Analysing cited references via the CRExplorer: A practically oriented short guide

The CRExplorer (www.crexplorer.net) is a bibliometric tool developed by Andreas Thor (Thor, Marx, Leydesdorff, & Bornmann, 2016), enabling the analysis of references cited within a given publication set retrieved from the Web of Science (WoS) database. The software was developed to support the bibliometric method known as Reference Publication Year Spectroscopy (RPYS), which was initially developed by Werner Marx (Bornmann & Marx, 2013; Marx, Bornmann, Barth, & Leydesdorff, 2014). The following hints and rules of thumb may be helpful for the use of the CRExplorer.

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 RPYS. On the other hand, a typical research field normally consists of much larger publication sets. Here, the size of the publication set used for the RPYS is limited by more practical considerations: WoS documents can only be downloaded in units of 500 records each.

The publication set under inspection 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. Some missing publications do not change the overall picture derived from the CRExplorer analysis and the peaks of the spectrogram (the graph produced by the 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 of the peaks.

After uploading the WoS records in the CRExplorer, one can analyse the complete range of reference publication years, 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 can analyse recent reference publication years (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).

Clustering of cited references

For a first overview, one should order the table of references (alongside the RPYS spectrogram) by reference publication years (CRExplorer: Cited Reference Year) in temporal order and concurrently by the reference count (CRExplorer: Number of Cited References). For example, this procedure helps to identify the earliest cited references and the number of references in more recent years. We suggest marking the following columns via the table settings: Cited Reference, Cited Reference Year, Number of Cited References, Percent in Year, Percent over all Years, Cluster ID, and Cluster Size. The other columns can be ignored for the moment.

Note that the reference counts of all references within a specific reference publication year are mutually comparable since they are field-normalized: All citing papers belong to the same research field. Thus, the cited references generally originate in 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 cited references (in particular with regard to the source name, volume, and page numbers). The automatic clustering procedure of the 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 normally leads to satisfactory results (see the options above the table of references in the CRExplorer).

However, these options 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 DOI (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 cited references. Therefore, the CRExplorer offers the possibility of cleaning-up the data manually. However, the manual cleaning-up of a dataset is only practicable with low numbers of cited references. This is normally the case for references which are published earlier than 1900 (and sometimes also for references published before 1950).

Manual Cleaning

If manual cleaning-up is applied, we suggest ordering the table items by the number of references per cluster (CRExplorer: Cluster Size). In a first step, the items of larger clusters (which usually comprise the majority of cited references) 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 cited reference to be checked appear one after another. In order to cope with a large number of cited references when using the 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 cited reference items within a given publication set, but are only a small fraction of the total of cited reference counts. These references should be excluded before one checks for reference variants and inspects the spectrogram.

Inspection of the spectrogram

Normally, 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 reference publication year 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 reference publication year 1900 are comparatively low. Whereas the average (and maximum) reference count (CRExplorer: Number of Cited References) increases with the passage of time, the share of reference counts accounting for a specific reference in a single year (CRExplorer: Percent in Year) tends to decrease. This is the result of the continuously increasing number of papers and cited references, respectively, leading to increasingly less pronounced peaks in the spectrogram.

The spectrogram may not exhibit distinct peaks unless the range of reference publication years is limited by excluding the more recent period (CRExplorer: Data / Remove by Cited Reference Year). If the analysis aims to detect influential early works, it is reasonable to remove all references with reference publication years 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) consecutively referenced publication year 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 cited references can be further adjusted and revised by selecting a minimum reference count (CRExplorer: Data / 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 cited references (red curve in the spectrogram) and the (absolute) deviation from the median (blue curve) 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 reference publication year, 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 reference publication year. This kind of analysis can best be performed via the table of references ordered concurrently by the reference counts (in temporal order) and the reference publication year (with the most-cited references at the top) (CRExplorer: Cited Reference Year and Number of Cited References).

Final word

In general, the strategy for using the 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.

References

Bornmann, L., & Marx, W. (2013). The proposal of a broadening of perspective in evaluative bibliometrics by complementing the times cited with a cited reference analysis. *Journal of Informetrics, 7*(1), 84-88. doi: 10.1016/j.joi.2012.09.003.

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.

Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (2016). Introducing CitedReferencesExplorer (CRExplorer): A program for Reference Publication Year Spectroscopy with Cited References Disambiguation. Retrieved January 18, 2016, from http://arxiv.org/abs/1601.01199