Liberating Data: How libraries and librarians can help researchers with text and data mining.

blogs.lse.ac.uk/impactofsocialsciences/2016/07/12/how-libraries-and-librarians-can-help-with-text-and-data-mining/

With advances in computational methods and the proliferation of data sources, text and data mining offers exciting new directions for research. **Neil Stewart**, **Jane Secker**, **Chris Morrison** *and* **Laurence Horton** *look at the role of libraries in providing support to researchers for these projects, particularly to help with rights issues and to digitise material for scholarly re-use. Librarians should be bold about the advice they give to researchers and encourage them to use the new copyright exceptions.*

Recent alterations to copyright legislation, and in particular changes in the UK following the Hargreaves Review of intellectual property law, have removed legal barriers for researchers who wish to perform text and data mining (TDM) on the research literature corpus. This change has opened up a previously prohibited form of research, and holds the promise of being able to answer entirely new research questions that were previously not amenable to enquiry; Ross Mounce wrote about these changes and what these new forms of research might look like on this blog. This post examines the ways in which libraries and librarians can facilitate the work of researchers who might want to apply TDM methods to library resources, be they electronic or hard copy sources. It also advocates for librarians to be bold about the advice they give to researchers to encourage them to use the new copyright exceptions. This may at times mean directly contravening the terms and conditions of a licence with an electronic resource supplier.

What might researchers want to mine?

TDM is generally performed on corpora of text or data in electronic form. For example, a body of chemistry research literature might be mined by running TDM software on it to discover then create a database of molecular structures, a valuable resource for molecular chemists, crystallographers and other scientists. Alternatively, to use an example from the Digital Humanities, a large body of newspapers from the Victorian era can be mined to extract jokes, which can be used to analyse aspects of Victorian culture and social history. It is not just electronic corpora that can be mined, though- and we provide an example of digitisation of hard copy for TDM purposes below.

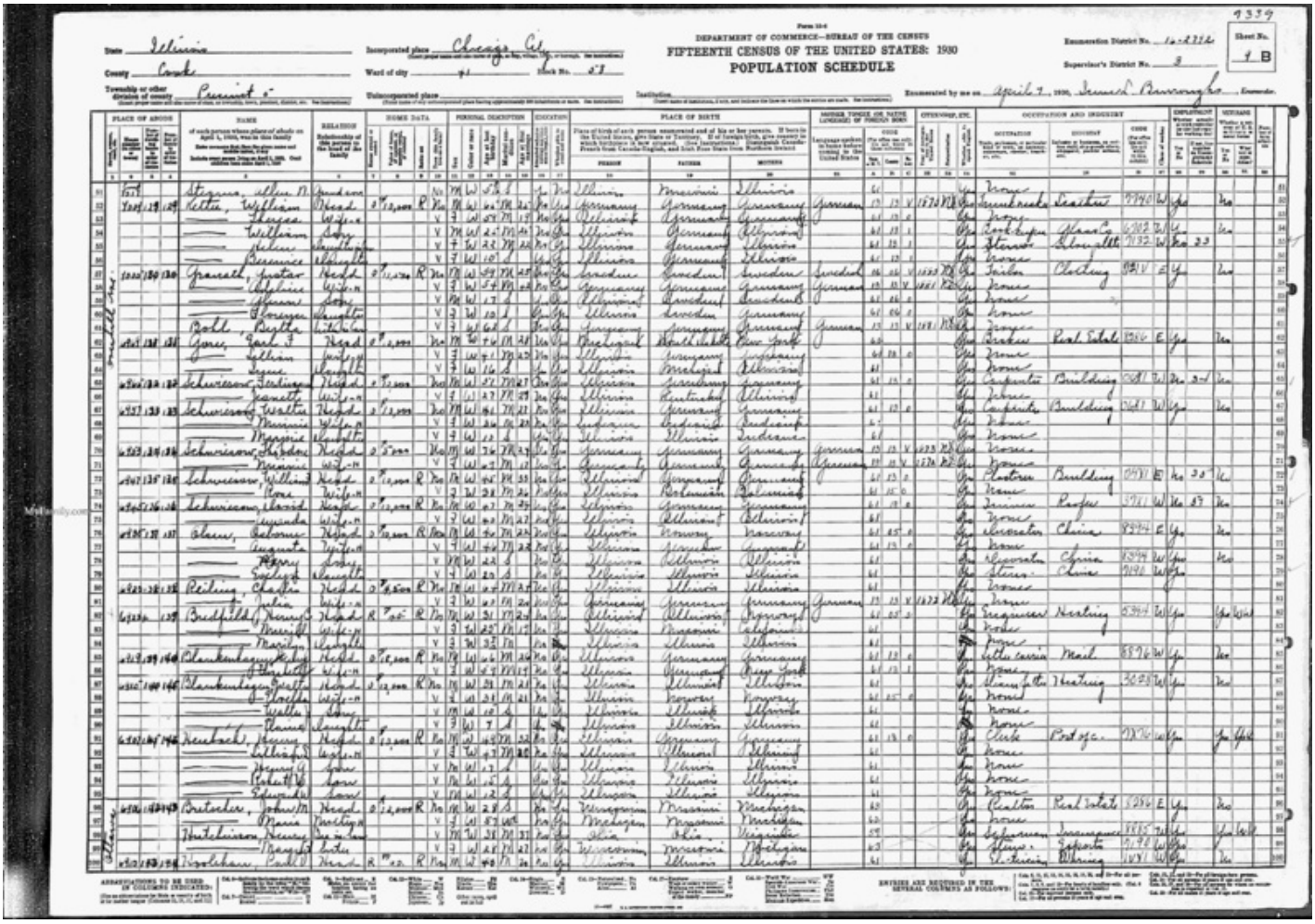

Image credit: 1930 US Census for Chicago, Illinois with Osborne Theomun Olsen (1883-1971) Wikimedia Public Domain

What problems might researchers encounter?

So how might a researcher get started with a TDM project? A good port of call is your library service, which will be able to offer ideas and pointers on firstly getting access to the data set you wish to mine, and then the other areas of your university that might also be able to assist. Some of the larger UK universities now have units dedicated to digital humanities, and of course Computer Science and Information Science departments and IT services are all likely to have in-house expertise.

There are a few specific problems that researchers might encounter when doing a TDM project. These relate to the way in which researchers get hold of material to perform analyses and include:

- Reaching a download limit (often arbitrary and relating to contractual stipulations made by the publisher) for papers from a particular publisher or database, at which point access can be cut off.
- Being served with an “unusual behaviour” report from publishers, for example when programmatically downloading material.
- Encountering Digital Rights Management (DRM) technologies that prevents the researcher using TDM methods, which then require a request for permission to have the DRM removed.

In each of these cases it is likely to be the library that manages access to subscribed sources, so researchers may well get in touch with librarians for support.

How can librarians help?

There are a number of things library services and librarians can do to assist with TDM services. These include:

- Advertise TDM support as a service, perhaps by coupling it with the related fields of Research Data Management or Digital Humanities support services.
- Encourage the development of partnerships with academic colleagues to work on TDM projects.
- Be clear about licensing and terms & conditions under which resources are made available to colleagues at your institution, and get to understand the legal exception that permits TDM for the purposes of non-commercial research [PDF link] regardless of contractual stipulations.
- Be firm with publishers, data providers and other content suppliers in the case of "unusual behaviour" reports or DRM blocks to protect the legitimate interests of researchers, as defined by law.
- Advise on the next port of call should local support not be enough, up to and including appealing to the Intellectual Property Office (IPO) if a rights-holder will not remove DRM locks (though to date take up of this has been very low). Information on making a complaint can be found on the IPO's website.
- Be aware of data protection issues: TDM is not exempted from data protection law, so content identifying a living individual must be processed in compliance with the Data Protection Act (1998).
- Be mindful of clauses in any new licence agreements you are signing for resources that might restrict TDM activities.

Digitisation projects for TDM

An example of a recent TDM project here at LSE is ongoing work to digitise a series of hard copy South African census data volumes. This series reaches back to the mid-19th century and carries on through to 1960. It includes volumes concerning individual provinces before the unification of South Africa in 1910 as well as volumes for the whole country post-1910, and the total run comprises some 25,000 pages across 35 volumes.

Dr Joachim Wehner, a colleague in LSE's Department of Government, was the project's instigator when he identified the series as un-digitised and potentially of great historical importance for the understanding of South Africa. Working in collaboration with Daniel de Kadt, a PhD candidate in political science at MIT, he formulated a research question that he intends to answer by performing TDM on the datasets: what occurred to quality of life indicators as the South African franchise was gradually restricted on the grounds of (needless to say, dubious) racial categories under apartheid, then widened post-apartheid? To answer this question he needed to "unlock" the data in the census volumes by digitising it then running Optical Character Recognition software over the digitised pages. This has created a set of tables as spreadsheets that are now amenable to further analysis, both by traditional methods and using computational and statistical methodologies.

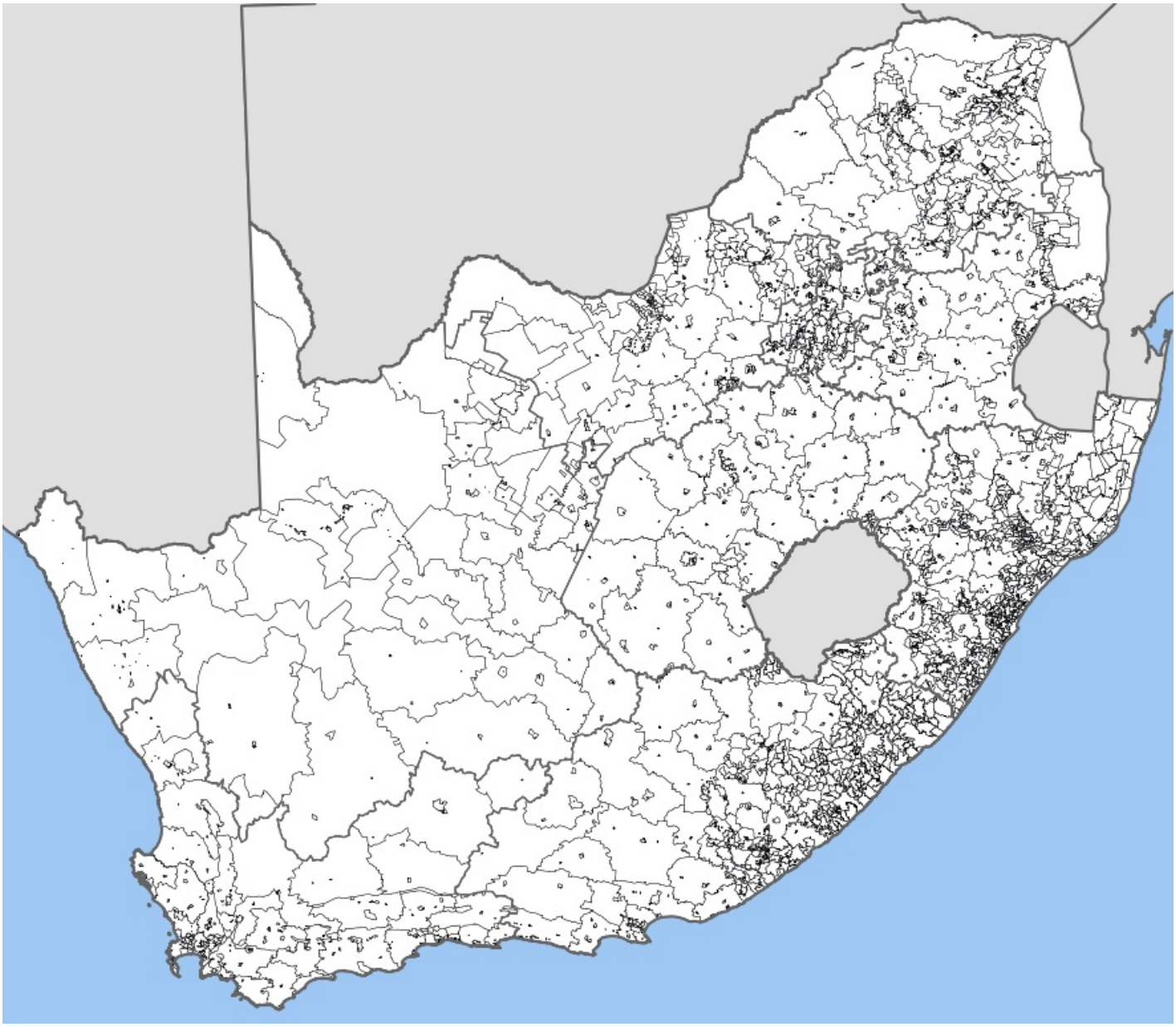

Image credit: Stats SA Map of South Africa divided into "Main Places" from Census 2001. CC BY

The project is ongoing, but the lesson to take away for the purposes of this blogpost is that Dr Wehner's research question could not have been approached without support from LSE Library. This support comprised arranging and helping manage the digitisation of the volumes and its subsequent OCR-ing, and providing a place for (some or all of) the digitised material and data tables to be made openly available: LSE's Digital Library. The (ultimately) open availability of the data is perhaps the most exciting thing about the project, given the plethora of possible questions that might be amenable to answer after the previously locked-away data is made available.

Conclusion

Researchers should look to their library services for assistance with TDM, and libraries should be ready and willing to assist with these projects, particularly to help with rights issues and to digitise material for scholarly re-use. In future, further harmonisation needs to occur across the EU to bring member countries in line with the recent UK legislation, thereby allowing multi-national research partnerships across Europe, something not currently permitted. The Libraries and Archives Copyright Alliance (LACA) and Universities UK have been keen to collect evidence of

problems researchers run into when trying to use the new copyright exception. In a recent example, LACA advised an academic to appeal to the IPO to have DRM removed from a site he wished to mine. Content Mine are also doing a lot of work in this area, to press publishers who restrict researcher's ability to mine through technical protection measures, and leading on TDM projects as well, for example a project to mine the research literature for information about the Zika virus. Finally, for librarians who are used to signing licence agreements, they must be mindful of the right to mine, and recognise the important role they can play in liberating data and knowledge to allow new scientific breakthroughs.

Note: This article gives the views of the author, and not the position of the LSE Impact blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Authors

Neil Stewart is the Digital Library Manager at the Library of the London School of Economics and Political Science (LSE). He manages LSE's Digital Library, an online repository of digitised and born digital materials from LSE Library's rich collection of social science holdings. He is interested in digital scholarship, digitisation for scholarly re-use, research data management, copyright and intellectual property, open access, open science and web technologies for libraries.

Dr Jane Secker is the Copyright and Digital Literacy Advisor at LSE where she coordinates digital literacy programmes for staff and students including copyright training and advice. She is Chair of the CILIP Information Literacy Group, a member of the Libraries and Archives Copyright Alliance and the Universities UK Copyright Working Group, which negotiates licences for the higher education sector. She is widely published and author of four books, including Copyright and E-learning: a guide for practitioners, which was published in 2010 by Facet. The second edition (authored with Chris Morrison) is due for publication in June 2016.

Chris Morrison is the Copyright and Licensing Compliance Officer at the University of Kent, responsible for copyright policy, education and advice, and is also a member of the UUK / Guild HE Copyright Working Group. He was previously Copyright Assurance Manager at the British Library and prior to that worked for copyright collecting society PRS for Music. Chris is the creator of Copyright the Card Game, a game-based approach to understanding copyright.

Laurence Horton is Data Librarian at the London School of Economics and Political Science. He is responsible for Research Data Management support in the School. He can be found on Twitter @laurencedata.

- Copyright 2015 LSE Impact of Social Sciences - Unless otherwise stated, this work is licensed under a Creative Commons Attribution Unported 3.0 License.