“A Trace of this Journey”: Citations of Digitised Newspapers in UK History PhD Theses

Paul Matthew Gooding
p.gooding@uea.ac.uk
University of East Anglia, United Kingdom

> “In two weeks, despite these notes, I shall no longer believe in what I am experiencing now. One must leave behind a trace of this journey which memory forgets” (Cocteau, 2013).

Academic citations are prostheses for the scholarly memory, providing traces of a text’s origins. They are also, taken collectively, a powerful source of information on scholarly influence, links between authors, and academic publishing trends. This paper will present work in progress to discover the extent to which citation patterns by UK historians have been affected by digitisation of historical newspapers. Previous studies into digital resources have used citation analysis for impact analysis (Meyer *et al.*, 2009), but faced problems gathering accurate data due to researchers’ unwillingness to cite digital resources. Text mining mentions of titles within a digital resource offers a solution to this problem; indeed, Milligan (2013) has successfully used techniques from Natural Language Processing to track citations of major Canadian newspapers in Canadian PhD theses. However, there are local variations around academic practice, and cultural heritage digitisation, and to date there has been no large-scale study of digital resource citations in the United Kingdom.

This paper will present my efforts to mine newspaper citation trends using over 6,000 history theses submitted at UK Higher Education institutions, from 1999 to 2015. It will also consider the implications of text mining using legal deposit library collections. Its significance is therefore twofold. It is the first study to use text mining to track citations in UK history theses, thereby providing insights into the effect of local digitised primary sources. Second, by collaborating with British Library Labs, it provides an important test case of the possibility of exploring the text mining exception in UK copyright law. This study therefore focuses on two key research questions:

- What does citation analysis of UK history theses tell us about the impact of historic newspaper digitisation on early career historians in the United Kingdom?
- How far does the ability to text and data mine copyrighted materials provide for data driven approaches when applied to legal deposit library collections?

Research context

In the last fifteen years several studies have proposed models for evaluating the impact of digital resources (Warwick *et al.*, 2006; Meyer *et al.*, 2009; Tanner, 2012). Citation analysis is commonly used in impact evaluation, and is a well-established bibliometric technique for impact analysis (MacRoberts and MacRoberts, 1989). As part of the wider field of bibliometrics, citation analysis has been used to judge the impact of academic publications and digital resources (Meyer *et al.*, 2009). Citations are proxy measures of how frequently a document or resource is used, founded on the assumption that there is a strong positive correlation between the number of citations that a resource or article attracts, and the quality of that resource (Smith, 1980). The reality, though, is that citation practices are not always followed or understood by academics. Smith (1980) noted several reasons a scholar may choose not to cite a document: inability to obtain the document; inability to read a foreign language; lack of relevance to their work; or lack of awareness of existing work. There are further reasons for digital resources; not least a lack of awareness of how to cite them (Sukovic, 2009), and disciplinary unwillingness to acknowledge their usage (Meyer, 2009). As a result, the traditional method of mining only citations provide an unreliable picture of citation levels of digital resources.

Furthermore, there is a need to account for the varied local and national context within which researchers operate. In Canada, for instance, there is a feeling among scholars that the “limited and fragmented” (Kheraj, 2014) newspaper digitisation programme lags behind nations such as the USA, Australia and United Kingdom. The prominence of *The Toronto Star* and *The Globe and Mail* mirrors the early years of newspaper digitisation in the United Kingdom, where the ubiquity of the *Times Digital Archive* encouraged some to overstate its representativeness (Bingham, 2010). Contemporary digitised newspaper resources in the UK by contrast, tend to aggregate dozens of newspaper titles into a single resource. This paper therefore explores the likelihood that this aggregation process may have caused different citation patterns among UK researchers, who are the largest group to access these resources (Gooding, 2014), providing a comparison with the differing Canadian context presented by Milligan.

Methodology

To achieve this, I will focus on newspaper titles from two British Library resources: *British Library Nineteenth Century Newspapers (BNCN)* and *The British Newspaper Archive (BNA).* I intend to identify mentions of newspapers by title within the full text of UK history theses. The dataset comes from EThOS, a national service which makes UK doctoral theses available online for searching and reading. EThOS contains approximately 440,000 records relating to theses awarded by over 120 institutions. It provides a comprehensive, systematically collected dataset for comparison of citation trends over time. Around 160,000 records provide access to searchable full text, and I have worked with British Library Labs to identify a subset of history theses published from 1999 to 2015; in total, over 6,000 theses were identified using Dewey Decimal Classifiers, covering eight years before and after the launch of BNCN in 2007. These theses will be searched using Natural Language Processing techniques to identify the incidence of specific newspaper titles that are included in BNCN and BNA, allowing me to identify how many theses use a given source, and how frequently each source was used.

This project also acts as a test case for text mining of British Library collections. In 2014, the UK government introduced an exception to copyright law to ensure that researchers undertaking text and data-mining for non-commercial purposes would no longer infringe copyright (Intellectual Property Office, 2014), without requiring that publishers took steps to guarantee the availability of suitable datasets for text mining. In reality, this means that there are many potentially valuable data sources that could be legally studied, but no infrastructure to do so. Starting with the British Library's PhD thesis holdings, I intend to work with British Library Labs to explore the possibility of opening up further datasets for text mining.

Conclusion

This paper will explore the ways in which historical newspaper digitisation have impacted upon historiography among early career researchers in the United Kingdom, by tracking citations of digitised newspaper titles over time in the full text of over 6,000 PhD theses. This is the first study to apply text mining of digital resource citations to the UK context. It also provides an important case study of text and data mining in legal deposit library collections, in the light of current limitations to the UK copyright exception. In doing so, this project will not only illuminate the ways in which digital resources can affect local research practices, but will demonstrate the utility of text mining in addressing the methodological limitations of citation analysis for digital resources. We must adopt new computational methods to ensure that the traces of this use are not wiped away by citation practices which continue to de-emphasise the role of digital resources in contemporary research.

References

Bingham, A. (2010). 'The Digitization of Newspaper Archives: Opportunities and Challenges for Historians.' *Twentieth Century British History*, 21(2): pp. 225–231. doi: 10.1093/tcbh/hwq007.

Cocteau, J. (2013). *Opium: The Diary of His Cure*. 3rd Edition. London: Peter Owen Publishers.

Intellectual Property Office (2014). *Exceptions to Copyright: Research, UK Government*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf.

Kheraj, S. (2014). 'Canada's Historical Newspaper Digitization Problem, Part 2.' *History Matters*. http://activehistory.ca/2014/02/historical-newspaper-digitization-problem/comment-page-2/.

MacRoberts, M. H. and MacRoberts, B. R. (1989). 'Problems of Citation Analysis: A Critical Review.' *Journal of the American Society for Information Science*, 40(5): pp. 342–349.

Meyer, E. T. (2009). *Software Tools for Bibliometrics, Toolkit for the Impact of Digitised Scholarly Resources (TIDSR)*. Available at: http://microsites.oii.ox.ac.uk/tidsr/kb/49/software-tools-bibliometrics.

Meyer, E. T., Eccles, K., Thelwall, M. and Madsen, C. (2009). *Usage and Impact Study of JISC-Funded Phase 1 Digitisation Projects & the Toolkit for the Impact of Digitised Scholarly Resources (TIDSR)*. Oxford: Oxford Internet Institute, University of Oxford. Available at: http://microsites.oii.ox.ac.uk/tidsr/sites/microsites.oii.ox.ac.uk.tidsr/files/TIDSR_FinalReport_20July2009.pdf.

Milligan, I. (2013). 'Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997-2010.' *The Canadian Historical Review*, 94(4): pp. 540–569.

Smith, L. (1980). 'Citation Analysis.' *Library Trends*, 30: pp. 83–106.

Sukovic, S. (2009). 'References to e-texts in academic publications', *Journal of Documentation*, 65(6), pp. 997–1015.

Tanner, S. (2012). *Measuring the Impact of Digital Resources: The Balanced Value Impact Model*. London: King's College London. Available at:

http://www.kdcs.kcl.ac.uk/fileadmin/documents/pubs/BalancedValueImpactModel_SimonTanner_October2012.pdf.

Warwick, C., Terras, M., Pappa, N. and Galina, I. (2006). *The LAIRAH project: log analysis of digital resources in the arts and humanities - Final report to the Arts and Humanities Research Council*. School of Library, Archive and Information Studies, University College London. http://www.ucl.ac.uk/infostudies/claire-warwick/publications/LAIRAHreport.pdf.