Developing text and data mining (TDM) support within a university research library

The introduction of the text and data mining (TDM) exception in 2014 led to researchers asking for support from staff within Library Services at the University of Birmingham. An initial involvement with a funded corpus linguistics project fostered an effective partnership between the Copyright and Licensing Team and the University's Research Infrastructure Team. This case study traces the TDM journey that Library Services has subsequently undertaken. The article will look at how staff in Copyright and Licensing and the Research Skills Team identified the original service gap. It will also look at issues impacting on supporting TDM and the results of a TDM survey that was sent to researchers. It concludes with a reflection on how the service might evolve in the future – from the creation and availability of TDM datasets, to the skills development of both librarians and the university communities they support, and the impact artificial intelligence (AI) developments might have on TDM practices.

Keywords

text and data mining; TDM; research support; research skills; copyright and licensing; service development

LISA BIRD

Copyright and Licensing Advisor
Library Services
University of Birmingham

JAMES BARNETT

Research Skills Advisor
Library Services
University of Birmingham

Introduction

'The Copyright, Designs and Patents Act 1988 was amended in 2014 to include Section 29A: Copies for text and data analysis for non-commercial research'

The Copyright, Designs and Patents Act 1988 was amended in 2014 to include Section 29A: Copies for text and data analysis for non-commercial research.[1] This new section allows researchers to carry out computational analysis of works they have lawful access to, where the analysis is solely for non-commercial research purposes. This article will discuss how Library Services at the University of Birmingham developed their services to meet the needs of researchers who wanted to take advantage of this new change in the law. The article will look at Library Services' initial involvement with a text and data mining (TDM) project, how TDM service gaps were identified and developed, and a reflection on how the service might develop in the future.

Background

A year after the change in legislation, Library Services received its first TDM enquiry from an academic who was joining the University with a funded corpus linguistics project analysing the use of language in literature. As part of the project, they needed to be able to access a newspaper archive, understand the legal and copyright issues and develop a solution for accessing and hosting the materials. The Head of Copyright and Licensing, based within Library Services, led the support. Issues that needed addressing were:

- legal access to the content
- IT infrastructure for hosting the content
- data security and preservation
- how to share the data internally and externally with research collaborators
- if the materials would be needed to be accessed by others for future research projects.

A hard drive containing the newspaper archive was purchased. The publisher understood the plan for using the content, and the licence was reviewed to see how it aligned with the new legislative exceptions. This has since become part of the normal workflow when purchasing new resources as terms relating to TDM are challenged if they try and restrain the rights of the researchers. We prefer that the licences we sign up to reflect the law rather than having to rely on the 'no contract' override provision.[2] Library Services worked with the University's Research Infrastructure Team[3] and the material was loaded onto the research infrastructure to give the academic access to the materials. Further researchers could be provided with access upon request. Library Services has since worked with the Research Infrastructure Team to create an online form so that other researchers at the University can request their own copy of the archive on the research infrastructure. Having the data in this environment resolved many of the initial issues mentioned above. Access permission is managed by Library Services' Copyright and Licensing Team.

'when purchasing new resources … terms relating to TDM are challenged if they try and restrain the rights of the researchers'

This means that we now have at least two versions of that archive on the infrastructure:

1) The original archive as it was uploaded onto the research infrastructure. This is a version that other researchers at the University can request a copy to use for their own research. This is the 'golden copy'.[4]

2) The version that the corpus linguistics research team processed through optical character recognition (OCR) and corrected text file (.txt). This would have taken many hours of work to complete manually.

Identifying the service gap

Once the work associated with the corpus linguistics project was under way, we began thinking more broadly about the role Library Services might have in supporting the TDM activity of the wider research community. Through responding to ad hoc TDM queries that have subsequently arisen, we have been able to identify gaps that need to be filled (either by the library, or other university services). The queries themselves were usually initiated by researchers desiring clarity around the copyright and licensing implications of being able to mine the content of some of the databases and archives we subscribe to. The nature of the content was multifarious and included newspaper archives, legal cases and citation data. However, through handling these queries it became clear that there were common issues repeatedly emerging that then became barriers to the researcher(s) being able to undertake TDM work. These issues have allowed us to build a picture of what service gaps would need to be filled for a scalable TDM service to be established. The gaps are discussed below.

Access to data

There have been queries around TDM support where the researcher approaches Library Services wanting to know what minable data is available to them via our current subscriptions. For example, a doctoral student approached Library Services for support around a project requiring access to newspaper archives, but first needed clarity on which newspaper archives (and the platforms they were hosted on) we had available in our collections. Once clarified, a conversation then ensued about how accessible those resources were for undertaking TDM activity. For example, which resources had a free application programming interface (API) available? Did the platforms have download limits to be considered? This is a clear area for the library to be a leader in given that promoting and embedding resources is a key tenet of any library service, but the gap here is in making it clear not just where the data is held, but how accessible that data is for TDM activity.

'making it clear not just where the data is held, but how accessible that data is for TDM activity'

Access to APIs and funding

While Library Services, via the Copyright and Licensing Team, is well-positioned to offer researchers clarity about what can and cannot be done with respect to the text and data mining of the resources it subscribes to, for researchers undertaking TDM activity it is often not what can be done that proves the biggest barrier, but how to do it. For example, while a researcher wanting to mine a database or archive might initially contact the library for clarification that our licence permits them to perform TDM work, they very often follow this enquiry with questions about whether the library is able to provide access to APIs that publishers or content providers set up for this work. Many suppliers require an additional fee for access to their APIs, and these are often prohibitively expensive (sometimes costing more than an annual subscription to the database itself). A researcher with grant money where the cost of obtaining access to a relevant API is built into the project costs may be able to afford the expense, but how do we offer an equitable service for those who are unfunded? For the library, investing in expensive APIs becomes a difficult strategic choice as the financial outlay must be weighed up alongside requests for other resources.

'Many suppliers require an additional fee for access to their APIs, and these are often prohibitively expensive'

Access to skills and support infrastructure

Some ad hoc queries stemmed from a need for support to develop the skills to successfully perform the necessary TDM activity. For example, where an API is available, support might then be needed to understand how to query the API to leverage the required data. Further, where large datasets are accessible for TDM purposes, but an API is not, some researchers then need support in developing coding skills to write their own API. These are not skills that we are able to provide support and training for in Library Services, so we rely on signposting support from colleagues in Research Infrastructure who run training in coding languages such as R and Python.

TDM project

Having identified the gaps, Library Services started a small project to look at:

- the other datasets that we had on hard drives and whether we could add those to the research infrastructure for preservation and future TDM use
- reviewing off-the-shelf TDM tools
- creating TDM web guidance
- support needed by researchers.

Datasets on hard drives

Library Services identified several hard drives containing archival collections which had already been purchased. These have been added to the research infrastructure. We are in the process of producing the online request forms and mechanism so that researchers can access these 'golden copies' for TDM based on our experience with the newspaper archive.

TDM tools

Library Services also looked at acquiring access to TDM tools, but there is currently no funding pot to resource these. A request was made by a corpus linguistics academic who required a tool to enable their students to easily work on small TDM projects. Library Services opted for the Gale Digital Scholar Lab[5] which would allow data mining across many archives that Library Services had already purchased.

TDM web pages

The Copyright and Licensing Team created the original draft of TDM web pages to guide researchers wanting to participate in this type of research. The Research Skills Team and Research Infrastructure Team reviewed the pages and provided feedback on how information about their services could be integrated into the pages. This allowed the pages to reflect the wider central support available for TDM across the university from one location. The pages are now hosted on the Library Services website[6] and provide information on:

1) What TDM is and the copyright issues associated with this work.
2) The resources that we already have available on the infrastructure for TDM. Plus, details of our access to the Gale Digital Scholar Lab and the archives that we have available for mining within that resource.
3) TDM and software tools including paid for and free and open source software (FOSS)
4) Guidance on how to get coding support at the University.

TDM survey

Library Services surveyed staff and doctoral students engaged in TDM to understand more about researchers' current TDM use and issues that they encountered.

Researchers were identified either because Library Services had already had contact with them due to TDM related queries or because they were listed on the University web pages as belonging to research groups with a TDM interest, e.g. the Centre for Corpus Research[7] and The Institute for Interdisciplinary Data Science and AI.[8] 150 individuals were identified and received a personalized e-mail asking them to complete the survey. The survey was short and had the purpose of finding out:

- if these individuals were already participating in TDM activity
- what tools they were using to complete that work
- if they were encountering any barriers undertaking (or planning to undertake) TDM research
- how useful they found Library Services' new TDM web pages.

The survey had a 15% response rate, and the responses were from subjects across the University. 78% of the respondents said that they were involved in TDM projects. Those that were not involved stated that they wanted to undertake research involving TDM but needed training.

Respondents used a variety of tools and skills to conduct their TDM research, including coding skills, software and hardware. They identified 35 different tools, but most were only used by one or two people.

R and RStudio was used by 39% of respondents with Python (mentioned by 30% of respondents) following closely behind. Some respondents were making use of both of R and Python, using Python to preprocess the data and R to complete the analysis. It was useful to understand which coding language researchers are using for these projects so that Library Services know where to focus their efforts if staff need to become familiar with these languages in order to better support TDM researchers.

It was clear that off-the-shelf tools were not always meeting researcher needs, as 13% of respondents had created in-house tools. This reveals that a certain level of technical and coding skill can be required by researchers when undertaking TDM projects.

'It was clear that off-the-shelf tools were not always meeting researcher needs'

The two largest barriers to undertaking TDM research (with both being mentioned by 17% of respondents) were skills/technical knowledge and data availability. When looking at the comments relating to skills/technical knowledge, it was clear that researchers struggled due to not having sufficient coding skills and were also unsure of the best methods for cleaning any data that they had managed to pull together. Due to the large number of tools available online they also struggled to identify which ones would be most useful for their needs and wanted more directional support. The main barriers concerning data were, unsurprisingly, related to requiring funding to support API access costs. This was of particular concern for those who had unfunded research. However, some respondents were just unclear which datasets the University had available for TDM research, highlighting that better promotion of the TDM web pages was required.

Respondents felt that the new TDM web pages were a good starting point for anyone new to TDM but could be improved by including case studies from current TDM researchers to show the possibilities of TDM as a research method. They also requested that more skills guidance be added to the pages.

Reflections and the future

The TDM Project – particularly the TDM survey – has highlighted that while Library Services offers much to its users in terms of TDM support, there is scope to consolidate the offer alongside that of other University services and develop it further. Some of these future areas of development are considered below.

'there is scope to consolidate the offer alongside that of other University services and develop it further'

OCR/machine-readable versions of data

The corpus linguistics project created two versions of the newspaper archive: a 'golden copy' and a TXT version that has been fully processed through OCR and corrected. A text file is potentially more useful for research because it is ready for TDM. However, researchers may need to edit either version to meet their specific needs. If non-golden copy versions are made available, they should be accompanied by information about how they were edited. Ideally, publishers would provide OCR-corrected versions of their archives, which would save time across the sector.

TDM education and skills development

The TDM survey showed that 22% of the respondents wanted to be involved in TDM research but would need training, which identifies a clear gap. Realistically, upskilling library staff to have advanced coding skills would be a significant long-term endeavour. However, equipping the staff likely to interface with TDM queries to have knowledge of the basics of coding would be a useful strategic investment. Those staff would then be ideally placed to thoroughly understand the nature of TDM queries and improve the experience of researchers, by seamlessly signposting queries between library staff and coding support available beyond the library.

There is potential for the Research Skills Team to use a tool like the Gale Digital Scholar Lab – which allows TDM on included resources to be carried out without the need for researchers to have access to advanced coding skills – as the basis for developing training material that introduces TDM to researchers unfamiliar with the concept and uncertain of the relevance of TDM to their own research. The hope here is that Library Services can play an active role in broadening the awareness and appeal of TDM techniques to the wider research community.

Further development and promotion of TDM web pages

The TDM survey recognized that the new web pages provided anyone new to TDM with useful information for getting started but could be improved by adding case studies, skills guidance and clarity around available datasets. Adding case studies is a definite development for Library Services to work towards. However, the TDM web pages already contain information about skills development opportunities and available datasets and tools. This suggests that the work required is to ensure that the information being relayed on the web pages is as clear and accessible as possible, and then amplifying this information to the research community via targeted promotion.

Developments in artificial intelligence

'the potential exists for GAI technology to be used to query large sets of data using natural-language prompting to carry out TDM tasks'

It is also necessary for Library Services to be aware of the impact on TDM activity that the explosion of developments in artificial intelligence (AI), particularly generative AI (GAI), might have following the release of ChatGPT in November 2022.[9] Should use of GAI tools become more commonplace in research processes,[10] the potential exists for GAI technology to be used to query large sets of data using natural-language prompting to carry out TDM tasks that previously would have required access to APIs or advanced coding skills. This could open up TDM to a wider set of researchers than before. However, this possibility is also fraught with risks – from issues of reproducibility, understanding the corpus against which the tool has been built ('the black box'), to the potential for copyright infringement[11] – so Library Services will need to be ready to educate the research community on the opportunities and risks that come with using AI both for TDM activity and as part of research more broadly.

Conclusion

'TDM support will not necessarily be limited to a library's research community'

As a research method, TDM represents a real growth area. Advances in big data and machine learning, alongside the rise of disciplines such as the digital humanities, mean that researchers are increasingly interested in how TDM techniques are applicable to their own research areas.[12] However, it should be noted that TDM support will not necessarily be limited to a library's research community. At the University of Birmingham we are beginning to field TDM enquiries from undergraduate students, and from academics incorporating TDM into their teaching practices. As such, the support a library develops around TDM will be likely to need to encompass teaching and learning support as well.

'supporting TDM activity will also present opportunities for libraries to evolve their own skills and support'

Our experience of supporting TDM shows that there is a role for academic libraries. Some of the TDM support required is a natural extension of the support libraries already offer their communities. This includes advice around copyright and licensing in relation to TDM, alongside promotion of resources and datasets available for TDM purposes and the information literacy skills needed to evaluate their suitability.

However, supporting TDM activity will also present opportunities for libraries to evolve their own skills and support. This includes librarians developing introductory skills in coding and use of APIs to ensure maximum proficiency in handling TDM queries. It also includes being at the forefront of educating researchers on the literacy required to meet the opportunities and risks that developments in AI will have in a TDM (and wider research) context. We welcome

the training opportunities from CILIP, RLUK and Jisc[13] that are already available for librarians looking to begin bridging that gap, but more will need to be done to ensure that librarians working in research libraries have the skills they need to support staff and students with their TDM journeys.

Acknowledgements
The authors would like to thank Alex Fenlon, Head of Copyright and Licensing, University of Birmingham, who has driven forward TDM support within our library service and kindly reviewed a draft text of this article before submission.

Abbreviations and Acronyms
A list of the abbreviations and acronyms used in this and other *Insights* articles can be accessed here – click on the URL below and then select the 'full list of industry A&As' link: **http://www.uksg.org/publications#aa**.

Competing interests
The authors have declared no competing interests.

References

1. "Copyright, Designs and Patents Act, 1988, c. 48, s. 29A," legislation.gov.uk, https://www.legislation.gov.uk/ukpga/1988/48/section/29A (accessed 23 January 2024).
2. legislation.gov.uk, "Copyright, Designs and Patents Act, 1988, c. 48, s. 29A (5)."
3. "Birmingham Environment for Academic Research (BEAR)," University of Birmingham, https://www.birmingham.ac.uk/research/arc/bear/index.aspx (accessed 23 January 2024).
4. "What is a golden copy?," University of Edinburgh, https://www.ed.ac.uk/records-management/guidance/records/retention/golden-copy (accessed 23 January 2024).
5. "Gale Digital Scholar Lab: Open New Research Pathways," Gale, https://www.gale.com/intl/primary-sources/digital-scholar-lab (accessed 23 January 2024).
6. "Text and data mining," University of Birmingham Intranet, https://intranet.birmingham.ac.uk/as/libraryservices/library/copyright/text-and-data-mining/text-and-data-mining.aspx (accessed 23 January 2024).
7. "Centre for Corpus Research," University of Birmingham, https://www.birmingham.ac.uk/research/activity/corpus/index.aspx (accessed 23 January 2024).
8. "The Institute for Interdisciplinary Data Science and AI," University of Birmingham, https://www.birmingham.ac.uk/research/data-science/index.aspx (accessed 23 January 2024).
9. Bernard Marr, "A short history of ChatGPT: how we got to where we are today," *Forbes*, May 19, 2023, https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/ (accessed 23 January 2024).
10. Jack Grove, "The ChatGPT revolution of academic research has begun," *Times Higher Education*, March 16, 2023, https://www.timeshighereducation.com/depth/chatgpt-revolution-academic-research-has-begun.
11. Alex Fenlon, "AI: what are the risks?," *UoB PGR Development* (blog), June 26, 2023, https://uobpgrdevelopment.wordpress.com/2023/06/26/ai-what-are-the-risks/ (accessed 23 January 2024).
12. Peter Findlay, "Why should libraries publish AI-ready collections?," *Jisc Blog*, November 9, 2023, https://beta.jisc.ac.uk/blog/why-should-libraries-publish-ai-ready-collections (accessed 23 January 2024); Andrew Cox, *Research report: the impact of AI, machine learning, automation and robotics on the information profession* (London: CILIP, 2021), https://www.cilip.org.uk/page/researchreport (accessed 23 January 2024).
13. "UKeiG CPD Workshop: Artificial intelligence for librarians, information and knowledge professionals," CILIP, https://www.cilip.org.uk/events/EventDetails.aspx?id=1772016&group= (accessed 23 January 2024); "RLUK Digital Shift Forum," RLUK, https://www.rluk.ac.uk/dsf/ (accessed 16 November 2023); "Artificial intelligence and ethics," Jisc, https://beta.jisc.ac.uk/training/artificial-intelligence-and-ethics (accessed 23 January 2024).

Corresponding author:
Lisa Bird
Copyright and Licensing Advisor
Library Services
University of Birmingham, UK
E-mail: l.s.bird@bham.ac.uk
ORCID ID: https://orcid.org/0000-0001-9092-3498

Co-author:
James Barnett
ORCID ID: https://orcid.org/0000-0001-5284-4246

To cite this article:
Bird L and Barnett J, "Developing text and data mining (TDM) support within a university research library," *Insights*, 2024, 37: 5, 1–8; DOI: https://doi.org/10.1629/uksg.646

Submitted on 16 November 2023 Accepted on 12 December 2023 Published on 12 March 2024

Published by UKSG in association with Ubiquity Press.