Maximilian Haeussler, PhD
Post-Doctoral Scholar in text mining and genome analysis
CBSE, University of California at Santa Cruz
Santa Cruz, CA

Comment 1
As soon as scientific results are openly available, research groups like the one where I work and IT companies will use the opportunity to digitize, index, analyze, reformat the information in them and make them better accessible, as Google has done it with the Internet. Closed-access publications cannot be mined by companies like Google as all scientific publishers have made it illegal to download articles and analyze them in a new way. Open-access articles (less than 10% of all articles) are accessible and can be easily archived by anyone. The size of the data is ridiculously small these days, it all fits onto a single blue ray disc, which otherwise stores only a single Hollywood blockbuster. The size of the data for analysis is not problem, any library can do that.

Without open-access, any data mining is impossible. My example: Nature Publishing Group, which manage less than 5% of biomedical articles, has today sent me a quote on $85.000, for merely allowing me to run my software on their articles (the access is already being paid in millions of dollars by the UC library system). $85.000 that is just for one single research group. There is no law that allows me to run analysis software on research articles, so I will not do it with Nature articles. There are hundreds of closed-access publishers like this.

I tried previously to mass-download research articles from publisher websites, which is really easy. Blackwell Ltd called my University IT department and threatened them to shut down my internet connection. I have no way of exercising any pressure on Blackwell to search cancer journal articles for mutations. They have not interest in mutations.

Laws to allow data mining or laws to mandate open-access don't cost a lot to taxpayers, as publishers will just be paid by the publishing researchers instead of the libraries of the readers. I am convinced that the creative power of the many people and companies that will analyze open articles will be welcomed in the end even by closed-access publishers, as it will attract more readers and ways to analyse the data in these articles. In addition, libraries will be easily able to archive these articles by just downloading them from the publishers' websites.

Comment 2
I know that Springer, Elsevier and Nature Publications, among the biggest publishers, all outsource typesetting and sometimes printing to India. These are mundane tasks and require no special protection or justify why publishers claim intellectual property.

Publishers have not contributed any significant intellectual input. Taxpayers fund the scientists, scientists produced the research and scientists act as peer-referees for publishers. Publishers only select articles based on the referees. There is no reason to overly protect publishers, as they only do the correction, layout and printing of articles.

Comment 3
I don't think that the government should get involved with building search engines or develop analytical tools. For articles that are available as open-access data, companies, libraries and Universities will start building search engines. With PubmedCentral, this has already happened. UKPMC has taken their content and added new searches. It will take more time to see what other innovative ideas emerge based on open-access articles.

For older articles that are not published as open content, it would be great if individual researchers were allowed to analyze them and aggregate statistics. A federal agency like the NIH or the national library could keep a version (PDF) of all of these articles on a computer system (it only requires a standard computer and standard harddisks these days) and researchers could apply for access. This would make it possible for me to find all articles that relate to a given cancer mutation. This type of indexing service is currently impossible, as closed-access publishers do not provide access to the fulltext of the articles or charge a prohibitive price, to discourage any indexing. In my particular case the price was $85.000 for less than 5% of all biomedical articles, for just one research group. These prices make it impossible to ever come up with new ways of analysing or indexing research articles.

Comment 4
I don't know of any innovative search or analysis functions of existing closed-access publisher archives. Elsevier has started sending their full content to interested researchers like me, but we are still at their mercy and they can cancel the project at any time. No other publisher that I know of is giving access to closed-access

content.

Comment 5
I do not believe in the value of expensive mandatory core metadata as research changes so quickly. Each discipline has very different requirements, cancer journals need cell line data, developmental biologists are more interested in species. Google Websearch works very well without any core metadata. Once access to articles is available in an open-access form, the different disciplines can think about the metadata that they need and can often extract them from the fulltext of the articles with software. CiteSeerX and Arxiv is are good examples of how this worked in computer science and in physics.

Comment 6
More common open-access would lower library costs extremely. But publishing has a price and publishing costs should be an accepted part of federal research budgets. My impression is that it is likely that the initial cost of publishing for a researcher will be higher than what some open-access publishers (e.g. PlOS) are charging at the moment.

Comment 7
Conference proceedings replace journal articles in many disciplines (e.g. computer science, engineering), so yes, conference proceedings should be covered.

Many book chapters, not all, include a significant contribution of the publishers. Layout, color figures are designed by professional artists. These book chapters are used by students and do not serve the same purpose as scientific articles, they make information more accessible but do not report the original research. They should not be covered by open-access provisions.

Comment 8
I do not see a justification why taxpayer-funded research would be published as closed-access to subscribing University libraries and only be available to taxpayers after a 1 year period.

If I can help with anything else or more details on how my text mining of closed-access publications is impossible, please don't hesitate to contact me.