Citation: Lin, Y. (2012) 'Trans-disciplinarity and Digital Humanity: Lessons Learned from Developing Text Mining Tools for Textual Analysis', in: Berry, D M (ed.), Understanding Digital Humanities, Palgrave Macmillan, Basingstoke, UK.

Trans-disciplinarity and Digital Humanity: Lessons Learned from Developing Text Mining Tools for Textual Analysis

Yuwei Lin
School of Media, Music and Performance
University of Salford

Abstract

This paper, based on a case study of developing a set of bespoke text-mining tools for researchers in humanity in the UK, provides an empirical account of trans-disciplinary research practices across the social sciences and humanity. Through looking at the user-participatory development processes of the text-mining tools, the piece improves our understandings of digital humanity in the context of academic research, and highlights its trans-disciplinary characteristic from a pragmatist perspective. The paper concludes with the discussion of some methodological and socio-technical challenges of “digital humanity” emerging in this shift towards trans-disciplinarity.

Keywords: trans-disciplinarity; inter-disciplinarity; Digital humanity; Text-Mining; textual analysis; knowledge production;

Introduction

In recent years, with the emergence of Information and Communication Technologies (ICTs) and other social and political factors, national and international research funding councils have increasingly emphasised that research in humanity should engage with data intensive and evidence-based academic activities, as others in natural sciences and engineering do. As stated in the description of cross-nation and cross-discipline “The Digging into Data Challenge” programme[1], a call for “data-driven inquiry” or “cyberscholarship” has emerged as a result of hoping to inspire innovative research methods, to transform the nature of social scientific inquiry and create new opportunities for inter-disciplinary collaboration on problems of common interest[2].

New types and forms of data, may it be born digital data, transactional data, digitised historical records, archived administrative data, linked databases, or data generated or shared by Internet users, is all considered to be valuable input for research. And in order to facilitate access to and process such a massive amount of data, information technologists and computer scientists have been involved to construct high-throughput, high-performance computing, grid computing or cloud computing for research in humanity. e-Research (or Cyber-Infrastructure in the USA) has been proposed as an umbrella term to describe such computationally enabled science that allows researchers from distributed

1 The “Digging into Data” initiative, launched in 2009, is sponsored by eight international research funders, representing Canada, the Netherlands, the United Kingdom, and the United States. The second round of the competition took place in 2011. For more information please visit the website http://www.diggingintodata.org/

2 See the 2011 Digging into Data request for proposals document at http://www.jisc.ac.uk/media/documents/funding/2011/03/diggingintodatamain.pdf

locations and diverse backgrounds to access, discuss, analyse data and work together. That said, such a shift to large-scale networked infrastructures for supporting research not only highlights “big data” and computational data analysis methods, but also suggests the importance of research collaboration across disciplines. The “Digging into Data” programme sponsored by eight international research funders shows that research funders have also recognised that the complexities of subjects in society are beyond what a single discipline can deal with, hence inter-disciplinary or multi-disciplinary collaboration is needed. To address these challenges, research councils have been encouraging social scientists to adopt collaborative approaches, to share and reuse data, to explore and exploit mixed methods, and to develop innovative methods. To these ends, not only have various novel e-Research tools and services been created over the past years, but also a growing number of large-scale collaborative inter-disciplinary research projects have been funded.

The development and implementation of these e-Research tools have signified and signalled a dramatic *computational turn* in conducting research in humanity. Digital humanity has been heralded as the future of humanity research. e-Research programmes often emphasise inter-disciplinary and/or multi-disciplinary (Schroeder & Fry, 2007). Although to some extent these existing observations are valid, I will argue in this paper that the kind of digital humanity facilitated by e-Research tools, if widely adopted, is in fact trans-disciplinary, a step further than multi-disciplinary or inter-disciplinary. The realisation of trans-disciplinary research can be illustrated through looking at the process of developing text mining tools for social and behavioural scientists in the case study to be introduced below. I will discuss the challenges and implications of such trans-disciplinary research in light of this case study. The empirical case study provided here also contributes to the ongoing and long-standing discussion about inter-disciplinariy and trans-disciplinarity.

Before introducing the the case study that demonstrate the development process of text mining e-Research tools, I will provide some context of and elaborate what I meant by trans-disciplinarity.

Trans-disciplinarity

Many terms have been proposed over the past decades (arguably since 1960s) to conceptualise contemporary scholarly activities. Inter-, multi-, and trans-disciplinarity are the three widely recognized categories used to measure, analyse or identify inter-disciplinarity in actual research efforts (Huutoniemi et al., 2009). They suggest approaches that differ from existing disciplinary norms and practices.

Multi-disciplinary and inter-disciplinary research have been growing over the last four decades. They are not new concepts in scientific research. In his seminal work, *The Social and Intellectual Organization of the Sciences* published in 1984, Whitley has argued that, in addition to what they study empirically, scientific fields are shaped and affected by the degrees and types of *mutual dependence* and *task uncertainty* they possess (Whitley, 1984, p. 88).

The book *The New Production of Knowledge* published by Gibbons et al. in 1994 proposes a *Mode 2* knowledge framework which has received far reaching influence, especially in setting out EU research agenda. It is said that there are three prerequisites are needed to produce Mode 2 knowledge: a context of application to allow knowledge transfer, trans-disciplinarity, a diverse variety of organisations and a range of heterogeneous practice, reflexivity, a analogical process where multiple views in the team can be exchanged and incorporated (Gibbons et al. 1994; Hessels & van Lente, 2008). *Mode 2,* which is context-driven, problem-focused and trans-disciplinary, involves multi-disciplinary teams with heterogeneous backgrounds working together. This differs from traditional mode 1 research that is academic, investigator-initiated and discipline-based knowledge production. Nevertheless, to mark the distinction of *Mode 2*, the trans-disciplinarity is the key, and according to Hessels & van Lente (2008), it "refers to the mobilisation of a range of theoretical perspectives and practical methodologies to solve problems" and goes beyond inter-disciplinarity in the sense that the interaction of scientific disciplines is much more dynamic." (p. 741).

Whitley's theory of '*mutual dependence*' and '*task uncertainty*' and the *Mode 2* theory proposed by Gibbons et al., and philosophical and sociological discussion on the production of scientific knowledge (now often termed "science and technology studies – STS" e.g., Latour and Woolgar, 1979; Knorr-Cetina, 1982; Latour, 1987; Klein, 1990) have inspired many scholars to explore how inter-disciplinarity, multi-disciplinarity, cross-disciplinarity, or even trans-disciplinarity approaches (Flinterman *et al*., 2001) are perceived and performed in different research fields, particularly in computer-supported environments. For instance, Barry *et al.* (2008) have conducted a large-scale critical comparative study of inter-disciplinary institutions based on ethnographic fieldwork at the Cambridge Genetics Knowledge Park, an internet-based survey of inter-disciplinary institutions and case studies of ten inter-disciplinary institutions in three areas of inter-disciplinary research: a) environmental and climate change research, b) the use of ethnography within the IT industry, and c) art-science. Fry (2003, 2004, 2006), whose research aims to understand similarity and difference in information practices across intellectual fields, has conducted qualitative case studies of three specialist scholarly communities across the physical sciences, applied sciences, social sciences and arts and humanities. Schummer (2004) examines the patterns and degrees of inter-disciplinarity in research collaboration in the context of nanoscience and nanotechnology. Mattila (2005) studies the role of scientific models and tools for modelling , and re-conceptualises them as "carriers of inter-disciplinarity" that enable the making of inter-disciplinarity. Zheng et al. (2011) examines the development process of UK's computing grid for particle physics (GridPP), a grid that is itself part of the world's largest grid, the Large Hadron Collider Computing Grid, within a global collaborative community of high-energy physics. Most of these studies use qualitative research methods (notably ethnography and interview) to produce case studies that focus on how members of a project that involves more than one discipline communicate, negotiate and cooperate, instead of measuring quantitatively the degree of heterogeneity of knowledge combined in research[3].

3 Beaulieu et al. (2007) have questioned the surplus of (ethnographic)case studies on e-Science to-date and urged a need for conceptualising and theorising existing cases, especially from a perspective of science and technology studies (STS). Parallel to this qualitative-based stream of research, quantitative research methods such as econometrics, statistics, or bibliometric methodology are also used in studying interdisciplinarity (e.g., Morillo et al., 2003;

In a similar fashion, this case study, based on my participatory observation, offers another channel of getting to know prospective users and involving them in the process of tool development. This will not only contribute to the continued discussion of what constitutes and inter-disciplinary work; more importantly, through understanding how that work is organised in the field of social sciences and humanity, it provides an empirical glance into trans-disciplinary and what it means by “digital humanity”.

However, it has also been noted that to date there remains an incoherence in the usage of these terms, which are largely “loosely operationalized” (Huutoniemi et al. 2009: 80). Fuzzy definitions of these words mean that these categories are “ideal types only” and serve mainly for theoretical discussion. Given this, before going on to present the case study, it is useful to make clear the working definitions of inter-, multi- and trans-disciplinarity in this paper. For the purposes of this paper, inter-disciplinarity is referred to as an approach that allows researchers to work jointly and to integrate information, data, techniques, tools, perspectives, concepts, and/or theories from two or more disciplines or bodies of specialised knowledge to tackle one problem. Multi-disciplinarity, instead, allows researchers from different disciplines to work in parallel with each other but still from disciplinary-specific bases to address common problems. Trans-disciplinarity radicalizes existing disciplinary norms and practices and allows researchers to go beyond their parent disciplines, using a shared conceptual framework that draws together concepts, theories, and approaches from various disciplines into something new that transcends them all (Rosenfield, 1992: 1351).

Here, for the purpose of this manuscript, I have adopted Hessels & van Lente's interpretation of Mode 2, that is, “the trans-disciplinarity proposed by Gibbons et al. implies more than only the cooperation of different disciplines” and “co-evolution of a common guiding framework and the diffusion of results during the research process” are central to trans-disciplinary research (Hessels & van Lente: 751). Against this framework, disciplines involved in inter-disciplinary or trans-disciplinary research possess richer dependency than those involved in multi-displinary research. Therefore, it drives a closer investigation into how researchers in different disciplines interact and transform over a period of inter-disciplinary or trans-disciplinary project.

Background of the Case Study

The case study below is based on a 18-month ethnography of an interdisciplinary collaborative project funded by a UK higher education funder[4].

With the information overload and data deluge, to be able to locate information within a short period of time and how to conduct a thorough literature review, collecting and analysing data smartly and efficiently is one of the important milestones of the next-generation computational tools. In light of existing examples in natural and life sciences

Schummer, 2004).

4 The data has been anonymised due to research ethics.

where scientists use text-mining and data-mining tools to identify continuities and discontinuities in large bodies of literature or datasets, the initial idea set out by the funder was to demonstrate the usefulness of text-mining for the purpose of facilitating knowledge discovery, elicitation and summarisation in humanity. If these techniques could be successfully applied to social scientific data, it was hoping that not only could the time-consuming and labour-intensive manual coding of qualitative data be replaced (at least to some extent), but also enable social scientists to explore larger amounts of such data in a shorter time.

The project was funded to customise a range of pre-existing text-mining tools for application in studies analysing newspaper texts to reveal how contemporary events and issues are framed to shape the perceptions of their readers. And in so doing, the demonstrator produced by the project would provide a use case to extend awareness and promote adoption of text mining across all social science disciplines.

The project was designated to be an inter-disciplinary collaboration where the pilot social science users (hereafter “domain users”) work with text mining developers (in short, text-miners). Instead of developing everything from scratch, customising pre-existing tools would allow the developers to demonstrate the functionality and applicability of text mining tools for to target users as well as to the funder in a relatively short period of time. The original plan included an activity that resembled *the Turing Test* - a competition between the text-mining (artificial-intelligence enabled) programs and ordinary human researchers to find out whether a computer can act “more efficient and more accurate” than a person. This was to be a comparison between computer-generated results and human-coded ones. As some participants in such kind of Turing Test have revealed (e.g., Christian, 2011), the march of technology isn’t just changing how humans live, it is raising new questions about what it means to be researching humanity and reading texts. Similarly, as will be discussed below, this 18-month project turned out to be more than a feasibility study on the technicality and performativity of text mining tools in the context of humanity research; more importantly, it sheds light on a methodological change and a shift of disciplinary practices.

Through closely participating in the project as a project manager, pilot user, as well as an ethnographer, my ethnography produced first-hand account of working with the stakeholders (including PI, co-Is, developers, other domain users and the funder) as well as close observation of the dynamics emerging in the development process and inter-disciplinary collaboration. The reflection from the auto-ethnography and traditional participatory observation offer fresher insight into the actual work practices in the cross-disciplinary or inter-disciplinary research projects for better understanding of how these text-mining computational techniques are actually implemented and situated in real-life projects.

Every development tasks and activities in this project, ranging from constructing a database/corpus for carrying out text-mining tasks and training the algorithms to meet the needs of the pilot users, selecting and filtering out meaningful human-comprehensible terms, to communications between different project partners, all suggest that text-mining or other e-Research tools do not emerge out of the blue; instead, its realisation is the result of a negotiation of different disciplinary methodologies, practices, epistemologies,

and sense-making. As such, implementing any e-Research tools like the text-mining ones discussed here would suggest the move towards a settled/agreed/presumed/prescribed way of conducting research. As we will see later, adopting text-mining tools insinuates a radical shift from allowing diverse methodologies and theories to co-exist in arts and humanities (hermeneutic readings) towards pattern-matching, statistics-led, algorithm-based practices which favour a statistical modelling-based mining paradigm. Given that, text-mining leads to a trans-disciplinary paradigm shift.

What is text mining?

The state of art and the way "text mining" is referred to is more than text search. According to M. Hearst, a Professor in the School of Information at University of California, Berkeley,

> "Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation."[5]

According to the UK JISC-funded National Centre for Text Mining (NaCTeM),

> "Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. These various stages of a text-mining process can be combined into a single workflow."[6].

These explanations suggest that text-mining is considered as a set of technologies for "extracting more information than just picking up keywords from texts: names, locations, authors' intentions, their expectations, and their claims" (Nasukawa and Nagano, 2001). It is so applied that IBM for example has developed it further into sentiment analysis that can be used in marketing, trend analysis, claim processing, or generating FAQs (frequently asked questions)[7].

Given that, text-mining can be understood as an umbrella term for incorporating and implementing a wide range of tools or techniques (algorithms, methods), including data mining, machine learning, natural language processing, artificial intelligence, clustering, knowledge mining and text analysis, computational linguistics, content analysis and sentiment analysis and so forth, onto a large body of texts (usually an enormous collection of documents) to support the users' decisions-making. Just like Lego units, there are a set of components in the field that can be assembled and reconfigured for the purposes of the tasks of the domain users.

To illuminate what text mining can do, the text miners demonstrated some existing

5 http://people.ischool.berkeley.edu/~hearst/text-mining.html
6 http://www.jisc.ac.uk/publications/briefingpapers/2008/bptextminingv2.aspx
7 http://www.trl.ibm.com/projects/textmining/index_e.htm

applications, notably in the bio-medical field, to the social scientific domain users in the beginning of the project. One of the examples is similar to what Uramoto et al. (2004) developed - an application named MedTAKMI, which includes a set of tools extended from the IBM TAKMI (Text Analysis and Knowledge MIning) system originally developed for text mining in customer-relationship-management applications, for biomedical documents to facilitate knowledge discovery from the very large text databases characteristic of life science and healthcare applications. This MedTAKMI dynamically and interactively mines a collection of documents to obtain characteristic features within them. By using multifaceted mining of these documents together with biomedically motivated categories for term extraction and a series of drill-down queries, users can obtain knowledge about a specific topic after seeing only a few key documents. In addition, the use of natural language techniques makes it possible to extract deeper relationships among biomedical concepts. The MedTAKMI system is capable of mining the entire MEDLINE database of 11 million biomedical journal abstracts. It is currently running at a customer site.

What is Textual Analysis

The domain users, in turn, also demonstrated how textual analysis is usually conducted, and how, in this case, the analysis of newspaper content is carried out.

The analysis of newspaper texts has been widely adopted for investigating how texts[8] are explicitly or implicitly composed and presented to re/present certain events in various forms of mass media and to shape the perceptions or opinions of the information's recipients. It is a labour-intensive form of analysis, typically relying on the researcher to locate relevant texts, read them very closely, often more than once, interpret and code passages in the light of their content and context, review the codes and draw out *themes*. As a result, research projects are often restricted to corpora of limited size.

It is not novel to use computer to assist human analysts to conduct textual analysis. Computer-assisted qualitative data analysis software (CAQDAS) is a competitive market; there are many mature CAQDAS packages available (Koenig, 2004, 2006). Although some claim that CAQDAS tools support mixed methods (i.e., combination of qualitative and quantitative methods) (e.g., Bolden & Moscarola, 2000; Koenig, 2004, 2006), the requirement of large amount of human labour in using CAQDAS for coding emphasises the importance of the more interpretative, qualitative elements of textual analysis. Given that the amount of textual social data is growing at an unprecedented speed, a scalable solution which can support automatic coding and clustering of text for the textual analysis of large corpora is desirable.

Developing Text-Mining Tools for Textual Analysis

Despite the effort of establishing a dialogue between the domain users and the text miners

8 Textual analysis can be applied to a variety of forms of texts including visual, textual or audio. However, in this project, text-mining techniques are being applied to the written text solely.

in this inter-disciplinary project, such mutual sharing seemed asymmetrical in this instance. The text miners were more interested in building up a large dataset of textual data and acquiring the code book of the domain user who was conducting a research on how certain governmental agenda was re-presented in UK national newspapers. The reason why the text miners were so keen on acquiring the domain user's code book was because they needed to use that document to categorise some exemplary documents in the corpus that contains thousands of news articles. As we will see below, such instrumental / pragmatic attitude of the text miners pose some issues in this inter-disciplinary collaboration.

The process of the tool development can be described in the following steps:

Step 1: Scoping and data preparation

The process of knowledge discovery and data mining has never been straightforward; it involves many steps, and some of them are iterative and contingent (Kurgan & Musilek, 2006). Data preparation (including scoping, data cleaning) is an important first step before processing the data (e.g. Fayyad et al., 1996).

This scoping stage allowed the domain users and the text miners to know the domain and the data better. Having the domain users on board meant that the inter-disciplinary team could have some quick access to this body of knowledge.

The scoping stage also involved the identification of datasets. For the domain user who was carrying out a baseline textual analysis using a popular CAQDAS package, a rather typical social research process was followed - she began by identifying a suitable topic and research question. These activities later on informed her of the generation of a set of keywords, which were submitted as queries to the search engine of a digital archive of UK newspapers. A corpus, including all newspaper items (news, comment, letters, sport and so on) containing the key words/phrases, was built by using the search facility of this newspapers archive.

On the other hand, the text miners devised an algorithm to build a corpus of nearly 5,000 newspaper items (or ‘documents’) by extracting them from the same archive. This was an order of magnitude larger than the domain researcher's corpus because text-mining tools work best on large corpora.

What's also interesting is the way the two corpora were built. The human analyst and the text miner took different steps and actions when building these corpora. The smaller corpus was built by the human analysts with a goal of having 200-300 items in the corpus. The human analyst, bearing the research questions in mind, went through the articles that came up from the keywords searches one by one, judged, selected, and then included them in this smaller corpus. The interpretation started even before retrieving articles from the archive. Decisions on which topic to study, which type of data (newspaper or other printed media, national papers only or tabloid included) to look at, which keywords to search for, which way to collect data all flag important steps in research processes. Contrasting to the human analyst's approach of building a small-but-beautiful quality

dataset after carefully reviewing the source of data, the text miner's indiscriminative method of building a corpus as large as possible signals a fundamental difference between the two. The text miner applied the same keywords that the human analyst used to retrieve data from the same database. Data retrieved except for those from the local newspapers were all included in this larger corpus.

The inclinations to different corpus sizes of the domain users and the text-miners is interesting. For the domain users, what corpus size should be considered as representative is mainly to do with one's research questions. But a text-mining/data-mining turn has made the size of a corpus independent of the research question. In fact, text miners usually claimed that some "unexpected" clustering results may come out of the data, and this aids the limitation of human interpretation. The text miners claimed that a bigger corpora with more documents would allow users to reduce noise by ignoring common words that carry little contribution to the analysis. If users wanted to find (lexical) patterns, the larger data set for training purposes the better. According to the text miners, a sensible clustering usually needs 2,000-4,500 documents (short articles with 10 sentences usually).

Step 2: Data analysis and training the algorithms

Once the smaller corpus – comprising 200-300 items – was constructed, the researcher undertook a 'traditional' textual analysis, reading the newspaper texts and analysing them in the light of her review of related documents and policy statements, using a CAQDAS tool to manage the hermeneutic coding process, and identifying themes through an iterative process of re-reading the full articles and examining the coded segments of the texts. The quality of the analysis was assured by presenting the substantive results at conferences; all these were well-received.

In order for the process of conducting the baseline textual analysis to feed requirements to the text miners, the domain researcher met with the text miners occasionally to brief them and demonstrate her use of CAQDAS. The domain researcher showed the text miner how she built her own corpus and how she used a CAQDAS tool to code it; the text miner showed the domain researcher what text-mining tools were available and how they functioned. Ethnographic notes were taken on most of these meetings. In addition to learning from each other, the domain users and the text miners attended a CAQDAS training course where several CAQDAS packages were introduced and their strengths and weaknesses were reviewed. It provided the text miners with an opportunity to extend their knowledge of how social scientists conduct qualitative research aided by CAQDAS packages, discover what kinds of data they commonly analyse and the databases available to them, how they import the data into the packages, and the extent to which the packages automate the process of hermeneutic coding. Lastly, the domain researcher also produced a short report on how coding was undertaken within the usage of the CAQDAS tool, and how themes emerged through inductive reasoning, together with a detailed codebook. These materials, produced by one single researcher, were used to train the text-mining tools to search the content of the documents in the corpus.

Step 3. Software development

One component in the text-mining tool set is automated term extraction, where a term in this context refers to a compound of two or more words (or lexical items). This tool automatically generates, for each document separately, a list of terms that are significant within it. The users had the option to select one of three levels of significance, high, medium or low, and this affected the number of terms appearing in the list, the minimum being five or six of high significance.

Another text-mining tool clustered documents in the corpus by estimating the degree to which their content fits together. When the users entered a query on the system's search screen, the system returned a list of cluster titles on the left hand side of the screen. Clicking on one of them brought up a screen listing all the documents relevant to the query phrase within that cluster.

A third text-mining tool was a named entity recogniser, that is, a tool to identify the names of, for example, people or organisations. The users had the option of displaying the named entities contained in all the documents returned by a search. These appeared in pre-defined groups, such as, country, location, person, and organisation.

A fourth component of this text-mining system was a sentiment analyzer, which calculates a positive or negative score of each sentence in a document according to values pre-assigned to each word it contains. Sentences on the screen were shaded from dark through light green to light and then dark red to represent the magnitude of the positive through to negative score.

To develop these tools, the text miners attached tags (terms in the domain user's code book) to a document or to a sentence so that meanings were inferred to a sentence or a document. Unlike the domain researcher's inductive way of coding, the text mining method appeared to be deductive and positivist. For example, the domain researcher started from zero code, and as soon as she found something as she read, she created new codes. This was part of a process of "reading". This intuitive interpretative flexibility cannot be found in the text mining process as it needs a text miner to infer a fixed meaning to the original documents in the large corpus.

At this stage, the text-miners worked mostly alone with few interaction with the domain users. The infrequent communication between the text-miners and the domain users also suggests an asymmetrical relationship in this inter-disciplinary collaboration (as mentioned earlier). Social scientific expertise was brought in to meet the practical purpose of computing development. The relationship with the domain users was disconnected temporarily once the text-miners collect enough information for their development, and this temporary decision of jointing and disjointing / re(arranging) domain disciplinary expertises during the course of the project poses a risk to the inter-disciplinary collaboration in this project. That is, the team members had a lack of trust and limited understanding of each other's work.

Step 4. Iterative development

The development of these tools underwent a series of iterative and continuous development (including fine-tuning) to ensure the software returned the right documents and highlighted the right/meaningful phrases desired by the domain researchers. This stage involved a series of user trials to identify the shortcomings and increase the accuracy, the quality of the software. The domain users typed a keyword into the search field of the designated text-mining interface, which appeared like a Google search page, and inspected the returned documents and results. At the Alpha testing stage, in the eyes of human readers the documents returned or the sentences highlighted by the software were inconsistent in terms of meanings and semantics. Often a word, a phrase or a sentence was highlighted not because of its meaning in the context but because of its lexical. It was challenging to produce coherent and mutually exclusive categories required more remedial action at the pre-processing stage or at the mining stage.

When the results were unsatisfactory, the domain users would like to know how the search results were produced, how relevance between words and phrases was calculated and perceived, whether it was because of word frequency or some other modelling techniques.

Impressions on Text-mining in Humanity

Although the text-mining software system described above was incomplete, the project was proved to be an excellent feasibility study of the opportunities for, and threats to, extending text mining to the textual analysis of newspaper texts and more generally to qualitative social research.

The domain users in the interdisciplinary project found that the text-mining system provided a user-friendly entry into text-mining, with the initial screen – the search interface – resembling mainstream search engines and the results appearing in a familiar layout: a paginated list of the titles of the returned documents, their authors and dates, and snippets from each document in which the query word or phrase was highlighted. However, the term extraction and clustering results were found wanting in two respects. First, users were reluctant to accept ‘black-boxed’ results; instead, they wanted to know how the terms were extracted and the clusters created by the text-mining tools, this knowledge being critical to their judgement of the validity of the results. This poses a quandary: the more complex the search algorithm, the more successful it is likely to be at classifying documents according to their main theme (summarised by a term), but the more difficult it is likely to be to explain how the algorithm works to a user who is not a text-mining expert. The current system, where no explanation was offered and the phrases used to represent the terms were often obscure, left users inspecting the contents of the returned documents in an attempt to infer why they were clustered according to a specific term. They then encountered the second problem: there were often several hundred documents clustered under one term and users found themselves opening each one and reading its contents. The potential efficiency of the system was therefore lost; users were reading large numbers of documents to ascertain their meaning, just as they would in a ‘traditional’ textual analysis. Moreover, the system lacked the data management aids common in CAQDAS packages, leaving users hampered by clumsy navigation.

During the development of the system, the performance of its term extraction tool was improved by one single domain researcher applying her knowledge gained in the baseline study to evaluate and edit a long list of terms generated by the software. It was an advantage to have a system that can be trained in this respect. However, it was a disadvantage to have the quality of the system's output dependent on the extent of the prior training effort put into it by a domain expert, although this would be mitigated if the subsequent users' confidence in the results were increased. In that case, quality assurance would be provided not by understanding how the term extraction algorithm works but by knowing that a domain expert had trained the system to validly identify the main topics of each article.

In the light of their experiences of using the system, the users reported that they could envisage a scenario in which document clustering would be valuable. This would be as a preliminary scan through a very large number of documents, because it would reduce the number to be inspected to those clusters that the investigators found of interest. Similarly, they reported that term extraction could be useful if it were based on a domain expert extensively training the system in a preliminary study and then it were used by others in a large-scale follow-up study. Alternatively, there might be scope to use both tools in the first stages of a new textual analysis to generate some preliminary ideas about topics appearing in a large corpus, though this would need to be followed by a 'traditional' textual analysis to examine the emerging ideas in full detail.

Although users found the named entity tool straightforward to use, and the results intuitively understandable, they reported that it had three limitations. The first was that it did not immediately appear to have any advantages over using a keyword search in a standard search engine or in a CAQDAS package, although they recognised that the advantages might become apparent were the tool used in research where its disambiguation functionality was particularly important. The second was that the names that appear in the categories were taken from a pre-defined dictionary and the tool would therefore miss some of the persons, organisations, celebrities etc appearing in the newspaper texts. The third was that there is little social research in which identifying named entities contributed significantly to the interpretive analysis of qualitative data.

The sentiment analysis tool was also straightforward to use but it found little favour among social scientists because they were aware of too many issues about language use and sentence construction that undermine the validity of scores for each sentence based on the individual words it contains.

Overall, using this pilot text-mining system raised two fundamental issues. One is the question of what semantic content mined from texts would be most useful to qualitative social researchers. A case could be made for the terms extracted by the text-mining system but that would involve explaining the routines that calculate their significance. The other, related issue is how to present the text-mining tools in a way that builds trust among domain researchers that the results are valid.

One of the potential benefits of the system was its capacity to process enormous amounts of texts very quickly. However, this benefit was compromised when searches produce

terms that were accompanied by long lists of results spread over dozens, even hundreds, of screens. Only if the user had confidence in how the terms were extracted would she be willing to take the results at face value. Yet users' confidence in the extraction of terms (alternatively described as coding the qualitative data), which lies at the heart of all qualitative analysis, was normally built up through an iterative process of reading and re-reading the texts until the analyst feels that she had fully grounded the codes in an interpretive understanding of the texts, recognising that there is an inevitable relation between the phrases coded and the contexts in which they appeared. The semantically richer the analysis that is sought, the more effort is invested by the analyst in extracting meanings. In general, the more unstructured the texts and the less limited the domain, the more difficult the task is. This might be expressed as a continuum from (A) highly structured text about a limited domain to (B) very unstructured text about an almost unlimited domain. A might be represented by bio-informatics journal articles, for which many existing text-mining tools were developed, through newspaper texts to informal interviews, conversations and blogs, representing B. At A, quantitative measures such as word counts, word proximities and so on might suffice to summarise the meaning of the text. At B, much more interpretive effort is required.

Although A to B has just been described as a continuum, it is a matter of continuing debate across numerous social science disciplines as to whether there is discontinuity or break somewhere between the two poles. In linguistics, this appears in the discussions whether semantics can be captured through syntax; in social research it appears in the arguments whether quantitative content analysis is fundamentally different from qualitative textual analysis.

Beyond inter-disciplinarity: A Methodological Transformation and Trans-disciplinarity

In light of Barthes (1977), inter-disciplinary research "must integrate a set of disciplines so as to create not only a unified outcome but also something new, a new language, a new way of understanding, and do so in such a way that it is possible for a new discipline to evolve over time" (Fiore, 2008: 254). Adopting this system for textual analysis indeed denotes trans-disciplinarity as set out in the Mode 2 knowledge production framework, with a distinct problem-solving framework, new theoretical structures, and research methods or modes of practice to facilitate problem solving, (Gibbons et al. 1994). And this change involves what a discipline constitutes, basically "the body of concepts, methods, and fundamental aims... [and] a communal tradition of procedures and techniques for dealing with theoretical or practical problems" (Toulmin, 1972, p. 142). Using text-mining for textual analysis leads to trans-disciplinarity where "a shared, over-arching theoretical framework which welds components into a unit" exists (Rossini and Porter, 1979: 70). However, given the state of art of text-mining, this shift to trans-disciplinarity raises some methodological and managerial challenges.

The fundamental methodological challenge derived from trans-disciplinairy is: to what extent the theoretical and methodological framework is shared and by whom?

To develop a text-mining system requires not only textual analysts (social scientists) but also text-miners (computer scientists) to be on board. As of models and modelling in science, some hypotheses would be formed to be tested with some factors pre-assigned and pre-categorised. The algorithms in the text-mining system would have learned the specific knowledge (reading and interpretation) of specific domain experts who participated in the initial development, and analyse, organise and sort data out lexically and statistically. The intuitive human semantics are artificially programmed and inferred. Whoever wish to understand the newspaper texts will be relying on this specific sets of concepts and methods developed through a small team of computer and social scientists. Although it may be claimed that there are some benefits (e.g., processing large amounts of data within a very short time, increasing inter-coder reliability), this has not taken the interpretative flexibility of texts into account. The same texts can be looked at from different perspectives, through different means and frameworks. That said, others may not want to use these text-mining tools which were initially developed for a specific team of researchers who were investigating different research questions and bearing different agendas.

To avoid such conditions, text-mining tools for textual analysis would need to be situated in each individual research projects. And that denotes the kind of small-group “team science” that Fiore (2008) describes. This kind of “team science” is not to be confused with “Big Science” with large-scale networked computing infrastructures. The vision of “big science” is well presented in the current research policies and strategies that the research councils in Europe and North America have been making (Jankowski, 2009). This tendency of generalising methods and theoretical frameworks in arts and humanity as in natural science and engineering is not new. Rob Kling, for example, is one of those prominent scholars who constantly reminded developers of “field differences” and the shaping of electronic media in supporting scientific communication (e.g., Kling and McKim, 2000).

Based on the findings from the above case study, the text-mining system embodied mostly the an engineering-driven mindset. Had the system been available for wider adoption, the disciplines involved would all need to be integrated and re-conceptualised. However, the disciplinary boundaries in the studied project remained rigid. Without integration and re-conceptualisation of disciplines, the mutuality and interaction will remain superficial even if a shared platform or tool has been developed.

With such a technology-driven attitude, future arts and humanities are facing a risk of being instrumentalised – big linked datasets and (semi-)automated data analysis tools (such as the text-mining ones portrayed in this paper). This seemingly asymmetrical and asynchronous assemblage of artificial intelligence for knowledge mining and knowledge discovery only privileges the knowledge that holds by a specific group of experts. And the knowledge that is summarised, in the context of humanity, is not going to be widely shared. Inserting the perspectives and desires of those e-Scientists, notably from the scientific domains such as genetics, physics, biology, and clinical medicine, into humanity has caused uneasiness of domain experts, as Pieri (2009) writes “many social scientists and scholars in cognate disciplines remain apparently unaware or unimpressed by the promises of linking up large-scale data sets of fieldwork, and having access to the new tools and technologies that are being developed to cope with this scaling up of data set

size." (p. 1103). To balance this "inescapable imperative" (Kling and McKim, 2000: 1311) and avoid black-boxing (Bijker & Law, 1992) the e-Science technologieand techniques and exaggerating the expectations and applications, she calls for a discussion about the limitations and drawbacks of these e-Science infrastructures and tools, and to "explore the extent to which these values are shared across sections of the research community, or the extent to which they may be specific of certain stakeholders only". (Pieri, 2009: 1103) The needs for transparentising debates and for negotiation of values in research policy making is interconnected with the need for better communication, as raised by many (Fiore, 2008; Bammer, 2008). This leads to the managerial challenge that trans-disciplinarity brings.

As emphasised in existing literature on inter-disciplinarity, collaboration is a key to the success of such conglomeration (Fiore, 2008; Bammer, 2008). In 1979, Rossini & Porter proposed four strategies for integrating disciplinary components: common group learning, modelling, negotiation among experts, and integration by leader. More than three decades later, Bammer (2008) proposes three strategies for leading and managing research collaboration: 1) effectively harnessing differences; 2) setting defensible boundaries; 3) gaining legitimate authorization. Reviewing the case study against these suggestions, organizational learning for harnessing the differences, negotiation amongst team scientists all took place. However, despite the leadership from the Principle Investigator and his effort of energising the team from time to time, disciplinary integration appeared to be difficult and that hinders trans-disciplinarity. Although it has been acknowledged by text-miners that technical processes of data and text mining are highly iterative and complex (Kurgan and Musilek, 2006; Brachman & Anand, 1996; Fayyad *et al.*, 1996), text-miners have paid relatively little attention to the dynamics in the collaboration processes between interdisciplinary team members. In our experience, the domain users and text-miners found it difficult to communicate their own taken-for-granted background assumptions about the data and methods, and this was a marked hindrance to the project. To the domain users, the miners appeared instrument-oriented rather than user-centred. To the text-miners, the users appeared interfering by wanting more explanation about the operation of their tools and their criteria for preferring one algorithm over another. To some extent, the lack of open communication between domain users and text-miners worsened once problems were encountered. Positive results might have strengthened trust between the team members but early failure undermined it. This demonstrates that collaborative strategies are not incidental to interdisciplinary projects but central to their functioning.

Conclusions

Based on a case study of an inter-disciplinary project that gathered text-miner and textual analysts together to develop a text-mining system for analysing newspaper articles, this paper 1) examines how different disciplinary expertises were organised, integrated, jointed and disjointed at different stages of the development process 2) extends existing examination of inter-disciplinary practices specifically to the context of the digital humanities, and 3) discusses the methodological and managerial challenges emerging

from a seemingly shift towards trans-disciplinarity. Such a practice-based view echoes what Mattila (2005) argues that interdisciplinarity is “in the making” as in Latourian metaphor “science in the making” (Latour, 1987: 7). This case study has offered an episode that explores what had not been known yet - “which does not carry ready-made definition or categorizations” (Mattila, 2005: 533) about what text-mining can do for arts and humanities.

More than four decades ago, Rossini and Porter (1979) noted that “Interdisciplinary research lacks the collection of paradigmatic success stories which accompany nearly every disciplinary research tradition. Not only are specific strategies for integration lacking, but the notion of integration itself has not been well-articulated.” (p. 77) The case study has demonstrated that it is not straightforward to re-purpose text mining tools initially developed for biomedical research and customise them for arts and humanities. Nor should the software development effort be under-estimated. Nevertheless, the knowledge exchange and mutual sharing did take place between the text miners and social scientists to some extent.

As to the Turing Test, who won in the end? The aim of this paper is not to judge whether text mining enabled automatic coding is more efficient, or human manual coding. As this work symbolises the beginning of digital humanity, any conclusion would be premature as we had by no means exhausted the options available. But, at the moment, in light of the experiences of some social scientists who use computer-assisted qualitative data analysis (CAQDAS) tools (e.g. Seale *et al.*, 2006a; Seale *et al.*, 2006b), show that even if coding processes can be automated by computers, human intelligence would still be needed to make sense of the results based on their research questions. While nobody could say that computers can replace human intelligence, efforts will continue to seek ways of harnessing what computers are good at – in particular, processing huge amounts of data systematically – to support social science research advances that would not otherwise be possible. And this will be a long-term commitment of observing how this shift towards trans-disciplinarity in humanity transpires.

Acknowledgement

This writing is largely benefited from the insightful discussion I had with Prof. Peter Halfpenny, Elisa Pieri, and Dr. Mercedes Arguello-Casteleiro during the period when I worked at the coordinating hub of the ESRC National Centre for e-Social Science (NCeSS) at the University of Manchester from 2006 to 2009.

References

Bammer, G., 2008. Enhancing research collaboration: Three key management challenges. *Research Policy* 37(5): 875-887.

Barry, A., Born, G. and Weszkalnys, G., 2008. Logics of interdisciplinarity. *Economy and Society*, 37(1): 20-49.

Barthes, R., 1977. Image Music Text. London: Harper Collins.

Beaulieu, A., Scharnhorst, A., and Wouters, P., 2007. Not another case study: a middle-

range interrogation of ethnographic case studies in the exploration of e-science. *Science, Technology, & Human Values* 32(6), 672-692.

Bijker, W. & Law, J., (eds.) 1992. *Shaping technology / building society: studies in socio-technical change*. Cambridge, MA: The MIT Press.

Bolden, R. & Moscarola, J., 2000. Bridging the quantitative-qualitative divide – the lexical approach to textual data analysis. *Social Science Computer Review*, 18(4), 450-460.

Brachman, R. and Anand, T. (1996). The process of knowledge discovery in databases: a human-centered approach, in: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds) *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press., pp. 37-58.

Christian, B., 2011. Mind vs. Machine. The Atlantic Magazine. URL: http://www.theatlantic.com/magazine/archive/2011/03/mind-vs-machine/8386/ (retrieved on 23 Apr 2011).

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996. From data mining to knowledge discovery in databases. *AI Magazine* 17(3), 37-54.

Fiore, S. M., 2008. Interdisciplinarity as Teamwork: How the Science of Teams Can Inform Team Science'. *Small Group Research*, 39(3), 251-277.

Flinterman, J. F., Teclemariam-Mesbah, R., Broerse, J. E. W., Bunders, J. F. G., 2001. Transdisciplinarity: The New Challenge for Biomedical Research. *Bulletin of Science Technology Society* 21(4): 253-266.

Fry, J., 2003. *The cultural shaping of scholarly communication within academic specialisms*. Unpublished Ph.D. Thesis, University of Brighton.

Fry, J., 2004. The Cultural Shaping of ICTs within Academic Fields: Corpus-based Linguistics as a Case Study. *Literary and Linguistic Computing* 19(3): 303-319.

Fry, J., 2006. Scholarly Research and Information Practices: A Domain Analytic Approach., *Information Processing and Management*, 42(1), pp. 299-316.

Gibbons, M.., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., Trow, M., 1994. *The new production of knowledge: the dynamics of science and research in contemporary societies*. London: Sage.

Hessels, L. K. and van Lente, H., 2008. Re-thinking new knowledge production: A literature review and a research agenda. *Research Policy* 37: 740-760.

Huutoniemi, K., Klein, J. T., Bruun, H., & Hukkinen, J., 2010. Analyzing interdisciplinarity: Typology and indicators. *Research Policy*, *39*(1), 79-88.

Jankowski, N. (ed.) (2009). *E-Research - Transformation in Scholarly Practice*. New York: Routledge.

Klein, J. T., 1990. *Interdisciplinarity: History, Theory and Practice*. Detroit: Wayne State University.

Kling, R. and McKim, G., 2000. Not just a matter of time: field differences and the shaping of electronic media in supporting scientific communication. Journal of the American Society for Information Science, 51(14), 1306-1320.

Knorr-Cetina, K., 1982. Scientific communities of transepistemic arenas of research? A critique of quasi-economic models of science. Social Studies of Science 12, 101-130.

Koenig, T., 2004. Reframing frame analysis: systematizing the empirical identification of frames using qualitative data analysis software, Paper presented *at the annual meeting of the American Sociological Association, Hilton San Francisco & Renaissance Parc 55 Hotel, San Francisco, CA,*, Aug 14, 2004.

Koenig, T., 2006. Compounding mixed-methods problems in frame analysis through comparative research. *Qualitative Research*, 6(1), 61-76.

Kurgan, L. and Musilek, P., 2006. A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, 21(1), 1-24.

Latour, B., 1987. *Science in Action – How to Follow Scientists and Engineers Through Society*. Cambridge, MA: Harvard University Press.

Mattila, E., *2005.* Interdisciplinarity 'in the making': *modelling* infectious diseases. *Perspectives on science*, 13 (4): 531-553.

Morillo, F., Bordons, M. and Gómez, I., 2003. Interdisciplinarity in science: A tentative typology of disciplines and research areas. Journal of the American Society for Information Science and Technology, 54: 1237–1249.

Nasukawa, T. and Nagano, T., 2001. Text analysis and knowledge mining system. *IBM Systems Journal*, 40(4), 967 – 984.

Pieri, E., 2009. Sociology of expectation and the e-social science agenda. *Information Communication and* Society, 12(7), 1103-1118.

Rosenfield, P. L., 1992. The potential of transdisciplinary research for sustaining and extending linkages between the health and social sciences. *Social Science Med.*, 35(11), 1343-1357.

Rossini, F. A. and Porter, A. L., 1979. Frameworks for integrating interdisciplinary research. *Research Policy* 8(1), 70-79.

Schroeder, R. & Fry, J., 2007. Social Science Approaches to e-Science: Framing an Agenda", *JCMC*, 12(2), http://jcmc.indiana.edu/vol12/issue2/schroeder.html.

Schummer, J., 2004. Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. Scientometrics, 59(3): 425-465.

Seale C, Charteris-Black J, Ziebland S., 2006a. Gender, cancer experience and internet use: a comparative keyword analysis of interviews and online cancer support groups. *Social Science and Medicine.* 62(10), 2577-2590

Seale C, Anderson E, Kinnersley P., 2006b. Treatment advice in primary care: a comparative study of nurse practitioners and general practitioners. Journal of Advanced Nursing 54(5), 1-8.

Toulmin, S., 1972. *Human understanding: Vol. 1. The Collective Use and Development of Concepts*. Princeton, NJ: Princeton University Press.

Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H., Takeda, K., 2004. A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal* 43(3), 516 – 533.

Weingart, Peter and Nico Stehr (Eds), 2000. *Practicing Interdisciplinarity*. Toronto: University of Toronto Press.

Whitley, R., 1984. *The intellectual and social organization of the sciences.* Oxford, Clarendon Press.

Zheng, Y., Venters, W. and Cornford, T., 2011. Collective agility, paradox and organizational improvisation: the development of a particle physics grid. *Information Systems Journal*, 21(3), doi: 10.1111/j.1365-2575.2010.00360.x