Opportunities and Challenges to Utilizing Text-data Mining in Public Libraries: a Need for Legal Research

Liane Colonna

1 Introduction

Libraries have been an invaluable part of human history, helping to support equal access to education and propagating culture over the centuries. However, the digital age has transformed information access in ways that few ever imagined and now many are predicting that libraries are becoming obsolete. After all, books are cheaper than they have ever been in real terms: most classic books are available for free from Project Gutenberg or for a small charge on Kindle and second-hand books can be purchased from across the world via Alibris. A massive amount of information is also available online and easily searchable with Internet search engines like Google, which means that the Internet is replacing the library as a repository for knowledge.

Research is necessary in order to consider the ways in which librarians, scientists, technologists, scholars, policymakers and ordinary citizens can work together to leverage massive computing power, digitization, and "big data" to reimagine libraries as centers of information and innovation. Specifically, there is a need to find legally compliant ways for libraries to open their collections to computation, allowing individuals to apply text data mining techniques to create new ways of interacting with the vast archives available within them.[1]

2 The Need for Legal Research

Individuals have sound and valid reasons for relying on the Internet for their information needs. Internet search engines provide information that is self-service, free, and available around the clock in one's own home.[2] Anderson states, "Google has succeeded wildly at finding its users the information they want in return for a minimum investment of time and energy."[3] Likewise, Timpson explains that for searchers Google offers a one-stop shopping experience and a highly usable interface.[4]

Unlike Google, however, libraries like the National Library of Sweden ("Kungliga biblioteket") have access to older, analog material that has been digitalized. As such, the library can offer much higher quality information to the researcher in terms of the authority and the credibility of the resources: there exists not only an imprimatur of excellence in library resources that simply does not exist with Google sources but also a potentially more interesting and relevant set of resources than that which is available online.

---

1 An example of a legally non-compliant use of advanced data processing techniques is the Cambridge Analytica scandal. For more, *see* Mark Bridge, Carole Cadwalladr and Emma Graham-Harrison, *Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach*, The Guardian (17 Mar 2018).

2 R. Anderson, *The (uncertain) future of libraries in a Google world: Sounding an alarm*, 10 (3/4) Internet Reference Services Quarterly 29 (2005).

3 Id.

4 H. Timpson and G. Sansom, *A student perspective on e-resource discovery: Has the Google factor changed publisher platform searching forever?* 61 Serials Librarian (2011).

Unfortunately, from the perspective of the researcher, it is difficult to search through library archives, especially when compared to Google. Google's search is more granular because it can search at the article level. Libraries' search engines are not as sophisticated: generally an individual can only search as deeply as the title of a book. Google also has full-text search capability. Essentially, as Anderson observes, Google can search the content and the library catalog can only deliver the container.[5]

Text-based data mining (TDM) represents the solution for the dilemmas facing libraries in the twenty-first century. TDM generally refers to "the process of extracting interesting and non-trivial patterns or knowledge from text documents." [6] TDM has enormous potential both for industry and non-commercial research to the extent that it allows researchers to analyze huge amounts of data in order to find answers to pressing challenges. More specifically, TDM can be used to discover correlations between materials produced in different scientific fields and to generate new knowledge.

Essentially, TDM is the added value for libraries of the future because the technology promises to help libraries leverage the massive amounts of data that they have in their repositories. TDM within the library context has a number of different current applications and, naturally, the possibilities are expanding with the development of the technology. For example, Yale University is applying TDM to analyze Vogue magazine articles (2,700 covers, 400,000 pages, and 6 TB of data) to provide deep insights into the evolving use of imagery and language in fashion throughout the 20th century.[7] Another example of TDM can be applied is to sift through a large body of newspapers from the Victorian era in order to extract jokes and better understand various aspects of Victorian culture and social history.[8] These projects demonstrate the possible benefits for researchers and society alike when library archives are opened up to data mining algorithms to catalog, mine, visualize and create new ways of interacting with these vast data sets.

Although TDM represents exciting opportunities for researches and libraries alike, data protection law has the potential to hamper the application of this technology. Pursuant to data protection law, processing of personal data must always be lawful, meaning that a data controller meets its legal obligations regarding the processing of personal data. In the EU context, this means, for example, that the data processing adheres to the conditions set forth in the Data Protection Directive and/or the forthcoming General Data Protection Regulation

---

5 R. Anderson, *The (uncertain) future of libraries in a Google world: Sounding an alarm*, 10 (3/4) Internet Reference Services Quarterly 29 (2005).

6 Gaikward, Sonali Vijay et. al., *Text Mining Methods and Techniques*, 85 International Journal of Computer Applications 42 (2014).

7 *See Robots Reading Vogue Project at Yale University*, available at "web.library.yale.edu/dhlab/vogue".

8 Neil Stewart et. al., *Liberating Data: How libraries and librarians can help researchers with text and data mining*, LSE Impact Blog (12 July 2016) available at "blogs.lse.ac.uk/impactofsocialsciences/2016/07/12/how-libraries-and-librarians-can-help-with-text-and-data-mining/".

("GDPR"), especially the articles setting forth the legal grounds for processing, as well as Articles 7 and 8 of the Charter of Fundamental Rights of the European Union and Article 8 of the European Convention of Human Rights.[9] The data processing must also comply with all of the other general principles of data protection law.

3 The General Data Protection Regulation (GDPR)

In April 2016, the EU passed the GDPR and it will be applicable on May 2018. A central goal of the GDPR is to promote a stronger and more coherent data protection framework at the EU-level. Intended as a wide-ranging and far-sighted reform to strengthen and harmonize data protection in the digital age, the regulation updates most of the existing rules and introduces new ones. The GDPR may to some extent be supplemented by national legislation and, currently, a government inquiry is examining what kind of additional national legislation is possible and required in Sweden, particularly within the library context.

At the very outset of the regulation, the GDPR states that "(t)echnology has transformed both the economy and social life, and should further facilitate the free flow of personal data within the Union and the transfer to third countries and international organi(z)ation."[10] At the same time, however, in Article 45 the GDPR sets forth the so-called "adequacy" requirement. This rule requires that Member States may only transfer personal data to a third country (i.e. a non-EU Member State) where the third country ensures an "adequate" level of protection.[11] The objective of the adequacy requirement is to prevent that the high level of data protection that is provided for within the EU from being undermined when data flows extend beyond the EU's territorial borders.

Research is conducted internationally and depends on the global network. The ability to conduct meaningful research utilizing TDM requires not only being able to access data remotely but also being able to share and further process results within the Digital Single Market and beyond. What kind of implications does the GDPR and the national legislation that is now being investigated have on researchers and libraries in Sweden and their interaction with other Member States as well as researchers and libraries in third countries?

It is further important to note that the GDPR eliminates the possibility for Member States like Sweden to apply a misuse model of data protection: an approach that seeks to enhance the efficacy of data-protection rules by simplifying and focusing them on preventing the misuse of personal data under

9 Data Protection Directive, Article 7; European Union, Charter of Fundamental Rights of the European Union, 26 October 2012, 2012/C 326/02, Articles 7 and 8; Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, 4 November 1950, ETS 5 , Article 8.

10 GDPR, Recital 6

11 GDPR, Article 45

certain conditions.[12] This means that if a data controller has used the exemption for the treatment of unstructured material, it is important that it now examine, among other things, whether it has a legal basis for processing and whether it has properly informed the data subject about the processing.

4 Suggested Research Questions

Assessing the tension between TDM and data protection leads to a number of different research questions: How can both researchers and libraries utilize TDM in order for libraries to improve information services and offer new insights into their collections while at the same time comply with data protection laws? To what extent do libraries process "personal data" within the meaning of data protection law in the first place? Which data protection principles or rules cause the main obstacles to TDM? How can TDM be made compliant with data protection law? What kind of data handling processes constitute "best practices" within the research / library context? Do some of these rules cause excessive obstacles for TDM, which can be avoided by using other legal protection mechanisms? How will the global flow of research data, e.g. using TDM be realized? Does the GDPR enable this, or is supplementary national legislation a solution?

5 Relationship to the Existing Literature

The current literature is focused on the applicability of copyright law to TDM. More specifically, the literature concerns whether exceptions to copyright laws for TDM are necessary to facilitate the use of the technology and the scope of such exceptions (e.g. whether they should apply only to non-commercial research institutions or commercial institutions alike). Without a specific exception to copyright law, researchers face copyright infringement-related liabilities whenever they engage in TDM of a copyrighted work because TDM is impossible without at least the temporary reproduction of the work that is analyzed.[13]

Here, it must be noted that text and data mining for research purposes is generally permitted in the US under the "fair use" exception, in Canada under the "fair dealing" exemption, and in Japan under a special statutory exception introduced in 2009.[14] Likewise, the UK copyright law allows researchers to

12 Pursuant to Sweden's misuse model, unstructured processing of personal data is allowed unless it constitutes a misuse of the privacy of an individual. *See* Personuppgiftslag (1998:204), section 5(a).

13 Arul George and Scaria Rishika Rangarajan, *Fine-tuning the intellectual property approaches to fostering open science: some insights from India*, 8 W.I.P.O.J. 109 (2016).

14 Ian Hargreaves, Lucie Guibault, Christian Handke et al., *Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining, Report from the Expert Group for the European Commission* (2014), 44–48.

make copies of any work protected under it for the purpose of TDM, as long as they have lawful access to the work and the purpose of the TDM is for non-commercial research.[15] Furthermore, under new proposals outlined by the European Commission, a new mandatory TDM exception may be implemented into EU legislation although the scope of such an exception remains unclear.[16]

While the applicability of copyright law to TDM is being considered, there is very little research being done concerning the applicability of data protection law to this new technology. Here, there is great legal uncertainty concerning the scope and applicability of data protection law to TDM, which may cause researchers to refrain from the use of this exciting new technology. This project seeks to make a contribution to the literature and to help clarify whether and to what extent data protection laws affect the use of TDM. It also seeks to analyze whether data protection laws require a similar exception to the rules as exists within the domain of copyright laws.

6 Conclusion

Today, the words that exist in books can be more than simply read: they can be sliced, diced and mined in ways never before imagined in order to extract limitless amounts of new knowledge through the application of advanced technologies like TDM. The problem, from the perspective of libraries and society alike, is that TDM is not exempted from data protection laws, so content identifying a living individual must be processed in compliance with the burdensome requirements set forth in laws like the GDPR. As such, there is a genuine need to consider how libraries can make themselves relevant in the computer age through the use of TDM while at the same time meeting the strict demands of European data protection law.

15 Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014 (UK); Copyright and Rights in Performances (Disability) Regulations 2014 (UK); Copyright (Public Administration) Regulations 2014 (UK).

16 *See* Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market - COM(2016)593 available at https://ec.europa.eu/digital-single-market/en/news/proposal-directive-european-parliament-and-council-copyright-digital-single-market