UKSG eNews
ISSN 2048-7746
CONNECTING THE KNOWLEDGE COMMUNITY

Data supporting research collaborations

UKSG eNews 369

15 Apr 2016

Neil Smyth, University of Nottingham

The Corpus Statistics Group was launched at the University of Birmingham in February. It is a new collaboration between the University of Birmingham and the University of Nottingham, bringing together researchers and librarians who are interested in investigating how large data sets, including library data, are used in corpus linguistics and related research. The event demonstrated research using library data, and highlighted some of the challenges facing researchers and, as a consequence, librarians: the need for further advocacy; new knowledge and understanding; and new digital literacy skills.

The library data that is the focus here is the data behind licensed databases. In June 2014, the Copyright, Designs and Patents Act 1988 was amended through *The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014*. New permissions were introduced for computational analysis of anything recorded in works where a person has lawful access to the works. The right to read became the right to mine in the United Kingdom, creating new opportunities to make library data – and other data – available for research.

Thinking about an everyday example of library data might help. *The Times* is a newspaper you might read, whether through holding physical pages in your hands or online – if you are lucky enough now to have a membership subscription. Your university library might have access to hundreds of years of *The Times* on microfilm, allowing you choose a reel from the many rows of shelves and read the content on a machine. Some of you may work in libraries were there are several digital choices for reading past content from *The Times*, such as *The Times Digital Archive* or other databases, like *Nexis*. Some libraries also have *The Times* content on a hard drive which is stored in a cupboard as a back-up. Using the exception in the new copyright regulations, librarians can provide access to *The Times Digital Archive* data where they have access to *The Times Digital Archive*, making this data available to be read and mined in new ways for research.

But the *The Times* data is just one news data set licensed from one publisher. There are growing volumes of library data across institutions. Many research libraries, for example, have access to multiple historical newspaper archives from different publishers – and many more databases that could all be mined in research projects. The British Library is recording multiple television and radio channels, including subtitles, creating an ever expanding news corpus of data and metadata. So, what new challenges does the availability of library data offer for librarians?

At the Corpus Statistics Group launch, the key note lecture by Professor Laurence Anthony from Waseda University in Japan was about skills. This question was aimed at academic researchers but it could also be aimed at librarians. Like many researchers, librarians are not familiar with the raw data behind databases or with methods and tools for manipulating that data. Librarians could

develop the programming data skills to work with data and produce new tools to answer specific research questions. This could be a specific way to support research projects. If the librarians do not have the data skills, they may need to work with others who do. These collaborations will be across and beyond organizational boundaries. Librarians in universities and at the British Library, for instance, may need to work together to find new ways to combine news data licensed at the universities with news data at the British Library to create new corpuses for research.

Maintaining knowledge and understanding of changing academic disciplines becomes a challenge. For those librarians with the skills, there are new roles in developing digital literacy programmes for future academics so they have the skills to develop the tools for data. Others will be able to concentrate on the complexity of the digital networked scholarly record, see The OCLC report on the Stewardship of the Evolving Scholarly Record: From the Invisible Hand to Conscious Coordination by Brian Lavoie and Constance Malpas. One example is software. Laurence Anthony is the creator of corpus analysis tools, such as AntConc, AntPConc and AntWordProfiler. But how many librarians know about this software and how it is used with data, and who is introducing graduate students to these information resources?

With new knowledge comes the awareness that researchers do not want to be limited to opportunistic data where the library just happens to have to have access. An academic may have research questions related to data from a database where you do not have licensed access. Librarians may need to provide access to these data sets on their own. This is an emerging advocacy opportunity for librarians, with the potential for publishers to licence and sell databases, databases and data or just the data, depending on identified needs. With a wider range of choices, librarians can develop new ways to support research, perhaps around loaning hard drives of data or making the data available on university networks.

Advocacy is extended to changing legislation and understanding emerging research outputs. The United Kingdom legislation has changed but harmonisation across Europe and expanding the UK legislation to commercial purposes are next steps. But just think about what researchers are doing with library data. Words become numbers and numbers become visualizations. How do you judge the extent to which data is transformed into something new? When new research outputs are created from library data, who owns the derived works created from the data? The researcher, the publisher, the University? If researchers are collaborating across institutions, such as in the Corpus Statistics Group, are librarians managing the data transfer agreements? Perhaps librarians are not even aware of the activity.

Complexity is the challenge, whether it is knowledge and understanding, technical skills or managing relationships for advocacy across and beyond institutional boundaries. The launch of the Corpus Statistics Group is one recent example of evolving inter-institutional collaborations around library data and other data sets. Data will continue to transform academic research and the role of the library is only just beginning to emerge in this changing context.

This UKSG Editorial is taken from the industry newsletter *UKSG eNews,* published every two weeks exclusively for UKSG members. The newsletter provides up-to-the-minute news of current issues and developments within the global knowledge community.

To enjoy *UKSG eNews* and other member benefits become a UKSG member.
To submit an editorial suggestion for *UKSG eNews,*
contact the editors: seneditor@uksg.org.