Contributors:
Seth Cayley (Head of Research Solutions), Cengage Learning EMEA
Mike Gardner (Analyst and Web Developer in Web Technologies), University of Nottingham
Kat Gupta (Corpus Protocols Researcher), University of Nottingham
Michaela Mahlberg (Professor of English Language and Linguistics), University of Nottingham
Neil Smyth (Faculty Team Leader – Arts), University of Nottingham
Stella Wisdom (Curator), Digital Research at the British Library

Presenter:
Neil Smyth (Faculty Team Leader – Arts), University of Nottingham

Title:
Corpus Protocols: digital transformations of commercial newspaper collections for text and data mining to support academic research

Abstract:
This paper reports on outcomes from the *Corpus Protocols* project that investigated opportunities and challenges for data and text mining commercial news content, including licensing, storage and accessibility of data, and communication between researchers and publishers[1]. Related research in digital humanities and digital libraries has concentrated on methodologies, overviews and accessibility[2345]. This paper focuses on research in corpus linguistics based on digital news data, including the *Declassified Documents Reference System* and the *Times Digital Archive*. *Modernising Copyright* and published draft regulations for UK legislation indicate new opportunities for libraries to support research using digital newspapers[67]. Some publishers have pledged to bring more content online[8], and others are exploring sustainable commercial models for widening access to big data. This paper explores issues of research data management, innovations in web technologies for big data, and how research based on this kind of data satisfies requirements for the security of commercial news data in the context of emerging legislation. We identify potential conflicts this raises for research libraries and researchers.

---

[1] Corpus Protocols is funded by Horizon Digital Economy Research (http://www.horizon.ac.uk), and is part of the DAM network: University of Nottingham, "Data-Asset-Method," http://www.nottingham.ac.uk/humanities/digital/dam.aspx.
[2] Debora Cheney, "Text Mining Newspapers and News Content: New Trends and Research Methodologies," http://library.ifla.org/233/1/153-cheney-en.pdf.
[3] Ann Okerson, "Text & Data Mining - a Librarian Overview," http://library.ifla.org/252/1/165-okerson-en.pdf.
[4] R.B. Allen, "Improving Access to Digitized Historical Newspapers with Text Mining, Coordinated Models, and Formative User Interface Design" (paper presented at the IFLA International Newspaper Conference: Digital Preservation and Access to News and Views, New Delhi, 2010).
[5] M. B. O'Donnell et al., "Exploring Text-Initial Words, Clusters and Concgrams in a Newspaper Corpus," *Corpus Linguistics and Linguistic Theory* 8, no. 1 (2012).
[6] Intellectual Property Office, "Modernising Copyright: A Modern, Robust and Flexible Framework; Government Response to Consultation on Copyright Exceptions and Clarifying Copyright Law," http://www.ipo.gov.uk/response-2011-copyright-final.pdf.
[7] "Data Analysis for Non-Commercial Research," http://www.ipo.gov.uk/techreview-data-analysis.pdf.
[8] European Commission, "Licences for Europe: ten Pledges to Bring More Content Online," http://ec.europa.eu/internal_market/copyright/docs/licences-for-europe/131113_ten-pledges_en.pdf.

Biographies:

Seth Cayley is the Head of Research Solutions at Cengage Learning EMEA, responsible for publishing the company's programme of digital archives, such as the Times Digital Archive. Seth is especially interested in research-informed teaching, and how primary source archives can be used for study in the classroom. His involvement in the Data-Asset-Method project has convinced him of the potential for transforming the humanities if text and data mining methodologies can be taught more widely.

Mike Gardner is an analyst / web developer in Web Technologies at the University of Nottingham. His works focuses on usability and GUI design. Current interests include online mapping (eg Key to English Place-Names, http://kepn.nottingham.ac.uk/), displaying digitised text/images (eg Electronic Gateway for Icelandic Literature, http://egil.nottingham.ac.uk/), Windows on War, (http://windowson-war.nottingham.ac.uk/) and faceted search. Previous projects encompass library catalogues, e-prints, reading lists, portals/intranets, online elections and mobile apps.

Kat Gupta is the researcher on the Corpus Protocols project. She is a PhD researcher in the School of English, University of Nottingham. Kat's background is in corpus linguistics with interests in critical discourse analysis, digital humanities and interdisciplinary work, particularly points of contact between linguistics and history, gender and queer studies, and politics.

Michaela Mahlberg is Professor of English Language and Linguistics at the University of Nottingham where she is also the Director of the Centre for Research in Applied Linguistics (CRAL). Michaela's research focuses on the analysis of patterns and meanings in 'corpora', i.e. large collections of computer-readable data designed for linguistic purposes. She is the General Editor of the *International Journal of Corpus Linguistics* (John Benjamins), and Co-editor of the series *Corpus and Discourse* (Bloomsbury). Her publications include *Corpus Stylistics and Dickens's Fiction* (Routledge, 2012), *Text, Discourse and Corpora. Theory and Analysis* (Continuum 2007, jointly with Michael Hoey, Michael Stubbs and Wolfgang Teubert), and *English General Nouns: a Corpus Theoretical Approach* (John Benjamins, 2005). Michaela is an elected member of the Executive Board of ICAME (International Computer Archive of Modern and Medieval English) and will be hosting the annual ICAME conference at the University of Nottingham in 2014.

Neil Smyth is Faculty Team Leader (Arts) at the University of Nottingham. Based in the Hallward Library, he is responsible for working with senior representatives in the Arts Faculty to shape the strategic direction of library services. Neil has previously presented at national and international conferences on a range of themes, including management and leadership, emerging technologies in academic libraries, innovative industry partnerships and the future of the book. Neil is interested in data, and the data behind databases, and how this data can be made available for research and teaching; new business models and new legal frameworks and how they will shape future library services are Neil's current professional interests.

Stella Wisdom is a curator in the Digital Research team at the British Library; her role explores and promotes new methods of research using both born digital content and digitised collections. Stella's professional interests include interactive fiction, gamification, web archiving, online user engagement and crowdsourcing. Stella has worked for the British Library for eight years and prior to working in Digital Research, she managed Collection Storage at the British Library's site at Boston Spa in Yorkshire. Stella has also previously worked at the Library and Information Statistics Unit based at Loughborough University, the Warburg Institute Library and the National Library of Scotland.