Text Data Mining Beyond the Open Data Paradigm:
Perspectives at the Intersection of Intellectual Property and Ethics

Megan Senseney
School of Information Sciences
University of Illinois at Urbana-Champaign
Eleanor Dickson Koehl
University Library
University of Illinois at Urbana-Champaign

Introduction

Copyright law and resource licensing complicate the application of **text data mining** for research (Brook, Murray-Rust, & Oppenheim, 2014). Scholars often use – or wish to use – web-based content, news media, scholarly journal articles, or large collections of digitized books. To work with these data, scholars must:

- interpret the **terms of use** for publicly available content,
- negotiate with content providers for access through **formal licensing**, and
- operate within an ambiguous **fair use** framework for materials that are in copyright.

The existing legal and socio-technical landscape gives rise to **ethically complicated situations**:

- researchers want to use text data but lack clarity on which uses are permissible;
- authors want to mine journal content but may not engage in publishing practices that make their own work mineable;
- universities want to benefit from the use of altmetrics but, in doing so, risk compromising and commodifying scholarly production.

Method

Convened a National Forum on Text Data Mining with Use-Limited Data that brought together 25 leading stakeholders selected from among researchers, librarians, content providers, legal experts and representatives of scholarly societies.

Conducted conventional qualitative content analysis on materials gathered before and during the Forum in Atlas.ti using a set of 26 thematic codes divided into six categories (Hsieh & Shannon, 2005).

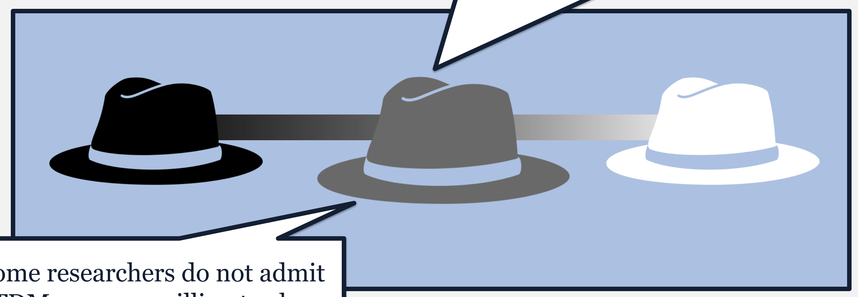
- Semi-structured interviews
- Participants' forum statements
- Participants' SWOT analyses
- Collaborative note-taking during Forum

Findings

Individual Level

Researchers who wish to utilize text data mining methods experience a chilling effect on their scholarship when faced with legal and ethical ambiguity.

"This legal ambiguity causes a great deal of uncertainty and disincentive in my work and makes it harder to collaborate. It also violates basic norms of research which are predicated on the transparency and reproducibility of scientific research."



"Some researchers do not admit to TDM or are unwilling to share their projects because of fear of being sued, leading to difficulty in reproducibility."

Fig 1. *One participant proposed a black hat, gray hat, and white hat model for understanding users' assumptions about access and use where "gray hat" practice operates "in a space where legal and ethical compliance is uncertain."*

Institutional Level

At the level of local policy, university administrators must re-examine the ways they license content and how they implement data governance policies in light of the text data mining practices of scholars and the vendors who wish to profit from scholarly production.

"A related concern is the ability of publishers to surveil uses of scholarly materials. [....] Only gradually are scholarly authors coming to realize that if you are not at the table, you are on the menu."

Data Governance

"Our libraries and larger institutions have antiquated and limiting policies regarding privacy and IRB matters"

Intellectual Privacy

Library Licensing

"It is poor stewardship of collections budgets to spend large sums licensing data that is restricted in ways that might hinder research. I believe that the research library community should leverage its collective consumer power by walking away from overly restrictive text- and data-mining licenses."

Conclusion

The current climate hampers research activity and undermines scholarly communication. While research is ongoing, a selection of preliminary recommendations for information professionals in higher education are presented below for discussion and debate.

1. Convene or participate in a campus level task force to address data governance and risk management
2. Centralize and share licensing agreements in a secure repository accessible by the entire campus community
3. Collaborate with professional organizations to commission best practices guide for fair use in TDM
4. Encourage data sharing practices that combine derived data and methods papers for TDM with use-limited data
5. Build infrastructure for facilitating peer review and reproducibility in secure environments

References

Brook, M., Murray-Rust, P., & Oppenheim, C. (2014). The social, political and legal aspects of Text and Data Mining (TDM). D-Lib Magazine, 20(11/12).

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative health research, 15(9), 1277-1288.

Hat icon by chiccabubble from the Noun Project

Acknowledgements

The authors gratefully acknowledge the Institute of Museum of Library and Information Services for their generous project funding (LG-73-17-0070-17), the Center for Informatics Research in Science and Scholarship, the iSchool and the University Library, Beth Sandore Namachchivaya, Bertram Ludäscher, and the National Forum participants.

ILLINOIS