A Shared Language for Building a Dataset of Sensitive Information *

Clare Llewellyn
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
C.A.Llewellyn@sms.ed.ac.uk

Laine Ruus
University Library
University of Edinburgh
Edinburgh, United Kingdom
Laine.Ruus@ed.ac.uk

Mark Smith
School of Social and Political Science
University of Edinburgh
Edinburgh, United Kingdom
Mark.Smith@ed.ac.uk

Steve Kirkwood
School of Social and Political Science
University of Edinburgh
Edinburgh, United Kingdom
s.kirkwood@ed.ac.uk

Ros Burnett
Centre for Criminology
University of Oxford
Oxford, United Kingdom
ros.burnett@crim.ox.ac.uk

Robin Rice
University Library
University of Edinburgh
Edinburgh, United Kingdom
R.Rice@ed.ac.uk

1. INTRODUCTION

Using text analysis tools to study large data sets is currently an area of popular interest. Prompted by the success of several big data research initiatives, researchers from a wide variety of disciplines may wish to gather and analyse data using text analysis tools created during these big data initiatives [2]. Communication between members of diverse teams can present a problem and developing a shared language and understanding of the task may provide a soultion [1].

In this work we look at the initial phase of an ESRC funded project involving academics from Social Work, Criminology, Informatics and the University Library. This project aims to secure a data set on allegations of sexual abuse made against the former disc jockey, Jimmy Savile. The Savile affair has taken place in a public and highly charged, arena. It has generated massive media attention and spawned several public reports, most notably that which was produced as a result of Operation Yewtree [6]. Major institutions such as the BBC have been affected by this case. The case has become a powerful driver of public policy around the investigation of current and retrospective child sexual abuse cases. Our research may either confirm or destabilise the narrative that has built up around Savile [3].

Early allegations against Savile emanate from former residents at Duncroft, a residential school for 'wayward but intelligent young women'. This project stems from data produced and collected by the blogger 'Anna Raccoon' herself a former resident at the school. Through her blogs on the subject of Savile and Duncroft she was contacted by others and has collected a variety of information on the subject. Unfortunately as Anna is in ill health the long-term security of this information resource was under threat; in fact the original blog was taken down prior to the start of the project. The initial research aim of this project was to secure documentary and interview data in relation to the Savile case before opening up our findings to wider consideration and scrutiny. The first point of data collection was the blog itself.

We found that in order to communicate effectively we had to develop a shared language between project participants. In this paper we discuss how we have come to a shared view of the challenge; about what it means to harvest and archive data, what types of analysis are possible and what are the ethical implications of working with controversial data taken from multiple sources.

2. CREATING A SHARED VISION

In order to design the archive and determine the types of access required each of the social science researchers involved in the project was interviewed. A set of questions was asked of each of the researchers to determine how they would work with the data and how they envisioned the archive. Each of the researchers was then shown a variety of online archiving tools. The aim of asking the questions in this manner was to initially allow complete freedom to the researchers to ascertain the researchers' image of what the system should be. The current tools were then used to demonstrate what type of functionality and behaviours were currently possible. It was hoped in this way to elicit a good fusion of what the researchers wanted and what currently avaliable tools could achieve.

The results of these surveys were interesting. It was clear that all of the researchers had a very clear idea of the data that was being collected. There was some disagreement as to whether the project would be focussing on the archiving or the re-use of the data and therefore both must be judged as equally important. Each interviewee had a very clear view about what they and others would do with the data and gave extensive descriptions of how they would interact with the data, from using print outs and highlighters to specialist qualitative data management systems are then used as a starting point for the system design. Throughout the interviews the idea of cross referencing and checking came out as a particularly important aspect of what the researchers will be doing. The sense of who the people, the locations and the dates are was important. Also important was to retain a bigger picture about what is stored in the archive as requested 'I'd like it to focus on the items in the story but not to ignore those telling the story.' It was clear that one of the concerns of the researchers was the controversial nature of the data. The data are unusual because the topic is sensitive and this means that any automatic analysis will need to be robust so claims made by researchers can be supported.

3. DATA COLLECTION

The initial component of the project involves capturing Anna Racoon's blog (The Racoon Arms). This is a WordPress blog and was only available on the Internet Archive's WayBackMachine (WBM). An active blog is a constantly evolving object, and therefore careful consideration needs to be given as to what version or versions should be harvested. Given that the blog is available via the WBM, one might question why it is necessary to download a copy at all. There are two main reasons for doing so. Firstly, the Internet Archive may at any time, and without notice, remove the objects from their archive. Secondly, to provide additional functionality to support qualitative analysis of the content of the blog, as well as indexing to support additional resource discovery not provided within the blog software or the WBM.

While harvesting the contents of a blog manually can be a long and arduous process, it can be simplified and automated using a software solution, such as *wget*. Apart from soliciting permission from the Internet Archive, decisions need to be made as to which version or versions should be harvested. Further decisions included to what level of recursion each harvest should be and whether just blog text or all files contributing to content and functionality of the blog should be gathered. Such decisions influence not only the size of the eventual object, but also the richness of the context. There are also concomitant draw-backs – the deeper the recursion, the greater the number of missing files (those that have not been harvested by the WBM). Given that WordPress blogs are based on HTML format files, apart from any images and other audio-visual files that may be associated with the blogs, the text portion is in as efficient a format as possible vis-a-vis file storage as well as capacity to use XML to provide value-added indexing and tagging. Storage capacity requirements depend largely on the number of snapshots of the blog that are harvested and the level of recursion specified in the harvests. The size of one snapshot can range from 53 MiB to 660 MiB (ranging from 1,500 to 88,000 files), depending on the options specified.

ESRC funding requirements relating to data deposit, confidentiality, and open access requirements will need to be applied differentially to the various research inputs that will comprise the research data for this project. Certainly the blog portion of the data, harvested as it is from a public web server, is anticipated to pose no difficulty vis-a-vis adherence, especially in its value added form. The e-mail and personal interview components of the project, will, however, pose additional issues involving privacy and confidentiality.

4. DATA ANALYSIS TOOLS

There is a clear requirement of the identification of specific people, places and dates from within the dataset. The data will be stored in the archive in the original format but in order to be processed a second set will be created by converting the data to XML. It will be processed using the Edinburgh Informatics information extraction tools which include LT-TTT2 and the Edinburgh Geoparser [4, 5]. These are well established tools used for text mining. These tools are used on various big data projects including Trading Consequences, a project from the 2012 Digging into Data Initiative [7].

5. LEGAL, MORAL AND ETHICAL IMPACTS

The subject matter of the data requires that an Advisory Group was set up from the outset, with representatives of various organisations, including those representing victims of historical abuse but also members of Jimmy Savile's family They will be consulted on how best to communicate difficult findings in a balanced way which is fair to the needs and perspectives of all parties and advances in criminal justice policies to prevent child sexual abuse and to bring offenders to justice. We have produced an introductory letter that outlines our proposed research. We require written consent at two different levels. Firstly, for the use the data as part of this research and secondly, for the data to be stored indefinitely and made available to other researchers for secondary analysis. The proposal has been considered and approved through the University of Edinburgh's School of Social and Political Science research ethics procedures. We have established direct discussion with members of the School's Ethics Committee and are maintaining ongoing contact with them to discuss unforeseen matters that might arise in the course of the research.

6. ADDITIONAL AUTHORS

Additional authors: Rocio.von-Jungenfeld University Library, University of Edinburgh, email:
`Rocio.von-Jungenfeld@ed.ac.uk`)

7. REFERENCES

[1] Langfield-Smith, Kim. Exploring the need for a shared cognitive map. Journal of management studies 29, no. 3 (1992): 349-368.

[2] Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela H. Byers. Big data: The next frontier for innovation, competition, and productivity. (2011).

[3] Clapton, Gary, Viviene E. Cree, and Mark Smith. Moral panics and social work: Towards a sceptical view of UK child protection. Critical social policy 33, no. 2 (2013): 197-217.

[4] Grover, Claire, and Richard Tobin. Rule-based chunking and reusability. (LREC. 2006)

[5] Grover, Claire, Sharon Givon, Richard Tobin, and Julian Ball. Named Entity Recognition for Digitised Historical Texts. (LREC. 2008)

[6] Gray, David, Watt, Peter. Giving victims a voice (2013)

[7] U. Hinrichs, B. Alex, J. Clifford, A. Quigley. Trading Consequences: A Case Study of Combining Text Mining and Visualisation to Facilitate Document Exploration (to appear at DH2014)