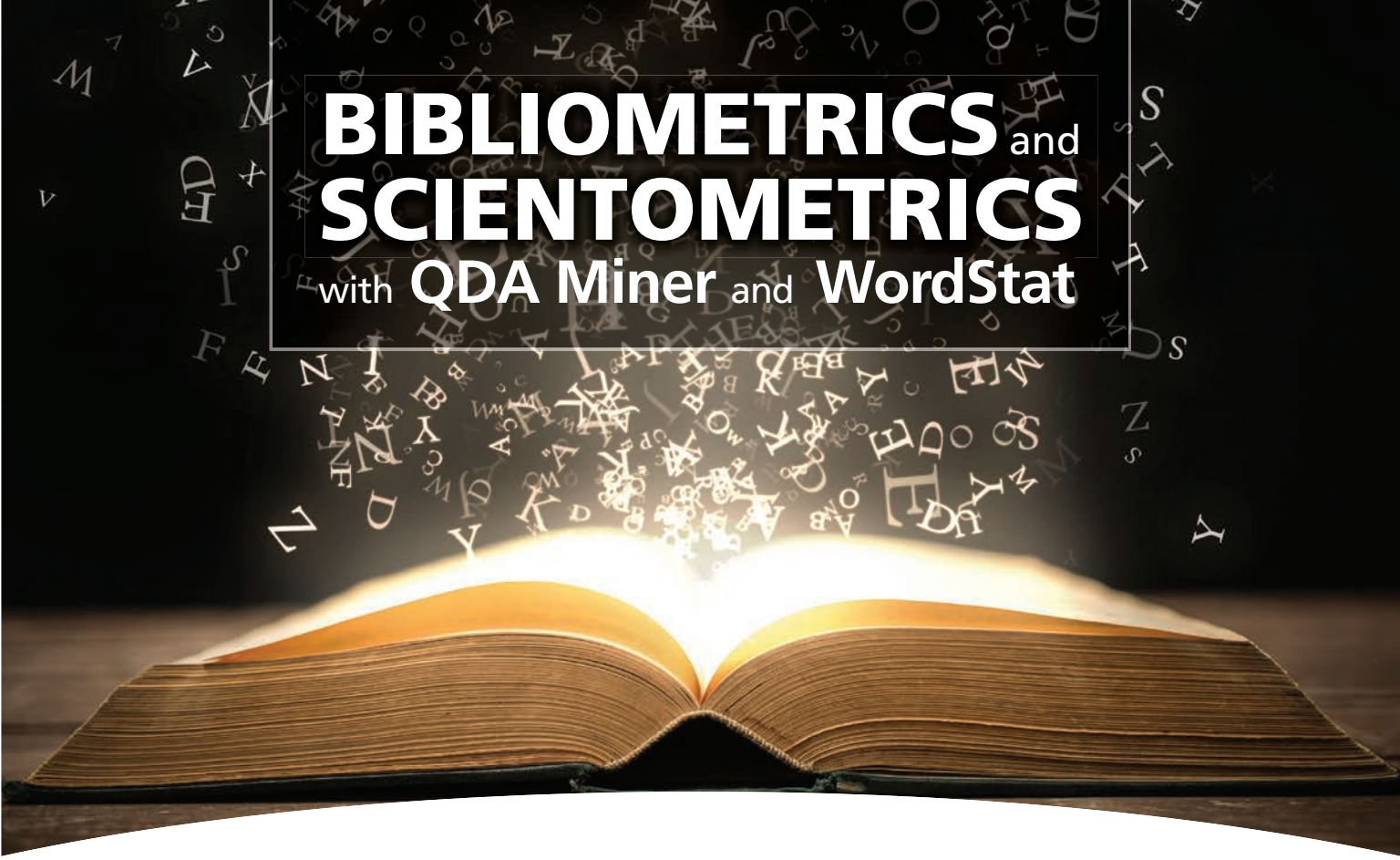

BIBLIOMETRICS and SCIENTOMETRICS

with QDA Miner and WordStat

What are Scientometrics and Bibliometrics?

Scientometrics and bibliometrics are methodological approaches in which the scientific literature itself becomes the subject of analysis. In a sense, they could be considered a science of science. Scientometrics researchers often attempt to measure the evolution of a scientific domain, the impact of scholarly publications, the patterns of authorship, and the process of scientific knowledge production. Scientometrics and bibliometrics involve the monitoring of research, the assessment of the scientific contribution of authors, journals or specific works, as well as the analysis of the dissemination process of scientific knowledge. Researchers in such approaches have developed methodological principles on ways to gather information produced by the activity of researchers' communications, and have used specific methods such as citation analysis, social network analysis, co-word and content analysis, as well as text-mining to achieve these goals. Many bibliometrics studies focus on authorship, or measure the contribution of journal and research organizations, but may also involve content analysis of words in titles, abstracts, the full text of books, journal articles or conference proceedings, or keywords assigned to published articles by editors or librarians.

PROVALIS RESEARCH

Examples of Studies using WordStat and QDA Miner

To celebrate the 25th anniversary of Human Communication Research, Stephen (1999) used WordStat and QDA Miner to analyze words in the titles of 634 articles and published his results in the final issue of the 25th volume. Using co-word analysis and hierarchical clustering, he was able to identify relationships among concepts as well as changes in topics studied over time. West (2007) did a similar analysis of words in the titles of 345 papers published in the International Journal of Advertising.

Abstracts are also good indicators to understand and grasp the content of publications. Lonchamp (2012) used WordStat to analyze abstracts from the International Journal of Computer-Supported Collaborative Learning (IJCSCL) to investigate and map its content. Other researchers have analyzed much larger corpora consisting of tens of thousands of journal abstracts on gene expression to identify ambiguous designation of genes in the published articles (Coimbra, Vanderwall and Oliveira 2010) or to reveal unexpected gene associations (Chaussabel & Sheer, 2002).

Keywords may be good indicators of a paper's content, and have often been used in scientometrics studies (Fratesi, 2008; Lonchamp, 2012; Reinhold, Laesser, & Bazzi, 2014). Fratesi (2008) used keywords and titles of 175,000 articles published in 68 geology journals between 1945 and 2000 to track changes in the influence of subdisciplines over time.

Many researchers have used WordStat to extract and analyze large corpora consisting of the full text of journal articles (Lonchamp, 2012; Anderson et al., 2007; Waismel-Manor, 2011). Waismel-Manor (2011) retrieved 1,317 articles from the Campaigns & Elections journal using Lexis-Nexis, obtaining a text corpus of 1,736,042 words, and was able to demonstrate the gradual professionalization of campaign consultants over time. Anderson, Joly and Fairhurst (2007) analyzed the full text of 149 articles published over a five-year period to document how retailers used business intelligence and data-mining tools to implement customer relationship management (CRM) in retailing.

Bibliometrics studies often focus on co-citation analysis, identifying the influence of authors and journals and the relationship among them. Jacobsen, Punzalan and Hedstrom (2013) used WordStat to analyze 165 articles on collective memory from four leading archival studies journals between 1980 and 2010 to identify which scholars and well-known works have been the most influential.

Scientometrics studies may focus on how concepts are defined over time or in different domains. Walterbush, Gräuler and Teuteberg (2014) identify similarities and differences in definitions of the word "trust" through literature of research spanning over 50 years. The authors collected a set of 121 definitions from various domains and analyzed those with a word-stem frequency analysis in WordStat as well as with a qualitative data analysis using QDA Miner.

QDA Miner: Import from databases, spreadsheets, References Information System (RIS) files created by online journal databases. Import full-text articles in PDF, HTML, or MS Word.

WordStat: Automatically identify topics and themes, or focus on specific dimensions with user-defined content analysis dictionaries. Compare across different sources or look for changes over time.

Software features for Scientometrics and Bibliometrics

Several key features of QDA Miner and WordStat are very useful for scientometrics and bibliometrics studies.

DATA IMPORTATION: The ability to import MS Word, RTF, HTML, as well as PDF files, and to associate metadata (such as dates, numerical and categorical data) with those articles, allows one to easily create a corpus of full-text articles with relevant variables. QDA Miner's ability to import Reference Information System (RIS) data files is convenient for importing information from journal databases like ProQuest or bibliographic software such as EndNote or Reference Manager. The Document Conversion Wizard is especially useful to import data from databases such as Lexis/Nexis, for splitting single documents into multiple ones, or for extracting variables from structured listings or reports.

EDITING, TAGGING AND ANNOTATING: Once imported into QDA Miner, documents may be edited, coded and annotated manually, allowing one to perform content analysis with WordStat while ignoring irrelevant sections or focusing on specific ones. For example, manually tagging the reference sections of journal articles could allow one to perform co-citation analysis, while by manually coding the research method sections of journal articles, one could also describe the evolution of research methods over time or compare methods used in different journals.

TEXT PRE-PROCESSING: WordStat's ability to transform words into stems, to lemmatize and remove words of little semantic value (like prepositions, conjunctions or pronouns) allows one to quickly focus on more relevant words and phrases.

WORDS AND PHRASES EXTRACTION: WordStat can process up to 300,000 words per second and quickly produce frequency counts of significant words, extract common phrases, and produce visual displays in the form of bar charts, word clouds, and more.

ANALYSIS OF CO-OCCURRENCE: The analysis of co-occurrences using statistical techniques such as hierarchical clustering, multidimensional scaling and visualization tools like the proximity plot, allows one to promptly identify topics and themes in a discipline. Such tools in WordStat have often been used for mapping scientific domains (Fratesi & Vacher, 2008; Friedman & Smiraglia, 2012; Lonchamp, 2012; Reinold, Lasser & Bazzi 2014).

COMPARATIVE ANALYSIS: The ability to compare frequencies of words, phrases or content categories across different sources (e.g., journals, countries) or to look for changes over time could be used to identify the evolution of a scientific discipline, the rise and fall of specific ideas or concepts, or to document the differentiation process of scientific publications or the geo-spatial distribution of scientific activities. One could compute simple statistical tests (like chi-square tests, F-tests or correlations), create presentation-quality visualizations (such as bar charts, line charts, bubble charts or heatmaps), and apply correspondence analysis.

APPLICATION OF CONTENT ANALYSIS DICTIONARIES: The possibility in WordStat to build dictionaries of key words, key phrases and proximity rules allows one to focus on specific dimensions. For example, one may easily build a dictionary of authors or journals and perform co-citation pattern analysis. One may also create dictionaries to group together key terms into broader concepts, to measure the prevalence of methodological traditions, theories, research topics, etc.

KEYWORD-IN-CONTEXT: The Keyword-in-Context (or KWIC) feature is an essential tool to test the validity of existing or user-built dictionaries by making sure words or phrases used to measure reference to specific topics are effectively capturing the intended meaning. When an item is found to be ambiguous, KWIC lists are also useful to identify proper disambiguation rules.

References of Scientometrics studies using QDA Miner and WordStat

Anderson, J., Jolly, L.D., & Fairhurst, (2007). Customer relationship management in retailing: A content analysis of retail trade journals. *Journal of Retailing and Consumer Services*, *14*(6), 394-399.

Araujo, R.F., & Olivera, M. (2015). Technological Basis for Information Science in Brazil: A Scientometric Study. *Qualitative and Quantitative Methods in Libraries*, *4*, 231-241.

Coimbra, R.S., Vanderwall, D.E., & Oliveira, G.C. (2010). Disclosing ambiguous gene aliases by automatic literature profiling. BMC Genomics 2010, 11(Suppl 5):S3.

Dabic, M., González-Loureiro, M., & Harvey, M. (2013). Evolving research on expatriates: what is 'known' after four decades (1970–2012). *The International Journal of Human Resource Management*, *26*(3), 316-337

Dabic, M., González-Loureiro, M., & Furrer, O. (2014). Research on the strategy of multinational enterprises: Key approaches and new avenues. *Business Research Quarterly*, *17*, 129-148.

de-Miguel-Molina, B., Chirivella-González, V., & García-Ortega, B. (2015). Corporate philanthropy and community involvement. Analysing companies from France, Germany, the Netherlands and Spain. Quality & Quantity, 1-26.

Fisher, I.E., Gernsey, M.R., Hughes, M.E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*. Online.

Forrester, A. (2015). Barriers to Open Access Publishing: Views from the Library Literature. *Publications*, *3*(3), 190-210.

Fratesi, S.E. & Vacher, H.L. (2008). Scientific journals as fossil traces of sweeping change in the structure and practice of modern geology, *Journal of Research Practice*, *4*(1), 1-23.

Friedman, A., & Smiraglia, R.P. (2012). Nodes and arcs: concept map, semiotics, and knowledge organization. *Journal of Documentation*, *69*(1), 27-48.

Jacobsen, T., Punzalan, R.L, & Hedstrom, M.L (2013). Invoking "collective memory": mapping the emergence of a concept in archival science. *Archival Science*, *13*(2): 217-251.

Kim, J. (2015) Growth and trends in digital curation research: The case of the International Journal of Digital Curation. *American Society for Information Science*, 1-4.

Lonchamp, J. (2012). Computational analysis and mapping of ijCSCL content. *International Journal of Computer-Supported Collaborative Learning*, *7*(4), 475-497.

Milojevic, S. (2012). Multidisciplinary cognitive content of nanoscience and nanotechnology. *Journal of Nanoparticle Research*, *14*(1), 1-28.

Milojevic, S. (2015). Quantifying the cognitive extent of science. *Journal of Informetrics*, *9*(4), 962-973.

Milojevic, S. & Leydesdorff, L. (2012). Information metrics (iMetrics): a research specialty with a socio-cognitive identity? *Scientometrics*, *95*, 141–157.

Milojevic, S., Sugimoto, C.R., Yan, E., Ding, Y. (2011). The cognitive structure of library and information science: Analysis of article title words. *Journal of the American Society for Information Science and Technology*, *62*(10), 1933–1953.

Oleinik, A. (2012). Publication patterns in Russia and the West compared. *Scientometrics*, *93*(2), 533-551.

Oleinik, A. (2014). Conflict(s) of Interest in Peer Review: Its Origins and Possible Solutions. *Science and Engineering Ethics*, *20*(1), 55-75.

Reinhold, S., Laesser, C., & Bazzi, D. (2014). *The intellectual structure of transportation management research: A review of the literature*. 14th Swiss Transport Research Conference. Monte Verità / Ascona.

Smiraglia, R.P. (2015). Domain Analysis for Knowledge Organization: Tools for Ontology Extraction. Amsterdam: Chandos Publishing.

Stephen, T. (1999). Computer-assisted concept analysis of HCR's first 25 years. *Human Communication Research*, *25*(4), 498-513.

Stephen, T. (2000). Concept analysis of gender, feminist, and women's studies research in the communication literature. *Communication Monographs*, *67*, 193-214.

Talamini, E. & Dewes, H. (2012). The Macro-environment for liquid Biofuels in Brazilian science and public policies. *Science and Public Policy*, *39*(1), 13-29.

Tseng, Y.-H. & Tsay, M.-Y. (2013). Journal clustering of library and information science for subfield delineation using the bibliometric analysis toolkit: CATAR *Scientometrics*, *95*, 503-528.

Walterbusch, M., Gräuler, M., & Teuteberg, F. (2014). How Trust is Defined: *A Qualitative and Quantitative Analysis of Scientific Literature*. 20th Americas Conference on Information Systems, Savannah, Georgia, USA.

Waismel-Manor, I. (2011). Spinning forward: Professionalization among campaign consultants. *Journal of Political Marketing*, *10*(4), 350-371.

West, D. (2007). Directions in marketing communication research: An analysis of the international journal of advertising. *International Journal of Advertising*, *26*(4), 543-554.

Zhang, H., Babar, M.A., & Tell, P. (2011). Identifying relevant studies in software engineering. *Information and Software Technology*, *53*(6), 625-637.

DOWNLOAD A TRIAL VERSION AT: provalisresearch.com/trial

PROVALIS RESEARCH

To schedule a web demo or for more information on our products,
contact us: TOLL FREE **1 855 355-5252** or **1 514 899-1672**
sales@provalisresearch.com, provalisresearch.com