



Poster

2012

Open Access

This version of the publication is provided by the author(s) and made available in accordance with the copyright holder(s).

---

## Text Mining and copy right laws : A case for change in the medical research field

---

Blanc, Xavier; Collet, Tinh-Hai; Iriarte, Pablo; de Kaenel, Isabelle; Krause, Jan Brice

### How to cite

BLANC, Xavier et al. Text Mining and copy right laws : A case for change in the medical research field. In: EAHIL 2012 Conference - Health information without frontiers. Bruxelles. 2012.

This publication URL: <https://archive-ouverte.unige.ch/unige:23172>

# Text Mining and copy right laws

## A case for change in the medical research field

Medical Library and Department of Ambulatory Care and Community Medicine, University Hospital Center and University of Lausanne, Switzerland

Xavier Blanc, Tinh-Hai Collet, Pablo Iriarte, Isabelle de Kaenel, Jan Krause

xavier.blanc2@chuv.ch, tinh.-hai.collet@chuv.ch, pablo.iriarte@chuv.ch, isabelle.de-kaenel@chuv.ch, jan.krause@chuv.ch

### Background

Mid 2011, the research team at the Department of Ambulatory Care and Community Medicine, University of Lausanne, decided to start a project on a new research topic: Shared Decision Making (SDM). The objective was to identify publication trends about SDM in 15 major internal medicine journals over the last 15 years. It was decided to use a "text mining" approach to systematically review all the articles published in these main journals and automatically search for the different occurrences of SDM. The research team turned to the medical library for help to collect the electronic publication files.

### Methods

The software applications used in text mining allow to search through large sets of unstructured texts. The results are then clustered to extract trends, facts and build new knowledge. In order to work consistently, all the text sources should be aggregated on one single local platform. However electronic scientific publications are currently stored as licensed materials on publishers' sites. Bulk download of thousands of articles are not commonly permitted by licenses. The library teamed up with researchers in order to get all the permissions to compile the files for research purpose.



### Request for permission sent to publishers of the major internal medicine journals



	Number of articles in 1996-2011	Publisher' response	Collecting full text strategy	Text mining allowed
Massachusetts Medical Society	19,778	No response	Manual PDF download after full-text search from web site	✗
Lancet Publishing Group [Elsevier]	45,325	No response	Manual PDF download after full-text search from web site	✗
American Medical Association (2 titles)	41,256	Refusal : suggested advanced search on web site	Manual PDF download after full-text search from web site	✗
American College of Physicians (2 titles)	13,206	Initial positive answer, reminder sent but no response	Manual PDF download after full-text search from web site	✗
British Medical Association	61,903	OK and protocol signed	Complete full-text in XML format provided by FTP	✓
Canadian Medical Association	18,686	Positive answer but no FT provided yet	Manual PDF download from PMC	✓
Blackwell Scientific Publications	2,544	Initial positive answer, reminder sent without response	Manual PDF download after full-text search from web site	✗
American Association of Professors of Medicine	6,780	Initial positive answer, reminder sent without response	Manual PDF download after full-text search from web site	✗
Mayo Clinic [Elsevier]	4,810	OK to download from website	Complete full-text in PDF format download from website	✓
American College of Preventive Medicine	3,442	OK to download from website	Complete full-text in HTML format download from website	✓
American Society of General Internal Medicine	4,267	OK and protocol signed	Complete full-text in PDF format provided by FTP	✓
Elsevier	3,368	No response	Manual PDF download after full-text search from web site	✗
American Academy of Hospice and Palliative Medicine	3,814	No response	Manual PDF download after full-text search from web site	✗



### Results

Contacts with publishers and exchange of information over the research project were particularly cumbersome and time consuming. After 6 months, only 5 out of 15 publishers had agreed to grant an license extension that gave the right to systematically download the articles for research purpose. Permission was usually granted under one main condition : all downloaded content for text based analysis should be destroyed when research is complete.

### Conclusion

Due to the rapidly expanding body of electronic biomedical literature, text mining should become an essential process for research in the medical field. To allow this new research method to expand, copyright law and licenses for electronic access have to be amended and new competences have to emerge in libraries and research centers.



### Bibliography

- Ramakrishnan C, Patnia A, Hovy E, Burns GA. Layout-Aware Text Extraction from Full-text PDF of Scientific Articles. Source Code Biol Med. 2012 May 28;7(1):7
- Agarwal S, Yu H, Kohane I. BioNØT: a searchable database of biomedical negated sentences. BMC Bioinformatics. 2011 Oct 27;12:420. PMID: 22032181
- JISC (2012) The Value and Benefit of Text Mining to UK Further and Higher Education. Digital Infrastructure. Available at: <http://bit.ly/jisc-textm>
- Kostoff RN. Expanded information retrieval using full text searching. Journal of Information Science.2010; 36:1. 104-113. 2010
- Van Noorden R. Trouble at the text mine : Computers can rapidly scan through thousands of research papers to make useful connections, but work is being slowed by publishers' unease. Nature. 2012 Mar 7;483(7388):134-5. doi: 10.1038/483134a. PMID: 22398532
- It just got easier to find open access research for textmining/redistribution. BioMed Central Blog, Wednesday Mar 07, 2012. Available at: [http://blogs.openaccesscentral.com/blogs/bmcblog/entry/it\\_just\\_got\\_easier\\_to](http://blogs.openaccesscentral.com/blogs/bmcblog/entry/it_just_got_easier_to)

### Links

- <http://text.soe.ucsc.edu/>
- <http://cameronneylon.net/blog/they-just-dont-get-it/>
- <http://caseybergman.wordpress.com/2011/06/>