

Galton 2012: the bibliometric journey continues

Juan Gorraiz¹, Christian Gumpenberger¹ and Martin Wieland¹

¹University of Vienna, Library and Archive Services, Bibliometrics Department, Boltzmanng. 5, A-1090 Vienna

Corresponding Author:

Juan Gorraiz, University of Vienna, Vienna University Library, Bibliometrics Department, Boltzmanngasse 5, A-1090 Wien, Austria – e-mail:

Abstract. Francis Galton was not only a universal genius; he can also be regarded as a scientometric pioneer. On the occasion of his 100th death anniversary in 2011, he was used as a role model for a bibliometric impact analysis of his works. This is the follow-up study of a previous citation analysis, which was now expanded by introducing ProQuest Dissertations & Theses (PQDT) as a complementary data source. Additionally all retrieved documents in WoS either citing or mentioning Galton were visualized using Bibexcel, Pajek and VOS Viewer. Furthermore Galton's h-index and g-index were calculated and found to be very high compared to other historic scientific personalities. The citation-to-oblivion ratio was different for PQDT in comparison to WoS or Scopus. Visualization allows better interpretation and understanding of the obtained results and is useful for the identification of eponyms.

Overall citation analysis and occurrence counting are complementary useful methods for the impact analysis of the works of "giants". This type of retrospective bibliometric studies presents an interesting and promising field of activity for librarians and information specialists.

Keywords: Francis Galton, citation analysis, oblivion, history of science, visualization, h-index, g-index, theses, dissertations, Web of Science, Scopus, Google Scholar, ProQuest Dissertations & Theses

1. Background

Francis Galton was an all-rounder of science that lived in the 19th century. Driven by his own curiosity he contributed to various fields like Geography, Meteorology, Psychology and Genetics. He was particularly enthusiastic about counting and quantifying everything. This obsession can not only be regarded as the stimulus for the foundation of scientometrics. It also radically changed social sciences which then increasingly relied on quantified measurements and

statistical methods. It was Galton who introduced regression, correlation and percentiles as new statistical concepts; simply to deal with the huge amount of data he accumulated (Obituary, 1911; Enciclopedia Italiana, 1950; Forrest, 1974; Gillham, 2001).

Galton was fascinated by the measurement of science (Godin, 2007). “Hereditary Genius” (1869) and “English Men of Science” (1874) are major contributions to this field. The first can be regarded as history’s first example of historiometry (Wikipedia, 2012), whereas the latter inspired Cattell to publish his directory “American Men of Science” (Cattell, 1906) more than thirty years later. Moreover Galton was a pioneer of mapping science. The results of his famous beauty map of the British Isles were compared to a beauty map of London recently compiled by Swami and Hernandez (2008).

Galton passed away in 1911, leaving more than 300 papers and almost 20 books for posterity.

This bibliometric analysis is the continuation of a previous study in appreciation of the crucial contributions of Galton to scientometrics.

2. Retrospective introduction

In a previous study (Gorraiz, Gumpenberger and Wieland; 2011) citation analysis of Galton’s works was done in Web of Science, Scopus and Google Scholar (Publish or Perish) in order to retrieve his most frequently cited books and journal articles. The retrieved book and journal article citations were extensively analysed. This citation analysis was then complemented by an analysis of references where Galton is rather mentioned than cited, a phenomenon generally known as obliteration by incorporation. In addition occurrences of Galton’s works were counted in major encyclopaedias, biographical indexes, in obituaries, Festschriften and the website Galton.org.

Correlation analyses of the most cited books with occurrences in biographical sources and encyclopaedias were performed using the Pearson correlation coefficient.

The major findings were that Galton’s works turned out to be increasingly cited or mentioned. The phenomenon of obliteration (i.e. the use of eponyms) applies to the remarkable proportion of up to one third of Galton’s works. Whether scientists are cited or rather only mentioned either depends on the respective subject field or on the country-specific cultural behaviour. Our findings suggest that obliteration is probably more common in selected subject areas like mathematics and statistics (formulas, processes, effects, etc. are named after a person) than e.g. in psychology.

3. Scope of the analysis

This subsequent bibliometric analysis comprises of the following new aspects:

- a. The data sources were expanded by introducing “ProQuest Dissertations & Theses (PQDT)” with regard to the aspect of “citation vs. obliteration”.
- b. All documents retrieved in Web of Science (WoS) citing or mentioning Galton were visualized in order to better illustrate the varied impact of

Galton's work and the significance of his individual contributions as well as to identify relevant networks. A selection of maps is presented in the results part.

- c. Correlation analyses between data sources WoS, Scopus and Google Scholar were expanded by PQDT for Galton's most cited books and journal articles.
- d. Calculation of the h-index and g-index in different databases and comparison with the h-index of other giants in similar studies was performed.

4. Methodology

4.1. Analysis in “ProQuest Dissertations & Theses (PQDT)”

One goal of the follow-up study was to extend our correlation analyses between occurrences and citations – so far only performed for monographs – also to journal articles. However, it became evident that these are hardly cited or mentioned in encyclopaedias and biographic indexes, whereas dissertations and theses have so far been neglected. “ProQuest Dissertations & Theses (PQDT)” recently allows searching for cited documents and therefore enabled the inclusion of theses and dissertations as important document types in our analyses (Andersen and Hammarfelt, 2011).

This complementary data source is advertised as being the world's most comprehensive collection of dissertations and theses. PQDT includes 2.7 million searchable citations to dissertation and theses from around the world from 1861 to the present day together with 1.2 million full text dissertations that are available for download in PDF format.

Searches in PQDT were done in March 2012 in order to identify the most cited and also the most mentioned documents (monographs and journal articles). The search included only the years 2006-2012, since citation data are not available in PQDT prior to 2006. Search strategy and manual disambiguation were similar to the procedures described in the previous study (Gorraiz, Gumpenberger and Wieland; 2011). Searches were done in title, descriptors, identifiers and abstracts, and despite of the fact that PQDT allows to search the full text, this feature was not used for comparability reasons.

4.2. Comparative view of “citation vs. obliteration”

The new results obtained from the PQDT analysis were combined with the results from the previous study performed in WoS and Scopus. A chart was produced in Excel to compare the citation-to-obliteration ratio for each database.

4.3. Visualization

Visualization was done with the freely available software packages BibExcel, VOSviewer and Pajek. BibExcel is a software intended to analyse bibliographic data to generate data files that can be imported to Excel (or any programme suitable to process tabbed data) for further processing (BibExcel, 2012). VOSviewer is intended to analyse bibliometric networks by creating, viewing

and exploring maps (VOSviewer, 2012). Pajek (Slovene word for Spider) is a Windows based programme for the analysis and visualization of large networks (Pajek, 2011).

Comparative maps were produced representing either WoS categories, WoS Author Keywords, WoS KeyWords Plus® (index terms created by Thomson Reuters from significant, frequently occurring words in the titles of an article's cited references) or WoS Abstracts for both the cited and the mentioned works of Galton.

4.4. Correlation analysis

The previous Pearson correlation analyses between the different data sources (WoS, Scopus, Google Scholar) for Galton's most cited monographs as well as journal articles were expanded by introducing PQDT as complimentary data source.

4.5. h-index and g-index

The h-index and the g-index were calculated in the different databases compared to the results obtained in similar studies (Marx, Cardona and Lockwood, 2011). For WoS data the h-index was furthermore calculated separately for the journal articles as well as for the monographs.

5. Results

Results from the comparison “citing” vs. “mentioning”

Table 1 shows the absolute numbers of retrieved documents in WoS, Scopus and PQDT, whereas Figure 1 is a comparative depiction of the citation-to-obliviation ratio expressed in percentages for each analysed database in the period 2006-2012.

50% of all mentioning documents in all databases refer to the “Galton-Watson-process(es)” and confirms that this is the most important eponym

Table 1: analysed documents in WoS, Scopus and PQDT

| | WoS (all) | WoS (2006- 2012) | Scopus (all) | Scopus (2006- 2012) | PQDT (2006- 2012) |
|-------------------------|--------------|------------------------|-----------------|---------------------------|-------------------------|
| citing documents | 4808 | 1234 | 3537 | 1826 | 326 |
| mentioning documents | 1215 | 363 | 994 | 345 | 19 |
| both | 214 | 74 | 194 | 101 | 2 |

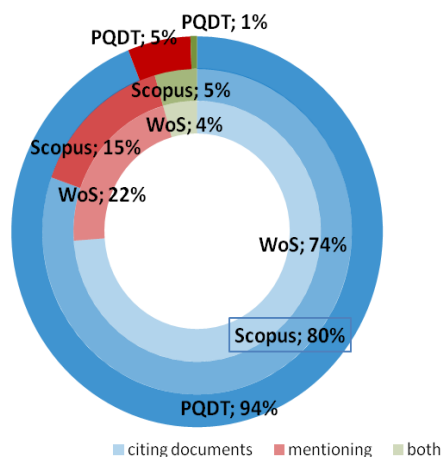


Figure 1: Comparison “citing” vs. “mentioning” for WoS, Scopus and PQDT (all for 2006-2012)

The citation-to-oblivation ratio is similar for both WoS and Scopus. Even fewer mentioning documents could be retrieved in PQDT.

Results from visualization:

Comparison citation vs. obliteration – WoS categories

The works citing Galton form two visible fronts as obvious from Fig. 2. The first (and main) cluster comprises of the WoS categories Psychology, Psychiatry, Behavioral Sciences and Genetics & Heredity. The second cluster includes Statistics & Probability, Computer Science, Social Sciences as well as Sociology/Demography. Both are linked to each other. In addition Anthropology, Zoology and Evolutionary Biology can be identified as isolated WoS categories.

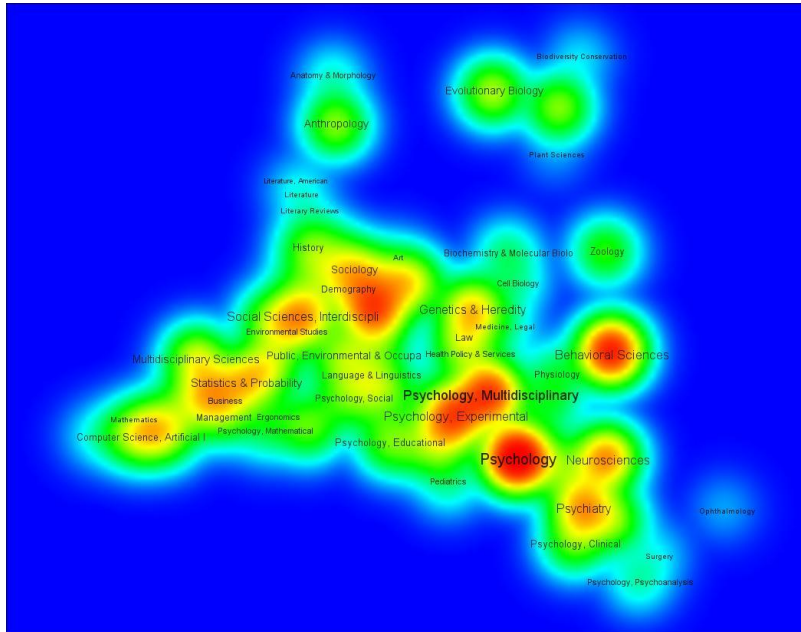


Figure 2. WoS categories of the works citing Galton (VOSviewer map)

Fig. 3 depicts the major WoS categories for the works mentioning Galton. Again two clusters are visible. The first cluster comprises of the WoS categories Statistics & Probability, Biology, Genetics and Social Sciences, the second one of Mathematics. Both clusters are not connected.

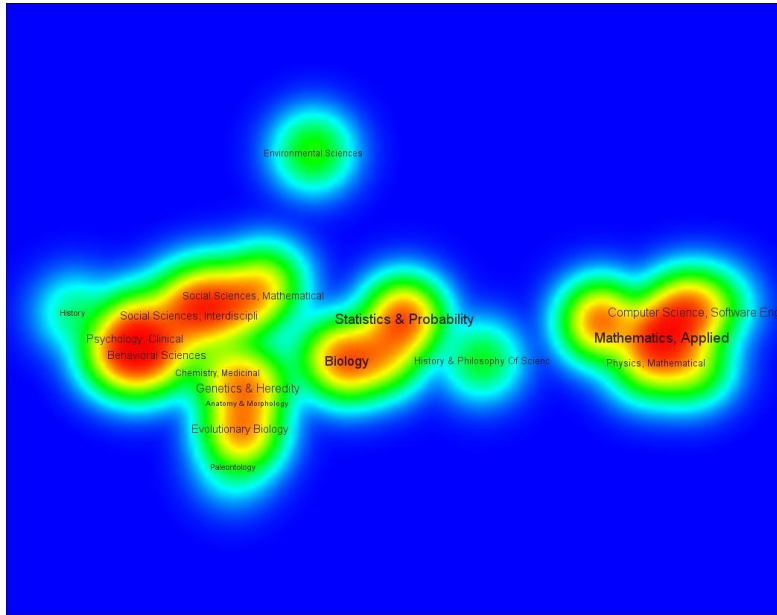


Figure 3. WoS categories of the works mentioning Galton (VOSviewer map)

Comparison citation vs. obliteration – WoS descriptors

Fig. 4 depicts the WoS author keywords of the works citing Galton. Intelligence, genetics, eugenics, correlation and biometrics stand out as concepts. In comparison Fig. 5 shows the major author keywords including the most important eponyms relating to Galton.

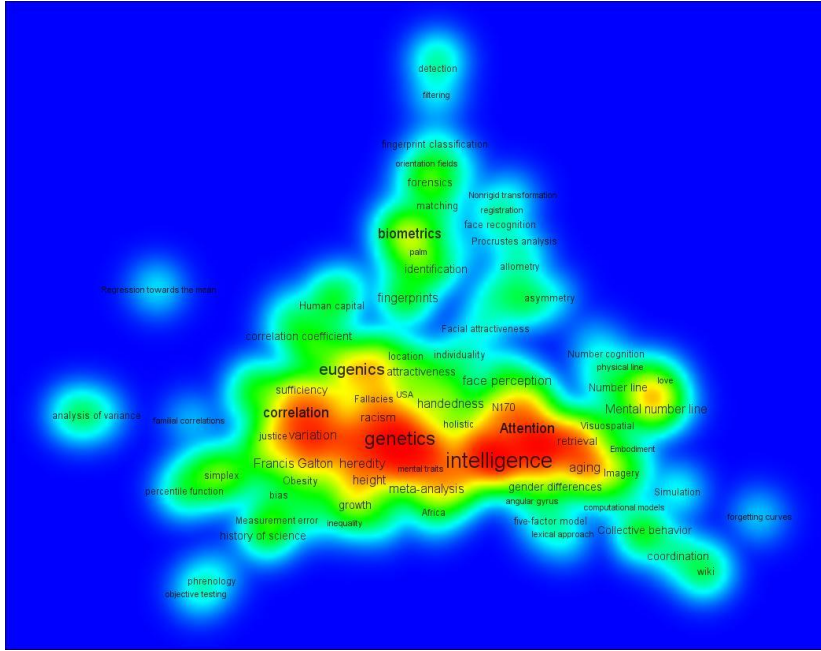


Figure 4. WoS author keywords of the works citing Galton (VOSviewer map)

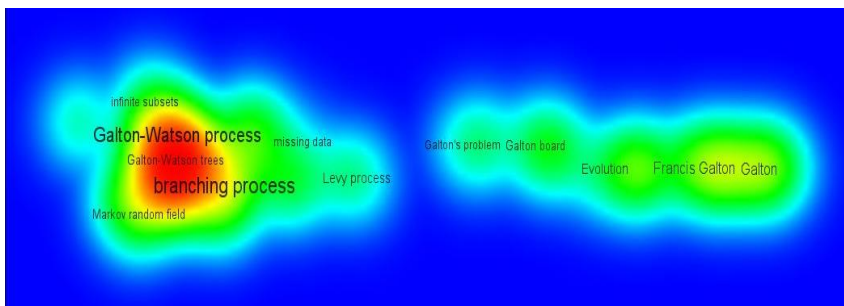


Figure 5. WoS author keywords of the works mentioning Galton (VOSviewer map)

Comparison citation vs. obliteration – WoS identifiers

Figure 6 and 7 allow a deeper and more complete insight by using the WoS KeyWords Plus® instead of the author keywords, which are not always available in WoS (included in WoS records of articles from 1991 forward). This way additional eponyms like “Galton's fallacy” can be identified (see Fig. 7).

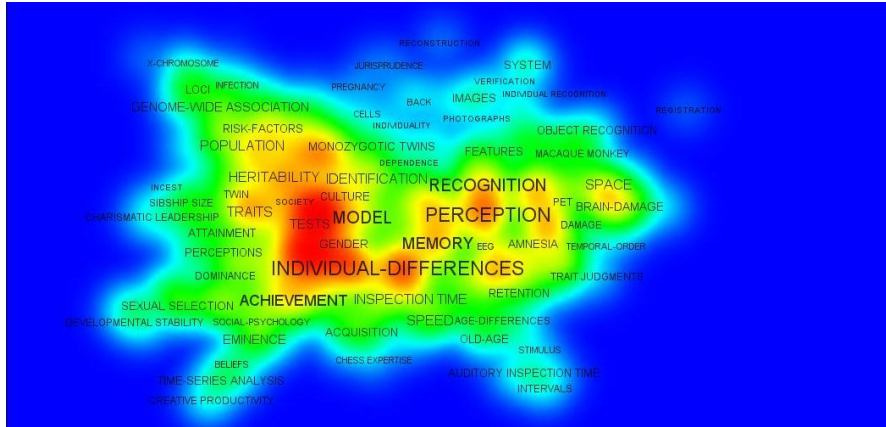


Figure 6. WoS KeyWords Plus® of the works citing Galton (VOSviewer map)

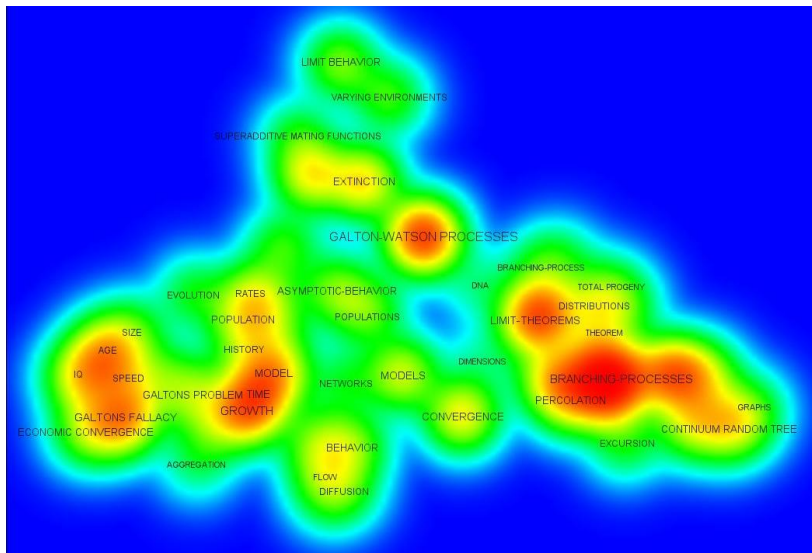


Figure 7. WoS KeyWords Plus® of the works mentioning Galton (VOSviewer map)

Results from correlation analysis of data sources for citations:*Monographs*

Table 2 shows the 15 most cited Galton's monographs in WoS, Google Scholar, Scopus and PQDT.

Table 2: Comparative view of Galton's top 15 cited monographs

| Rank WoS | Title (Abbreviation) | WoS | | PY MC E | Google Scholar | | | Scopus | | PQDT | |
|-------------|------------------------------|--|-------------------------------|---------------|------------------------------|-----------|-----------|-----------|-----------|------|----|
| | | Most correctly cited Ed. (MCE) | Most cited Ed. (MCE) | | Cits (C) to all Eds | Rank C | Rank C | Rank C | Rank C | | |
| 1 | INQUIRIES HUMAN FACU* | 673 | 839 | 1883 | 1066 | 11 | 13 | 2 | 351 | 2 | 40 |
| 2 | HEREDITARY GENIUS | 274 | 503 | 1869 | 912 | 1 | 2004 | 1 | 363 | 1 | 93 |
| 3 | NATURAL INHERITANCE | 274 | 356 | 1889 | 387 | 2 | 657 | 3 | 156 | 5 | 9 |
| 4 | ENGLISH MEN SCITHEI | 114 | 222 | 1874 | 252 | 3 | 377 | 4 | 80 | 4 | 15 |
| 5 | FINGER PRINTS | 154 | 213 | 1892 | 250 | 4 | 338 | 5 | 74 | 3 | 17 |
| 6 | MEMORIES MY LIFE | 54 | 111 | 1908 | 142 | 5 | 185 | 6 | 49 | 7 | 8 |
| 7 | ESSAYS EUGENICS | 31 | 48 | 1909 | 55 | 7 | 85 | 7 | 33 | 6 | 12 |
| 8 | NARRATIVE EXPLORER T** | 18 | 28 | 1853 | 52 | 6 | 101 | 8 | 19 | 9 | 3 |
| 9 | FINGER PRINT DIRECTO | 6 | 20 | 1895 | 24 | 9 | 21 | 14 | 1 | 12 | 0 |
| 10 | ART TRAVEL SHIFTS CO | 6 | 8 | 1855 | 24 | 8 | 31 | 11 | 5 | 8 | 5 |
| 11 | METEOROGRAPHICA METH | 8 | 12 | 1863 | 12 | 15 | 3 | 9 | 6 | 10 | 1 |
| 12 | GENIE VERERBUNG*** | 6 | 10 | 1910 | 10 | 10 | 15 | 12 | 4 | 12 | 0 |
| 12 | NOTEWORTHY FAMILIES | 9 | 10 | 1906 | 10 | 12 | 6 | 9 | 6 | 10 | 1 |
| 14 | RECORD | 3 | 3 | 1884 | 3 | 12 | 6 | 15 | 0 | 12 | 0 |

| | | | | | | | | | | | |
|----|---------------------|---|---|------|---|----|---|----|---|----|---|
| | FAMILY FACULTY | | | | | | | | | | |
| 15 | DECIPHERMENT BLURRS | 3 | 3 | 1893 | 3 | 14 | 5 | 13 | 2 | 12 | 0 |

Journals

Table 3 lists the 15 most cited Galton’s articles in WoS, Google Scholar, Scopus and PQDT. Considering Galton’s preferred publication strategy, i.e. to publish his results in one subject-specific as well as in one popular journal (Gorraiz, Gumpenberger and Wieland; 2011), publications corresponding to the same title were aggregated.

Table 3: Comparative view of Galton’s top cited journal articles

| Title of article | Rank WoS | Citations WoS | Rank GS | Citations GS | Rank Scopus | Citations Scopus | Rank PQDT | Citations PQDT |
|---|----------|---------------|---------|--------------|-------------|------------------|-----------|----------------|
| Regression towards mediocrity in hereditary stature | 1 | 176 | 1 | 347 | 1 | 120 | 3 | 9 |
| Psychometric experiments | 2 | 162 | 2 | 313 | 2 | 95 | 8 | 3 |
| The history of twins, as a criterion of the relative powers of nature and nurture | 3 | 156 | 3 | 237 | 5 | 34 | 5 | 5 |
| Hereditary talent and character | 4 | 126 | 4 | 220 | 4 | 57 | 2 | 12 |
| Visualised numerals | 5 | 121 | 8 | 131 | 3 | 85 | >10 | 2 |
| Composite Portraits | 6 | 97 | 5 | 162 | >10 | 6 | 6 | 4 |
| Correlations and their measurement | 7 | 80 | 7 | 157 | 9 | 27 | 6 | 4 |
| On the probability of the extinction of families | 8 | 75 | 6 | 159 | 8 | 28 | 4 | 7 |

| | | | | | | | | |
|---|-----|----|-----|----|-----|----|-----|----|
| The geometric mean | 9 | 68 | >10 | 88 | 10 | 26 | >10 | 0 |
| Typical Laws of Heredity | 10 | 63 | >10 | 84 | 11 | 25 | >10 | 1 |
| Statistics of mental imagery | 11 | 55 | 10 | 94 | 12 | 22 | 10 | 3 |
| Measurement of character | 12 | 54 | 0 | 0 | 6 | 33 | 1 | 16 |
| Statistical inquiries into the efficacy of prayer | 13 | 47 | 9 | 97 | 7 | 32 | 10 | 3 |
| Eugenics: Its Definition, Scope, and Aims | >13 | 15 | 10 | 94 | >12 | 13 | 6 | 4 |
| Personal Identification and Description | >50 | 1 | 11 | 90 | >12 | 8 | 6 | 4 |

The corresponding Pearson correlations between all used data sources are presented in Table 4.

The highest correlation is observed between Scopus und WoS when considering monographs, and between WoS and Google Scholar when considering journal articles.

Table 4: Pearson correlation coefficients between data sources

| | WoS / GS | WoS / Scopus | GS / Scopus | WoS / PQDT | Scopus / PQDT | GS / PQDT |
|------------------|----------|--------------|-------------|------------|---------------|-----------|
| Monographs | 0.610 | 0.992 | 0.689 | 0.849 | 0.890 | 0.875 |
| Journal Articles | 0.856 | 0.784 | 0.721 | 0.153 | 0.189 | 0.047 |

d) h-index and g-index

Table 5: comparative overview of h-index and g-index values

| | WoS | | Scopus | | GS | |
|---------|---------|---------|---------|---------|---------|---------|
| | h-index | g-index | h-index | g-index | h-index | g-index |
| overall | 25 | 69 | 21 | 43 | 30 | 81 |

| | | | | | | |
|-------------------|----|----|----|----|----|----|
| journals | 21 | 39 | 16 | 28 | 26 | 54 |
| monographs | 11 | 56 | 8 | 34 | 11 | 62 |

6. Discussion & Conclusions

Retrospective bibliometric studies always come with limitations. Citation analyses become challenging due to typos in publication years, different editions of books, different spellings of titles, titles changes or journals volumes covering two publication years (Gorraiz, Gumpenberger and Wieland; 2011). To make things worse the data sources used for the analyses are also far from perfect. Limited journal coverage, limitations of search fields, data base errors, translation errors, misspelled citations, complex author names, and complicated journal names have all been explicitly mentioned (Marx, 2011) and always need to be taken into account.

Since dissertations can be regarded as an underdeveloped source of analysis in bibliometric research (Andersen and Hammarfelt, 2011), PQDT was added to this analysis as a complementary data source. As outlined before, certain limitations are also evident for this data base. The reference search is not ideal, as citing documents are sometimes retrieved without available references (approximately 17%). These would only be accessible after purchasing the documents of interest.

On the positive side, PQDT would allow searching the full text, which could be a valuable feature to better understand and analyze the phenomenon of obliteration. However, the procedure is cumbersome and can only be explored in further studies.

In spite of the fact that the majority of works are preferably cited, no study would be complete without including the mentioned part. Visualization finally allows a better interpretation and understanding of both aspects and is especially helpful for the identification of relevant eponyms.

Maps can be interesting alternatives to other lengthy data processing procedures. They are valuable to show the relationships (networks) between the analyzed criteria and therefore certainly appeal to historians and biographers. However, not all maps are equally useful, e.g. in this study maps based on WoS Abstracts turned out to be meaningless for citing documents, and almost similar to the ones based on WoS author keywords for mentioning documents. It is furthermore very helpful to exclude meaningless terms before finally creating the maps.

The outstanding achievements of Galton become evident by simply recognizing the high impact for only his books. Galton's overall h-index of 25 is surprisingly high in comparison to the values determined for other historic scientists, e.g. 20 for Rutherford, 13 for Planck. According to a similar reference multiplier of 30

or 40 as introduced by Marx, Cardona and Lockwood (2011), Galton's present day scaled h-index would rather range between 113 and 116. Separately determined h-index values for monographs and journals articles (based on WoS) need to be taken with a pinch of salt due to the different amount of citations for both publication types. These shortcomings have therefore been addressed by also calculating the g-index based on the citation distribution.

Citation analysis and occurrence counting in biographical sources are considerable methods to study the history, philosophy and sociology of science. Both approaches support the retrieval of the most relevant or most influential works of outstanding scientists, and their combination even better allows the retrospective unmasking of a "giant's" publication strategy. The complementary use of both methods results in a merger of the objective nature of citations and the subjective peer perspective of a biographer.

Overall this type of retrospective study should prove librarians and other information specialists that bibliometric activities do not necessarily need to be restricted to evaluative purposes. Expertise can also be successfully applied to other interesting fields of research like the one presented in this study.

References

- Andersen, J. P. and Hammarfelt B. (2011). Price revisited: on the growth of dissertations in eight research fields. *Scientometrics* 88(2), 371-383. doi: 10.1007/s11192-011-0408-8
- BibExcel (2012). Retrieved March 22, 2012 from <http://www8.umu.se/inforsk/Bibexcel/>
- Cattell, J. M. (1906). *American Men of Science: A Biographical Directory*. New York: The Science Press.
- Enciclopedia italiana di scienze, lettere ed arti . Roma: Istituto della Enciclopedia Italiana , 1.1929 - 36.1939; App.1.1938; 2.1938/48(1948) - , -2000.
- Forrest, D. W. (1974). *Francis Galton: The Life and Work of a Victorian Genius*. New York, NY: Taplinger. 340pp.
- Galton, F. (1869). *Hereditary Genius*. London: Macmillan.
- Galton, F. (1874). *English Men of Science: Their Nature and Nurture*. London: Macmillan.
- Gillham, N. W. (2001). *A Life of Sir Francis Galton: From African Exploration to the Birth of Eugenics*. New York: Oxford University Press. 416pp.
- Godin, B. (2007). From eugenics to scientometrics: Galton, Catell, and men of science. *Social Studies of Science*, 37 (5), 691-728.
- Gorraiz, J., Gumpenberger, C., and Wieland, M. (2011): Galton 2011 revisited: a bibliometric journey in the footprints of a universal genius, *Scientometrics* 88(2), 627-652. doi: 10.1007/s11192-011-0393-y
- Historiometry (2012). Retrieved March 22, 2012 from <http://en.wikipedia.org/wiki/Historiometry>
- Marx W., Cardona M., and Lockwood D. J. (2011). Rutherford's scientific impact from a bibliometric perspective. *Australian Physics* 48(3), 78-83.
- Marx W. (2011). Special features of historical papers from the viewpoint of bibliometrics. *J Am Soc Inf Sci Technol* 62(3), 433-439.

- Marx W. and Cardona M. (2009). The citation impact outside references - formal versus informal citations. *Scientometrics* 80(1), 1-21.
- Obituary (1911). Sir Francis Galton D.C.L. D.Sc F.R.S. *Journal of the Royal Statistical Society*, Vol. 74, 3, 314-320.
- Pajek – Program for Large Network Analysis (2011). Retrieved March 22, 2012 from <http://pajek.imfm.si/doku.php>
- ProQuest <http://www.proquest.com/en-US/catalogs/databases/detail/pqdt.shtml>
- VOSviewer (2012). Retrieved March 22, 2012 from <http://www.vosviewer.com/>