PUBLISHING

Text-mining spat heats up

Scientists and publishers clash over licences that would let machines read research papers.

BY RICHARD VAN NOORDEN

It is seen as the future of computer-based research — if only the gatekeepers would let scientists in. Researchers have complained bitterly over the past year that publishers do not allow them to use computer programs to download and crawl across the text of research articles, a methodology known as text mining that can reveal large-scale patterns in the studies (see 'Uses of text mining').

Fearful that their content might be freely redistributed, publishers tend to block programs that they find crawling the full text of articles, making no exceptions for users who have paid for access. They give permission only on a case-by-case basis to those who negotiate agreements on access and use. Now, the European Commission (EC) and publishers' consortia are trying to craft clearer rules. But complaints last month to an EC group set up to discuss text and data mining suggest that disagreements are still rife.

"Data- and text-mining techniques … could hold the key to the next medical breakthrough, if only we freed them from their current legal tangle," Neelie Kroes, vice-president of the European Commission, told a Brussels intellectual-property summit last September.

Publishers say that so far, few researchers are asking permission to mine text. Still, Amsterdam-based publisher Elsevier says that computer robots crawling its ScienceDirect site made up 4% of total web traffic on the platform in 2012, almost twice the level in 2011. Whatever the intention of such programs, the figure suggests that machines, not just humans, are increasingly poring over articles.

Raul Rodriguez-Esteban, a computational biologist at drug company Boehringer Ingelheim in Ridgefield, Connecticut, says that he ran 160 text-mining queries in 2012. In one, he searched more than 23,000 articles to pick out hundreds of proteins that could relieve a mouse model of multiple sclerosis. He then sketched a network of other proteins that interacted with them, and found new potential drug targets. Academic researchers covet this capability, but say that it takes months or years to negotiate agreements. It took Max Haeussler at the University of Santa Cruz, California, three years to get the rights to download 3 million articles, from which he extracts DNA data to annotate an online map of the human genome (see *Nature* **483,** 134–135; 2012).

Later this year, the United Kingdom will make text mining for non-commercial purposes exempt from copyright, potentially allowing scientists to mine any content they have paid for. Text miners want the EC to do the same. "The slogan that is doing the rounds among users currently is: 'the right to read is the right to mine'," says John McNaught, deputy director of the National Centre for Text Mining at the University of Manchester, UK.

But the EC's working group to discuss text and data mining, set up this year, has already run into controversy. After a meeting on 4 February, researchers and librarians complained that the group was discussing only how to work with text-mining licences, not how to exempt text mining from copyright. "This will only raise barriers to the adoption of this technology and make computer-based research in many instances impossible," they wrote on 26 February in a letter to Kroes and three other European commissioners, who have not yet responded.

A second meeting on 8 March provided little assurance that exceptions would be seriously considered, says Ross Mounce of the University of Bath, UK, who is using text mining to extract trees of evolutionary relationships from the literature. The EC working group hopes to reach a conclusion by the end of the year.

The situation is better in the United States, where some lawyers think that text mining might be permitted by 'fair use' rights, which allow snippets of text to be freely copied. But no one knows for sure, and many researchers are wary of testing the bounds of this legal grey area.

Some publishers say that unrestricted text mining could strain their servers, and so agreements will always be needed to specify when and how articles may be downloaded. CrossRef, a non-profit collaboration of thousands of scholarly publishers, is developing a system to let researchers agree to standard text-mining terms by clicking a button on a publisher's website. CrossRef's Geoff Bilder hopes that the system will roll out by the end of the year.

The Copyright Clearance Center (CCC) in Danvers, Massachusetts, which works with publishers on rights licensing, is pursuing a more ambitious effort. It would act as an intermediary, collecting publishers' terms and content and storing them on a website for researchers, says the CCC's Roy Kaufman. It is working with six publishers (including Nature Publishing Group) and with drug and chemical firms eager to mine the literature.

Heather Piwowar of the National Evolutionary Synthesis Center in Durham, North Carolina, who studies how researchers use data, says that it is unfair that large firms such as Google are allowed to crawl across content to index it — yet scientists are restricted. "Is this defensible on the grounds that Google knows what it is doing but The Rest Of Us Can Not Be Trusted?," she blogs. "I sure hope not." ■

A RICH RESOURCE

Uses of text mining

Linking genes to research papers. The text2genome project pulls out DNA sequences from around 3 million research papers to produce an online genome map in which each region is linked to relevant articles. go.nature.com/iupijx

Mapping the brain. The NeuroSynth project extracted brain-scan data from almost 4,400 research articles, allowing users to link locations in the human brain with associated research terms and topics. neurosynth.org

Chemistry data. SureChem (owned by the same parent company as *Nature*) harvests and makes freely available data on molecules from some 20 million patents. surechem.com

Drug discovery. Researchers searched free abstracts from more than 20 million articles in the MEDLINE database, and discovered an indirect link between E-cadherin (a cell-adhesion molecule) and Parkinson's disease. go.nature.com/fsv4az

CORRECTION

The News story 'DNA tool kit goes live online' (*Nature* **495,** 150–151; 2013) wrongly located Randy Rettberg at the Massachusetts Institute of Technology. He is at the non-profit organization the iGEM Foundation in Cambridge, Massachusetts. The News story 'Sticky problem snares wonder material' (*Nature* **495,** 152–153; 2013) said that the 2010 Nobel prize was awarded for graphene's discovery, but it was for experiments involving graphene. And the World View 'The unlikely wisdom of Chairman Mao' (*Nature* **495,** 143; 2013) wrongly stated that Mao Zedong was China's premier.