.

Callendar, G. S. (1949). Can carbon dioxide influence climate? *Weather, 4*, 310-314.

Chamberlin, T. C. (1899). An attempt to frame a working hypothesis on the cause of glacial periods on an atmospheric basis. *Journal of Geology, 7*, 545-584, 667-685, 751-787.

Comins, J. A., & Leydesdorff, L. (2016). RPYS i/o: software demonstration of a web-based tool for the historiography and visualization of citation classics, sleeping beauties and research fronts. *Scientometrics, 107*(3), 1509-1517. doi: 10.1007/s11192-016-1928-z.

Fourier, J. B. J. (1827). Mémoire sur les températures du globe terrestre et des espaces planétaires. *Mémoires de l'Académie Royale des Sciences, 7*, 569–604.

Marx, W. (2011). Special features of historical papers from the viewpoint of bibliometrics. *Journal of the American Society for Information Science and Technology, 62*(3), 433-439. doi: 10.1002/asi.21479.

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology, 65*(4), 751-764. doi: 10.1002/asi.23089.

Marx, W., Haunschild, R., Thor, A., & Bornmann, L. (2017). Which early works are cited most frequently in climate change research literature? A bibliometric approach based on Reference Publication Year Spectroscopy. *Scientometrics, 110*(1), 335-353. doi: 10.1007/s11192-016-2177-x.

McLevey, J., & McIlroy-Young, R. (2017). Introducing metaknowledge: Software for computational research in information science, network analysis, and science of science. *Journal of Informetrics, 11*(1), 176-197. doi: http://dx.doi.org/10.1016/j.joi.2016.12.005.

Plass, G. N. (1956). The Carbon Dioxide Theory of Climatic Change. *Tellus, 8*(2), 140-154.

Small, H. G. (1977). A co-citation model of a scientific specialty: a longitudinal study of collagen research. *Social Studies of Science, 7*(2), 139-166. doi: 10.1177/030631277700700202.

Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (2016a). Introducing CitedReferencesExplorer (CRExplorer): A program for Reference Publication Year Spectroscopy with Cited References Standardization. *Journal of Informetrics, 10*(2), 503-515.

Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (2016b). New features of CitedReferencesExplorer (CRExplorer). *Scientometrics, 109*(3), 2049-2051.

Tyndall, J. (1861). On the absorption and radiation of heat by gasses and vapours, and on the physical connection of radiation, absorption, and conduction. *Philosophical Magazine, 4*(22), 273-285.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538. doi: 10.1007/s11192-009-0146-3.

Wasi, N., & Flaaen, A. (2015). Record linkage using Stata: Preprocessing, linking, and reviewing utilities. *Stata Journal, 15*(3), 672-697.