REPORT

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Making commercial data available for research

Author(s):	Neil Smyth, Thomas Parsons & Katharina Lorenz
Audience:	Project stakeholders & JISC MRD
Published:	05/06/2013

Contents

1. Introduction

The ADMIRe project started in October 2011 and aimed to complete research data management pilots in all five Faculties at The University of Nottingham (Nottingham). Of these the Faculty of Arts was identified as key:

"The Faculty of Arts provides a stimulating and welcoming community where you can discuss thoughts, debate ideas and develop your thinking. Our great range of facilities will allow you to nourish your interest in the arts - a theatre, galleries, museum, recital hall and libraries are all here on our inspirational campus for you to use and enjoy." Professor Stephen Mumford, Dean of the Faculty of Arts

As would be expected, there is a significant body of publications arising from the Faculty on a yearly basis and significant funding revenue from RCUK. The primary funding body is the Arts and Humanities Research Council (AHRC), who along with the Nottingham Horizon Digital Economy Research Institute[1] awarded funding for a novel project aimed at exploring the notion of data within the arts. The project was called the *"Data – Asset – Method: Harnessing the Infinite Archive"* (DAM) and is an international collaboration led by Nottingham. It aims to bring together university partners in the UK and US, along with cultural institutions and industry partners in the UK, the network explores the shaping of digital archives and, in turn, technology and knowledge. Academic partners and publishers include: De Montfort University; University of Exeter; University of the West of England, Pervasive Media Studio; University of North Carolina, Chapel Hill; University of Virginia; Broadway Cinema, Nottingham; ProQuest; Cengage Learning EMEA/ Gale International Limited; the British Library and the National Archives.

ADMIRe was approached to both inform and participate in the areas of research data management (RDM) with the AHRC project, with contributions of:

1. Presentations or participation in workshops
2. Advice on research data
3. Practical support where possible.

As is the case for ADMIRe JISC MRD projects, DAM is part of a larger AHRC Digital Transformation theme, which examines the potential for digital technologies to transform research in the arts and humanities. The key goal of the DAM project was to run a series of collaborative workshop events that will lead to a sandpit event and potential new assets for research.

[1] http://www.horizon.ac.uk/Home

2. The use of research data

Of interest for our ADMIRe pilot study, was the identification that arts researchers utilise and purchase significant amounts of historical data. As expected, these take the form of manuscripts, papers and books, yet digital archives are also widely used. These would include text corpus such as The Electronic Text Corpus of Sumerian Literature which is maintained by The University of Oxford[2], through to commercial archives such as State Papers Online offered by Cengage[3]. Meetings with DAM project members and Libraries and Research and Learning Resources staff identified that a researcher will:

1. Frequently utilise text corpus
2. Create and share their own corpus
3. Licence and pay for corpus via project budgets
4. Utilise Nottingham Library resources.
5. Want to use data from licenced information resources.

Hence within the use of text corpus, it was acknowledged that RDM could have a role to play in identifying where: text corpuses are being used, helping Nottingham researchers identify what resources are available and safe-guarding the data against accidental loss. With the overall aim being to identify and make relevant research data sets available for research across Nottingham.

Archive data is often supplied through the purchase or subscription to commercial databases and Nottingham has licenced information resources from Cengage and Proquest. Nottingham has seen an accelerated adoption of digital information resources in the arts and humanities and a growing demand for access to commercial archives - thereby raising the costs to Schools and LRLR through the procurement of commercial archives on an ad-hoc basis. Alongside the cost implications, there are a range of technical, legal and other financial issues that inhibit or prevent data being made available for research. Of interest to this pilot was the licencing around data and whether or not it could be stored, shared or even catalogued if it had been purchased by Nottingham researchers or LRLR. One idea that arose from the discussions with researchers was to store commercial data on the Nottingham Research Filestore. This would avoid data loss through reliance upon USB drives and safeguard the data from accidental release or disclosure- all aspects of good RDM practice.

[2] http://etcsl.orinst.ox.ac.uk/
[3] http://gale.cengage.co.uk/state-papers-online-15091714.aspx

2.1. Licence restrictions

LRLR Arts Faculty Team Leader Neil Smyth worked with researchers and commercial corpus providers (ProQuest & Cengage) to understand how licence restrictions may impact good RDM practice. The following clauses were identified from different but current licences:

- Data mining is strictly prohibited[4].
- Automated searches against ProQuest's systems are not permitted.
- Downloading of all or parts of a Product in a systematic or regular manner or so as to create a collection of materials comprising all or a material subset of a Product is strictly prohibited whether such collection is in electronic or print form[5].
- You may use the Licenced Material to perform and engage in text mining/data mining activities for academic research and other Educational Purposes.

Therefore in certain cases it appears that it is legitimate to extract data and store it for academic use, while in other it is strictly forbidden unless an additional payment is made. A recent DAM network event at the British Library showed that there were different understandings of the licence agreement language. "Authorised Users," "Commercial Use" and many other words and phrases are explicitly defined in some licence agreements. Other phrases, such as "data mining," "text mining" and "automated searches" are not defined. There is a lack of shared understanding around how research can be used. In many instances researchers are limited to using the web interface provided by the supplier, which therefore limits the analysis that can be carried out. Nottingham has a specific department with the School of English[6] who use corpus linguistics software to analyse large text corpuses, so being unable to download the data and analyse it as a whole can impair the depth of research they can carry out.

2.2. Costs of accessing purchased data

The initial goal of the pilot from an ADMIRe perspective was to open up data that was held by Nottingham to other Nottingham departments and researchers. However, in some cases Nottingham own a licence for the data, but would still need to pay additional fees to achieve this, i.e. double licencing. Conversations with publishers produced a range of quotes:

- £5,000 for the data on hard drives for each historical newspaper archive, and where the University already owns the data and pays to access the data through the publisher interface. This has been described as a Cost Recovery Fee.
- In another instance Nottingham subscribes to an online historical newspaper archive, yet has been asked to pay around £50,000 for the publisher to provide the database and the data on a hard drive.

[4] http://www.proquest.co.uk/en-UK/site/terms_conditions.shtml
[5] http://www.proquest.co.uk/en-UK/site/terms_conditions.shtml
[6] http://www.nottingham.ac.uk/cral/index.aspx

Sometimes we have a licence that explicitly permits text and data mining but the data cannot be used for research because of technical reasons. In some cases backup hard drives will be provided upon termination of annual access, but we do not have the local platforms to make the data available.

2.3. Recommendations for RDM

What appeared to be a relatively simple technical pilot to copy data from one medium to a secure filestore has met with considerable licencing problems. It should be noted that future considerations for librarians or academics when choosing to invest in either online access or complete datasets, should include:

1. If the licence is for online access, is a usable backup of the data physically provided as well?
2. Is there an additional licence fee payable to extract online data?
3. Are we being charged anything extra for data, and is this a reasonable cost?
4. Are there any additional costs necessary within the institution before we can make the data available for research? E.g. new infrastructure, software etc.
5. Can the dataset be indexed within an institutional data catalogue? Or made available to other institutional researchers via an internal data repository?

3. Conclusion

The current licencing models favour access to a licenced dataset based upon individuals or new research project, with the costs being typically borne by the library or as a direct cost from the research grant. Due to the closed nature of research and the flux of researchers leaving and joining Nottingham, it is not uncommon for datasets to be purchased and never used outside of that research project, let alone catalogued. Therefore encouraging purchasers of datasets to explore and question the licence issues, should in theory allow greater access to datasets for interested researchers and increase the depth of research being carried out.

The Government has recognised this and proposes to amend the Copyright, Designs and Patents Act 1988, so that it is not an infringement of copyright for a person who already has a right to access a work (whether under a licence or otherwise) to copy the work as part of a technological process of analysis and synthesis of the content of the work for the sole purpose of non-commercial research. Thereby permitting text mining and in theory, allowing datasets to be stored within an institutional data repository.

European Technology SMEs, Open Access Publishers and the Research Sector have withdrawn from the "Licences for Europe" framework[7], because double licencing is not a solutions when there is an urgent need to remove existing legal, technological and skills barriers that prevent text and data mining.

In short, this pilot has shown the need for new assets for research: a legal framework so that we can access and use data; and a range of technical tools so that we can extract, store and use data.

[7] http://www.libereurope.eu/sites/default/files/Letter_of_withdrawalL4E_TDM_May%2024_1.pdf