Text and Data Mining in EThOS

Kiera McNeice, 2 October 2014

Contents

1 Introduction

1.1 EThOS and PhD Theses

A PhD thesis is an unusual kind of document. Its content, structure and format can vary dramatically – indeed it is not always a "document" in the traditional sense. Exempt from the Legal Deposit Libraries Act of 2003,[1] there is also some ambiguity about the status of a thesis as a "published" work. Theses are considered "grey literature", outside the realm of commercial publishing – they are not created with any expectations of direct material gain, but for less tangible rewards like a doctoral qualification, career advancement, and improving one's reputation as a researcher.

It has been estimated that as much as US$10 billion is spent funding doctoral research worldwide each year,[2] much of which comes from public sources. Often a PhD thesis is the only product of this research (other than the researcher themselves); there is therefore a strong incentive to ensure that any value within PhD theses is made discoverable and accessible to help drive research and innovation for the public good.

Historically theses have been made available mainly through copies held in libraries at awarding institutions. While in theory these theses are available to anyone who has access to the institutional library, in practice this means that they are not easily discoverable or accessible by the general public.

In recent years, parts of the higher education community have been moving towards collecting theses in electronic form and beginning to make theses available online. The British Library's Electronic Theses Online Service, EThOS, has played a key role in this process by curating a central record of UK PhD theses that anyone can access. Uptake of participation in EThOS has varied; as there is no mandate for deposition of theses in the British Library, theses are included on a voluntary basis only. The fact that theses are created without expectation of direct profit means that unlike in other areas thesis authors are not concerned with issues such as "lost sales" when making their work more freely accessible, but there are still a variety of reasons that researchers and other stakeholders in the higher education community are wary of the implications of making theses available online (see section 4).

Even so EThOS widely been recognised as a valuable service that benefits both researchers and institutions by making potentially valuable information more discoverable to researchers, and helping institutions both to preserve and raise the profile of theses created by their students. EThOS already contains over 360,000 records, including an estimated 80% of theses published in the last five years. It has been estimated that by 2015 over 99% of theses created in the UK will be collected in electronic form;[3] with over 20,000 doctorates now awarded in the UK each year (a number that has more than doubled since 1995[4]) EThOS can be expected to grow rapidly in coming years.

[1] Legal Deposit Libraries Act 2003 s 1(3)

[2] Researcher and advocate for open access and content mining, Peter Murray-Rust cites this figure as a "back-of-the-envelope estimate" of money spent on doctoral research worldwide. Although the figure itself is of dubious accuracy, it is probably about the right order of magnitude.

[3] Influencing the Deposit of Electronic Theses in UK HE (Brown, Sadler & Moyle, 2010). It is worth noting however that this estimation is based on aims stated by higher education institutions in 2010 to implement

If a doctoral thesis averages between 100 and 200 pages,[5] this represents a *vast* store of information that at this point is still significantly underutilised. One of the goals of EThOS is to ensure that this knowledge is made available for researchers and other interested parties to use to its full extent, to the benefit of researchers, higher education institutions, and all other vested interests.

Part of the reason for underutilisation of theses may be that they have not historically been easily discoverable or accessible, and researchers are still adapting to the idea of being able to easily download and read theses through EThOS. Scepticism about the value of theses as a resource may also have an influence: When asked whether they would ever refer to a PhD thesis for information, one post-doctoral biomedical researcher commented, "Anything useful in a thesis would have been published elsewhere already." This attitude is not uncommon; among many in the academic community theses are seen as a less valuable and less reliable information source than other publications like peer-reviewed journals.

However there are many scenarios in which the information contained in theses may never be reported elsewhere. For example many negative results that are included in theses will not be published in journals, and as PhD theses do not have the strict limitations on length of academic journals they often include more in-depth detail than other publications. In other cases researchers may simply leave academia with no intention of pursuing further publications after completing their thesis.

The growth in access to theses through EThOS[6] indicates a growing recognition of the value of theses as an information source – but thus far access to EThOS has been designed around the use case of an individual researcher accessing specific theses on an individual basis. The sheer potential of theses as a source of information is much broader than this: little has been done to explore the potential of theses as an *aggregate* source of information, or what new knowledge might be generated by analysing the set of UK theses as a whole. And it will be difficult to know just how vast that potential value is until and unless theses are made available to researchers to start to investigate and analyse in aggregate.

1.2 The National Compound Collection (NCC) Initiative

The National Compound Collection is an initiative of the Royal Society of Chemistry to create a national, comprehensive, searchable library of chemical compounds for the UK. Each record would include a synthetic method – a stepwise description of the reactions and procedures that can be followed to synthesise a chemical compound – so that researchers wishing to work with compounds found in the database would be able to recreate those molecules themselves.

PhD theses were identified as one possible source of molecules. In order to determine whether theses contained molecules that were novel and valuable, the RSC in collaboration with researchers from Bristol University planned a pilot project to extract molecular structures from theses, compare them to existing databases of molecules, and investigate how valuable a set of such molecules might be to researchers in academia or industry.

---

electronic deposit policies by 2015; the extent to which theses policies have since been carried out (and adhered to) in reality has not been assessed.

[4] Based on figures from the Higher Education Statistics Agency

[5] How long is the average dissertation? (Retrieved 2014-10-02)

[6] Currently 11,500 theses are accessed each month, up 19% in the last year

The pilot project began with a team of ten "data collectors" who worked with 15 universities around the UK to identify molecules in theses. Over four months they collected over 45,000 molecules that were entered into ChemSpider, a free online database.[7] The actual process of data collection generally involved data collectors approaching chemistry supervisors or departments, explaining the goals of the project, and asking if they could supply any potentially useful theses.

Data collectors would then turn to the experimental section of the thesis – where synthetic methods are usually reported – to look for suitable molecules. ChemDraw[8] was used to enter molecular structures into ChemSpider. Synthetic methods were not copied into the database; instead metadata explaining where to find the source thesis was attached to these new molecular structures (along with metadata about the thesis author, supervisor, year awarded, awarding institution, funding bodies, etc.) so that any researcher looking to investigate a particular molecule would know where to find the source thesis with the synthetic method.[9]

Searches of other databases of molecules found that up to 50% of molecules extracted from theses by the pilot project were novel – that is, they were not found in any existing database. Based on a small sample of theses used, lead data collector Laura Broad found that a further 20-30% of molecules were only reported online once – that is, published only once by the author of the thesis, with no evidence of further investigation.

One immediate benefit to making so many new molecules easily discoverable online through ChemSpider is to reduce the risk of duplicating research. This saves the research community time, money and resources that would otherwise be unwittingly spent on repeating work already reported in theses.

One other proposed method of demonstrating the value of a collection of novel molecules was to screen them *in silico* against biological targets – that is, to use automated processes to identify molecules that might make interesting biological targets. This process is discussed in further detail in section 2.2.3.

Data collectors reported that reactions from the academic community were generally positive – that most supervisors and other researchers were interested in supporting the NCC pilot project and positive about the idea of unlocking data in theses that was essentially undiscoverable in its current form. Only one or two were outright opposed to the idea; a larger minority were concerned about the implications of making parts of their work more easily accessible online, but agreed that the potential benefits of the project outweighed those concerns. (For more on community views about thesis access and re-use, see section 4.)

The RSC was therefore keen to investigate possible ways to scale up the pilot project and expand the collection of thesis-derived molecules in ChemSpider – and in particular, to learn whether EThOS records might be used as a source of more theses to work with. The methods used in the pilot

[7] ChemSpider is a free chemical structure database containing molecules, their properties, and other related information.
[8] ChemDraw is an industry standard molecule editor (software for creating and manipulating standard representations of molecular structures).
[9] An example record can be seen on ChemSpider here, showing all molecules that have been extracted from a particular thesis. (Retrieved 2014-10-02)

project to extract molecular structures would be prohibitively time-consuming and expensive at scale, so another goal was to investigate the possibility of automating the process to expand it to larger numbers of theses.

1.3 Text and Data Mining

One of the major challenges with a project like the NCC is the sheer volume of information to be analysed. EThOS currently contains records of over 360,000 theses published in the UK; to manually search through even the roughly 32,000 records currently identified as chemistry theses for molecular structures would be economically and practically prohibitive.

This is a problem that goes far beyond PhD theses. The exponential growth in publications is fast outstripping any individual researcher's ability to keep up with even a small selection of relevant publications. As far back as 1963, Derek de Solla Price calculated that the doubling time for scientific outputs such as number of journals and number of papers was fairly constant, on the order of 10-15 years.[10] More recent figures show that the number of research publications worldwide grew by 73% from 1,223,094 in 2001 to 2,121,740 in 2012.[11] There is a present and growing need for effective ways to analyse and extract important information from large datasets of information.

A class of processes known as "text and data mining" (alternatively "content mining" or "content analytics") is emerging to address this need. These umbrella terms refer to a wide variety of automated techniques for analysis and interpretation of information on extremely large scales, the potential benefits of which go far beyond science and even research as a whole. In both the commercial and public sectors more effective analysis and use of "big data" could generate huge savings and productivity gains. In 2011 the consulting firm McKinsey estimated that the US healthcare industry could create over $300 billion in value every year through better use of big data,[12] and an Expert Group reported to the European Commission in 2014 that text and data mining have the potential to add tens of billions of Euros to the aggregate GDP of the European Union.[13]

In the case of research, text and data mining activities can broadly be split into two categories: information extraction and generating new knowledge.

The NCC project is an example of information extraction: identifying and collecting data from multiple documents within a dataset. In this case the data to be collected (molecular structures) already exists in theses, and the main benefits of using text and data mining techniques are the time and cost savings associated with automating the process of extraction.[14]

By contrast, other text and data mining techniques can be used to generate new knowledge – information that does not exist in any of the data within the dataset, but is derived from analysing links and trends among the original data. As an analogy, information extraction can be compared to looking for clues within a case to solve a particular crime, while analysing trends across all crime

[10] de Solla Price, D. J. *Little Science, Big Science*. Columbia University Press, 1963.
[11] Response to the Public Consultation on the review of the EU copyright rules conducted by the European Commission, Directorate General Internal Market and Services (2014)
[12] Big data: The next frontier for innovation, competition, and productivity (Retrieved 2014-10-02)
[13] Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining: Report from the Expert Group (2014)
[14] Some purists would say this is not "true" content mining, but rather a new way of carrying out a traditional literature search.

statistics to identify which areas of a city are statistically more dangerous than others would be an example of generating new knowledge.

One example of new knowledge that can be generated by analysing information sources in aggregate is literature-based discoveries. These are relationships identified by finding and connecting concepts and results published in disparate parts of existing literature, an idea that was pioneered by scientist Don R. Swanson in the 1980s. Swanson identified a link between dietary fish oil and a circulatory disease known as Raynaud's Syndrome; it was later proved that the former could indeed be used to treat the latter.[15] Using automated processes to analyse and link data on larger scales opens up the potential for many more similar literature-based discoveries, as well as for answering questions that researchers have never previously thought to ask.

Even within the British Library, EThOS is just one example of many large datasets the Library holds and curates. Text and data mining techniques could be used not only to make the information within these datasets more accessible and useful to readers and researchers, but also to better curate the datasets themselves, for example by using mining processes to generate metadata (see section 5.1.4).

However barriers still exist to carrying out text and data mining on large scales. Often these are technical: infrastructure that impedes bulk access to information, lack of interoperability among information silos, information formats that make large-scale analysis difficult or costly, and other issues. However legal and ethical questions are also a concern, as text and data mining may involve copying protected works or exposing sensitive information.

1.4 Aims of This Report

The aim of this report is to comprehensively explain the potential technical, legal and ethical challenges to carrying out text and data mining on EThOS records, with specific reference to the NCC project as a case study.

Section 2 of this report will describe the technical processes of text and data mining, and the potential challenges therein. Section 3 will then go on to address the relevant legal concerns, and section 4 will discuss attitudes to thesis access and re-use within the higher education community. Finally section 5 will draw together these learnings and present a framework with which to consider any prospective text and data mining projects involving EThOS in future.

2 The Mining Process

Although text and data mining refers to a diverse class of processes, many key concepts and challenges are broadly applicable across the field of content mining and are discussed below.

2.1 Assembling a Dataset

A question that must be asked before beginning any text and data mining project is: Which data will be mined? Preparing documents for analysis can be a time-consuming and expensive process

[15] Swanson, D. R.; Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 1986, **30(1)**, 7-18

(see section 2.2), and limiting those efforts to only the most relevant and useful documents helps to minimise the costs associated with text and data mining.

The question of which data will be mined may of course be limited by which data are accessible. In the case of large-scale text and data mining operations accessibility refers not just to access to individual documents, but access to the dataset as a whole. That is, being able to access individual documents within a dataset is not sufficient – accessing hundreds or thousands of theses one at a time in order to collect a dataset for analysis would simply not be practical.

The EThOS database contains a mixture of record types including full text theses held by EThOS, links to full text theses held in institutional repositories, records of theses yet to be digitised, and some records of theses for which even digitisation is not an option. Bulk access is not possible for any of these records – not even for the full texts of theses that are held within EThOS. Each individual thesis may be accessed only after clicking through and agreeing to the EThOS Terms and Conditions for access. This presents an immediate challenge to any prospective text and data mining project intending to analyse theses on a large scale.

The next challenge is how to identify the data that are most relevant and useful in a way that maximises results, but minimises false positives.

In the case of the NCC project, the goal is to identify and extract novel molecular structures for which synthetic methods are reported. During the pilot project most theses were chosen manually by data collectors who approached individual chemistry PhD supervisors, described the goals of the project, and asked if the supervisors could suggest any theses which might be useful. Obviously this is not a method that would scale well.

Several options exist for carrying out large-scale, automated identification of useful records. In particular these include:

- **Keyword search**, where part or all of potential targets for mining are searched for instances of a specific word or group of words
- **Filtering by Metadata**, where metadata about potential targets for mining is used to subdivide them into more or less relevant categories
- **Content analytics** used to identify some relevant aspect of potential targets for mining

2.1.1 Keyword Search

Keyword search can be an effective way to identify references to a specific, known entity. Factors that affect the effectiveness of keyword search include the consistency of terminology used for the target entity, and the amount of text available for searching – performing keyword searches on the full text of articles increases the matches found by an order of magnitude or more when compared to searching abstracts alone.[16] While keyword searchers for more common words will return a greater number of false positives, in some cases it may be possible to narrow the parameters of the search to make searches more targeted. For example, many records within the

[16] Kostoff, R. N.; Expanded information retrieval using full-text searching, *J. Inf. Sci.* 2010, **36**, 104-113

Europe PMC corpus[17] can be searched for keywords within specific document sections such as abstracts, methods, or results.

In the case of the NCC project the targets for mining are a *class* of entities, the names of which are not known. In cases like this keyword search is essentially useless; it is difficult to conceive of a group of keywords that could be used to target this class effectively. Looking for "molecules" would return an extremely large number of false positives (records that would not be useful for this project). As the questions researchers ask of large datasets become more and more complex, keywords become less useful as a searching tool.

2.1.2 Filtering by Metadata

In some cases metadata may offer a better alternative to keyword search. For example, theses reporting novel molecules with synthetic methods generally fall into the subject of organic chemistry, and filtering by subject could therefore narrow down the corpus of records to analyse. However filtering by subjects recorded in metadata is also limited by the comprehensiveness and accuracy of the metadata available. Within EThOS around 10% of records lack a subject classification, and the subject classifications that exist have been assigned by various different methods that may not always be consistent.

A second issue with filtering by metadata is achieving the level of precision required. Using the example of subject classification again, "organic chemistry" is a fairly broad field and many organic chemistry theses cover topics that would not include novel molecules – for example, studies of the mechanisms of already-known reactions. In the case of the NCC, when an entire corpus of chemistry theses was examined only 10% were found to be useful for extracting molecular structures.[18] Conversely, there may be novel molecules in theses from subjects such as biology that would be excluded from theses identified in this way.

Filtering by narrower categories of metadata would help to reduce false positives. For example, theses written under the same supervisor generally come from the same research group and cover closely related topics of research. Searching for theses connected to supervisors whose work had already proved useful for the NCC would likely return a very relevant set of results – but filtering down to this level of granularity would require significantly more manual effort in identifying and curating a list of relevant supervisors. Furthermore, theses are even less likely to have more specific details like thesis supervisors recorded within the metadata.

These issues – incompleteness of metadata, inaccuracy of metadata, and the trade-off between broad categories with many false positives and narrow filters that require significant investment in curation – demonstrate that while filtering by metadata may be more effective than keyword search at sorting data, it still has significant shortcomings.

[17] Europe PubMed Central offers free access to a range of biomedical literature resources.

[18] As described in section 1.3, in most cases data collectors approached supervisors and directly asked for theses so the corpora they worked with were not representative of chemistry theses as a whole. However in one case St Andrews University offered data collectors access to all theses within the chemistry section of their online repository that were not subject to embargo. Of the 91 theses made available, 9 were found to contain molecules useful for the NCC (9.9%). This is the only data from the pilot project that gives an indication of what proportion of chemistry theses as a whole may be useful for the NCC project.

2.1.3 Content Analytics

An emerging solution to help categorise and search documents within datasets is to use algorithms to analyse content – essentially to use text and data mining techniques to facilitate text and data mining. Algorithms for content analytics have the potential to identify more complex concepts and targets within documents, and to more accurately discover sources of specific classes of information that may not be easily identified by keywords or metadata.

As with keyword search, access to the full text of a thesis is an advantage when employing algorithms to analyse content: more data to analyse generally leads to more accurate results. However useful results may be achieved even when access is limited to metadata.

Alongside the NCC, the RSC had previously asked EThOS for a set of theses that might be used to test a tool to classify how "chemistry-like" they were. By analysing information such as word frequency this tool was applied to the metadata of a set of around 30,000 EThOS records and assigned each a score to indicate its similarity to RSC publications. The top 7,000 to 10,000 ranked theses were considered the most likely to be chemistry theses, and the records were further classified by the RSC-like tool according to the RSC's twelve categories of chemistry publications.

A rough look at the effectiveness of this tool was carried out by comparing the results with the set of chemistry theses offered to the NCC pilot project by St Andrews.[18] Fifty-six of the St Andrews chemistry theses were among those analysed by the RSC-like tool, and all fifty-six were ranked within the top 7,000 RSC-like results. Data collectors extracted molecular structures from six of these (10.7%); of the theses used for extraction five were in the top 4,000 RSC-like results and classified as organic chemistry, while the remaining thesis was ranked around 6,500 and classified as biochemistry.[19]

Whereas only 10% of the entire set of St Andrews chemistry theses turned out to contain useful molecules for the NCC project, five of the thirteen St Andrews theses (38.5%) classified as organic chemistry by the RSC-like tool were used by data collectors. Although the sample size here is very small, this indicates that the RSC-like tool might be used to identify highly RSC-like organic chemistry theses as potential targets for the NCC project. This would cut the initial corpus of over 30,000 records analysed by the RSC-like tool down to around 800 primary targets for the NCC project – and the process is effective even without access to the full texts of theses for analysis.

Of course, one disadvantage of using content analytics to identify useful data is that the development, training and testing of algorithms can itself be a challenging and costly process. But as the field of text and data mining develops, no doubt so too will the availability of such tools.

2.2 Analysing Data

Once a dataset of relevant records has been assembled, the next challenge is developing software that can interpret the information within it.

Most electronic theses are stored in PDF format (or in the case of some digitised theses, image formats such as TIFF). The information in these files is considered *unstructured data*; that is, it is data without any semantic meaning attached to it and cannot be easily "understood" by computer

[19] It is worth noting that due to the nature of the tool, the less RSC-like a record is the less reliable its subject classification is likely to be.

software. Where an image of a line graph saved as a JPEG file is an example of unstructured data, the Excel spreadsheet used to generate that line graph is a structured representation of the same data – and only the latter can be easily manipulated and analysed by computer software.

Interpreting unstructured data is one of the major technical challenges of text and data mining. To illustrate the difficulty involved with interpreting a scanned copy of a thesis saved as an image, consider the cognitive steps required for a human being to read and understand a sentence printed on a piece of paper:

1. Visual information must be interpreted as representing strings of characters
2. Strings of characters must be grouped together into words
3. Semantic meanings must be assigned to words
4. Grammatical roles must be assigned to words using contextual information
5. Strings of words must be interpreted as signifying relationships between semantic meanings
6. A grammatical interpretation of the whole sentence must be parsed
7. Finally, this interpretation may be linked to other relevant information and concepts stored within memory

For a computer to interpret text saved in an image format it must have ways to emulate all of these processes – and similar challenges apply to interpreting other types of information such as images and graphs. While much progress has been made towards meeting these challenges, interpreting unstructured data is still very much a developing field.

2.2.1 Creating Structured Data

2.2.1.1 From Text

Converting a digitised image of text into structured data largely follows the same process as the steps described above.

Converting images of text into strings of characters is commonly referred to as Optical Character Recognition (OCR), and many tools have been developed for carrying out OCR with varying levels of success. It is now possible to automatically convert an image of English language text into strings of characters fairly accurately, although unusual characters used for example in chemical names or mathematical formulae can cause difficulty. Even as theses are increasingly deposited in electronic form in the UK, improvements in OCR technology will remain important for the digitisation of older theses recorded in the EThOS database.

Deriving semantic meaning and grammatical syntax from these strings of characters is the purview of a field of computer science known as Natural Language Processing (NLP). NLP can be used to generate a copy of the original text which is tagged (annotated) to indicate semantics and syntax; this structured data can then be used for analysis and comparisons with other data to generate new knowledge. Although a wide range of tools and approaches to NLP exist, the general process involves:

1. **Tokenising** or splitting the strings of characters into meaningful chunks (e.g. words)
2. Identifying **semantic entities** (e.g. by comparison with domain-specific dictionaries)
3. Assigning **grammatical roles** to tokens and entities

4. Generating an **annotated** copy of the original document (structured data)

Tools to perform NLP generally need to be domain-specific to function effectively, as conventions for the use of grammar, vocabulary and even punctuation can vary significantly among different fields. In most cases, for example, a comma can be used as an easy identifier as the end of a token (word) – but in chemistry commas can be used in the middle of chemical names. This is just one example of why most tools would need to be developed or adjusted for specific domains.

Identifying semantic entities is usually achieved by comparing tokens (words) to existing dictionaries. The process is imperfect – even when expert humans are asked to identify or index chemical entities within a text agreement is around 91%, which can be considered an upper bound for accuracy.[20] Still, in the field of chemistry tools such as ChemTagger[21] have made significant progress towards accurate tagging of chemical entities – and new tools are being developed that approach human levels of agreement for locating (87.39%) and indexing (88.20%) chemical entities.[20]

Such processes generally rely on building and curating extensive and accurate dictionaries of words and synonyms, a time-consuming process. In the fields of chemistry and biology some such dictionaries are already available for use, such as ChEBI[22] (Chemical Entities of Biological Interest). However these tools still have their limitations. "Pyridine" for example can refer either to the molecule known as pyridine, or more generally to any molecule which has pyridine as part of its substructure, a distinction that even the most extensive dictionaries would have trouble making.

2.2.1.2 From Images

In the case of images existing tools are even more narrowly specialised, and are generally most effective on images that follow a standardised structure. For example some progress has been made in generating structured data from phylogenetic trees[23] and simple line graphs.[24]

In the case of the NCC pilot project, software known as CLiDE (Chemical Literature Data Extraction)[25] was considered as a possible way to identify and interpret standard drawings of molecular structures. CLiDE is advertised as the "chemistry intelligent equivalent of OCR", and the pilot project team requested and obtained a temporary license to assess CLiDE's potential for automating the data extraction process.

Although CLiDE was able to identify molecular structural drawings well, it was found that the software had several limitations. These ranged from simple failures of optical interpretation (e.g. mistaking double bonds for single bonds in some cases) to more complex issues. For example, in some cases a thesis would report a range of molecules all based on the same structure, differing only in one small part. Rather than draw out each small variation separately, the basic structure would be drawn with a placeholder indicating the part of the molecule that varied, and the different

[20] Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 2 (Krallinger et al., 2013)
[21] ChemTagger is a name-tagger for chemical and other entities.
[22] ChEBI is a freely available dictionary of molecular entities focused on 'small' chemical compounds.
[23] Content Mining: we can now mine images (of phylogenetic trees and more) (Retrieved 2014-10-02)
[24] Content Mining; Extracting Facts from Plots – 2; we find errors in the paper (Retrieved 2014-10-02)
[25] CLiDE is developed by SymBioSis Inc.

variations would be listed separately. CLiDE currently lacks the sophistication to interpret such situations.

Overall it was found that the amount of time necessary to manually assess and connect results obtained from CLiDE meant that using the automated tool did not save time overall. However it was suggested that something similar to CLiDE may be used simply to *find* drawings of molecular structures within a thesis and indicate their location, potentially saving manual data collectors some time.

2.2.2 Interpreting and Linking Data

After semantic entities have been identified and tagged, algorithms can be employed to interpret relationships among those entities as well as relationships with other sources. In text, interpreting the relationships among words requires an understanding of grammar and syntax – another facet of natural language processing.

Chemical synthetic methods make particularly promising targets for attempts to parse grammatical structure as their formal, standardised grammar helps to reduce ambiguity. Tools already exist for making fair sense of chemical synthetic methods – for example OSCAR (Open Source Chemical Analysis Routines)[26] is an open-source tool for semantic annotation of chemistry papers. OSCAR is able to take a paragraph of text describing a synthetic method and identify not only the chemical entities within it, but things such as quantities relating to those entities, verbs describing actions carried out in a synthetic procedure, and time periods over which actions were carried out.

But the most exciting potential benefits of text and data mining – generating new knowledge from large datasets – require not only interpreting relationships within a document, but identifying new relationships *among* different documents within a dataset. This requires the generation of "linked data" – that is, entities and relationships identified within a text must be related to entities and relationships in other texts, and to broader, external concepts.

This can be done by making use of existing structured, linked databases which include information about connections among related concepts and entities. One such database is DBpedia,[27] which is essentially based on Wikipedia and the linked data that already exists within it (i.e. the links from one article to another).

The British Library is currently collaborating on a project called EnviLOD (Environmental Linked Open Data),[28] which aims to demonstrate the value of linked data for enhancing information discovery. Working with the Envia corpus,[29] tools have been generated to interpret, annotate and link information using DBPedia. The end result allows researchers to search, for example, for information specifically about floods in the last ten years which affected cities with a population greater than 100,000 in the UK.

[26] OSCAR is software for the semantic annotation of chemistry papers.
[27] DBpedia contains structured information based on Wikipedia.
[28] EnviLOD is a collaboration involving the British Library, Jisc, The University of Sheffield and HR Wallingford.
[29] Envia is a collection of "grey literature" related to environmental science, particularly information about floods, including governmental reports, PhD theses, and other data resources.

Such queries would be impossible with traditional methods of searching, and the potential value of being able to carry out research in this way is huge. However several of the problems encountered in the creation of EnviLOD are typical of the challenges of generating value from linked data.

Firstly there is the problem of irrelevant connections. In the case of EnviLOD, basing links between data on DBpedia generated a large number of connections to what turned out to be rock bands. These rock bands had used various words derived from environmental concepts and phenomena in their names, which meant that working with DBpedia, EnviLOD erroneously identified them as relevant to environmental concepts.

The second problem EnviLOD faced was that although the potential for asking questions of linked data is huge, the process of *generating* queries requires an understanding of complex "query languages" to specify all the terms and parameters of a search. Wikipedia gives the following example of SPARQL code used to return people's names and emails from within a dataset:[30]

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?email
WHERE {
  ?person a foaf:Person.
  ?person foaf:name ?name.
  ?person foaf:mbox ?email.
}
```

The majority of researchers are not familiar with this kind of code, which constitutes a significant hurdle to making use of linked data. However EnviLOD and others are working on generating new smart graphical user interfaces to simplify the query process from the end-user's perspective.

2.2.3 High-Level Complexities

As well as the specific limitations of various tools and processes used to carry out various aspects of text and data mining, there are often higher-level challenges that at this point seem far beyond the capabilities of any automated process to understand.

One example is the issue of near-useful data, which might be recognised by an expert human as potentially valuable but not by a computer (nor indeed by a layperson). For example, as mentioned in section 1.2 one of the NCC's proposed ways of demonstrating value of novel collections of molecules would be to carry out *in silico* screening of molecules against biological targets such as proteins.

Proteins are complex biological structures with various "binding sites" – parts of the molecule which because of their specific structure and reactivity will allow other molecules of specific shapes and reactivities to bind with them, like fitting a key into a lock. Small molecules that can bind to proteins in this way are potentially useful biological targets, and computational methods exist for predicting whether small molecules will be able to bind with known protein structures.

[30] http://en.wikipedia.org/wiki/SPARQL (Retrieved 2014-10-02)

However, binding in this way requires small molecules to be chemically reactive – and in chemistry theses, often molecules are reported in "protected" forms. In order to carry out very specific reactions, chemists will often add unreactive "protecting groups" onto one part of the molecule, to ensure that only *other* parts of the molecule will react. These protecting groups can easily be removed by other chemical reactions; a trained chemist would be able to identify protected parts of a molecule as being potentially chemically reactive. However a computer would require an incredibly sophisticated understanding of chemistry and chemical reactions to recognise the same potential.

Although this example is specific to one very particular application of data extracted by text and data mining, problems of this kind – complexities that require even humans to have expert knowledge of a field to understand – are likely to exist in many other areas.

3 Mining and Intellectual Property Rights

Several aspects of text and data mining processes require considerations of the intellectual property rights that apply to the content being analysed. New knowledge created by text and data mining projects are considered the work of their creator, meaning that a researcher carrying out text and data mining would generally be the first owner of any rights their new work attracts.[31] However in some cases the process of generating that new work may involve potentially infringing activities. This section discusses the intellectual property rights that may be relevant to text and data mining of theses.

3.1 Copyright

Copyright is an automatic protection granted for creative works. The precise details of what constitutes a "creative work" are decided by courts on a case-by-case basis, but in written works copyright has been found to apply even to creative expressions as short as a newspaper headline.[32] Where a work is protected by copyright it is an infringement to copy part or all that work without the permission of the copyright holder unless exceptions apply.

Text and data mining tools and processes may entail potentially infringing activities in two ways:

- **Preparing data for analysis** may involve digitisation (creating a digital copy) or annotation (creating an annotated copy) of the works to be analysed.
- **Publishing outputs of analysis** may involve reproducing excerpts copied from the original works being analysed.

Text and data mining projects must therefore consider whether the specific processes they intend to use involve making or reproducing copies of copyrighted content within the original work, and if so, whether permission from copyright holders must be sought.

3.1.1 Exclusions from Copyright

Copyright applies only to fixed creative works, and as such not all content within a thesis is protected by copyright. In particular ideas, know-how such as procedural knowledge, and scientific

[31] The only exception to this is when work is created by an employee in service of their employer (Copyright, Designs and Patents Act 1988 s 11(2)).
[32] NLA v Meltwater (Court of Appeal) Neutral Citation Number: [2011] EWCA Civ 890

facts and data do not attract copyright. For text and data mining projects this means that anything falling into one of these categories can be copied from a thesis and re-published without restriction.

3.1.1.1 Ideas and Knowledge

The precise distinction between an idea and a creative expression of that idea is difficult to define and is decided by courts on a case-by-case basis. Using the NCC project as an example, one interesting consideration would be whether the text describing a chemical synthetic method would attract copyright – whether the particular wording of that synthetic method could be considered a creative expression. If this were the case then re-expressing the information contained within the synthetic method in another form would still be non-infringing, as the procedural knowledge contained within that synthetic method does not attract copyright. However copying the actual *text* of that synthetic method would be a potential infringement.

It is questionable whether synthetic methods would attract copyright, as the process described by the synthetic method is knowledge that does not attract copyright and synthetic methods are described with formalised and fairly universal standards. But as this case has not been explicitly tested in court it is still an example of something that could potentially be challenged.

3.1.1.2 Facts and Data

Facts and raw data are not considered creative works and therefore do not attract copyright. Again the precise extent of what is considered a fact is decided by courts on a case by case basis, but examples of facts include mathematical relationships and molecular structures. For text and data mining projects this means this kind of content may be copied from a thesis and re-published without restriction.

In the NCC's pilot project, extraction of molecular structures from theses entailed data collectors looking at structural drawings in theses then re-drawing representations of the same structures using ChemDraw software. That is, the only content copied from the thesis was the factual molecular structure which does not attract copyright. This process therefore does not infringe on any copyright held within the thesis, and the molecular structural data extracted may be re-published without limitation.

3.1.2 Exceptions to Copyright

Several specific exceptions to copyright allow for copies to be made without permission that would otherwise infringe on copyright. Of particular relevance to text and data mining of theses are the exception for non-commercial research or private study, and the recently-implemented exceptions for non-commercial text and data mining and for quotation and parody.

3.1.2.1 Exception for Non-Commercial Research and Private Study

This exception allows for "fair dealing" with a work for the purpose of private study, or for non-commercial research so long as use of the work is accompanied by a sufficient acknowledgement. The precise definition of fair dealing with a given work is decided by courts on a case by case basis, and takes into account both objective facts and subjective judgements about the degree and intent of use of the work. The Intellectual Property Office's guidance about fair dealing describes

considerations of fair dealing as depending on the answer to the question, "How would a fair-minded and honest person have dealt with the work?"[33]

In practice this generally means that for a copy to be considered "fair dealing" it must be:

- Limited to a **part** of the work that is reasonable, appropriate and necessary for the intended use (usually only copies of a small part of the entire work will be considered fair dealing), and
- Not likely to affect the **market** of the original work, or act as a **substitute** for it.

This exception's limitation to fair dealing with a work has significant implications for projects intending to use theses for text and data mining. In particular, the restriction to copying only a (usually small) *part* of the work excludes the possibility of digitising or annotating an entire work for analysis.

This means that in most cases, text and data mining projects will not be able to rely upon the exception for non-commercial research and private study to carry out analysis of theses. Furthermore, as this exception applies only to personal use it may not be used to re-publish excerpts of data copied from the original works being analysed.

3.1.2.2 Exception for Non-Commercial Text and Data Mining

In June 2014 new UK legislation came into force permitting certain acts of copying that would otherwise constitute copyright infringement.[34] Specifically, this exception allows copies of copyrighted works to be made for "computational analysis of anything recorded in the work". The key aspect of the text and data mining exception is that it allows for copying of an *entire* work, not just for fair dealing with the work.

This exception is subject to the following limitations:

- The person making the copy must have **lawful access** to the work,
- The purpose of the text and data mining being carried out must be **non-commercial research**,
- The copy must be accompanied by **sufficient acknowledgement**, where practical, and
- The person making the copy may not circumvent any **technical protection measures**[35] that restrict access to the work.

In particular the first two limitations could have significant implications for projects intending to use PhD theses for text and data mining.

3.1.2.2.1 Limitation to Lawful Access to Works

The ways in which theses may be lawfully accessed vary significantly among (and sometimes even within) awarding institutions.

[33] Fair dealing (Retrieved 2014-10-02)
[34] Copyright, Designs and Patents Act 1988 s 29A
[35] Examples of "technical protection measures" include region locks, digital rights management tools, and restricting the volume of traffic to content to avoid undue server load.

In the past, theses were generally only accessible as physical publications stored in institutional libraries. In recent years many awarding institutions have been moving towards collecting theses in electronic form and making them available to access online, either through EThOS or through other online repositories. However policies and practices regarding access to theses vary dramatically from institution to institution.

In some cases, copyright holders sign clear licensing agreements which grant explicit rights to institutions to make theses freely accessible via online repositories. Online access to these theses is lawful as explicit permissions have been granted by the copyright holders, and copies of these theses may be made for non-commercial text and data mining without reservation under the text and data mining exception.

In other cases, institutions have adopted a risk-management approach to online thesis access in which theses are made available online *without* obtaining explicit permissions or licences from copyright holders. This is more likely to be the case where older, paper theses have been digitised in bulk as in most cases the copyright holder of a thesis is the author of that thesis, and obtaining permission from individual authors of every thesis to be digitised may be considered impractical or impossible.[36]

Making theses accessible in this way is considered by many institutions to be a low-risk approach (see section 4.1.1 for discussion of stakeholder attitudes and concerns). However digital copies of theses made without permission from copyright holders are not strictly lawful copies, which may mean that accessing such theses to make copies for non-commercial text and data mining may not be permitted under the text and data mining exception.

3.1.2.2.2 Limitation to Non-Commercial Research

There is no explicit set of rules in copyright law delimiting precisely which purposes are considered commercial; such cases are decided by courts on a case-by-case basis. A key factor for assessing whether a purpose can be considered commercial is the *intent* of the party involved, but this still leaves room for considerable uncertainty and the definition of "commerical" has been interpreted broadly in other areas of copyright law. In the case of the new text and data mining exception to copyright this uncertainty is exacerbated by the fact that the exception has only recently come into force and therefore has not been thoroughly tested in case law.

In the case of the NCC project, or indeed any large-scale text and data mining project, costs are expected to be considerable. In order to pursue the goals of the NCC project on a larger scale it is the intention of the RSC to develop a business model that would monetise the project's outputs and make it self-sustaining, if not necessarily profitable. It is unclear whether this kind of business plan would constitute commercial intent in the context of the text and data mining exception; certainly in

[36] Jill Russell, Head of Digital Assets at the University of Birmingham, reported that when attempting to contact authors for permission to digitise older theses the university did not have current contact details to up to two thirds of authors. A further two thirds of those contacted never responded, resulting in explicit consent being obtained for only about 10% of theses digitised.

other areas of copyright law concern has been expressed that limiting uses to "non-commercial" purposes effectively rules out the option of monetising outputs even to recoup costs.[37]

This could constitute a significant barrier to text and data mining projects, as without the certainty that a project's purpose would legally be considered non-commercial, it may not rely upon the text and data mining exception to copyright. Parties intending to monetise text and data mining projects may need to seek specific legal advice to understand whether permission from copyright holders is necessary.

3.1.2.3 Exception for Quotation

In October 2014 new UK legislation came into force permitting certain acts of copying that would otherwise constitute copyright infringement.[38] This exception allows for quotation from a work with the following limitations:

- The work must be **available to the public**,
- The use of the quotation must be **fair dealing**,
- Only as much of the original work as is **required for a specific purpose** may be quoted, and
- The quote must be accompanied by **sufficient acknowledgement**, where practical.

Like the exception for non-commercial research and private study discussed in section 3.1.2.1, the exception for quotation is limited to fair dealing with the original work. This precludes text and data mining projects from relying on the quotation exception to make digital or annotated copies of an entire work for analysis.

However the quotation exception may prove useful to text and data mining projects wishing to publish excerpts from original works as part of the outputs of analysis, provided those excerpts are fair dealing with the works in question.

3.2 Database Rights

Database rights are distinct from the rights that apply to each item within a database, and automatically apply to any public database in which there has been a substantial investment in obtaining, verification or presentation of its contents. The precise details of what constitutes "substantial investment" are decided in courts on a case-by-case basis.

The owner of the database has the exclusive right to extraction (permanent or temporary transfer of all or a substantial part of the database to another medium) and re-utilisation (making all or a substantial part of the database available to the public) of the contents of a database. Text and data mining projects that copy a large part of an existing database to analyse may therefore run the risk of infringing database rights (exceptions apply for non-commercial teaching and research[39]).

However the higher education institutions consulted during the course of this project did not express any intention to enforce their database rights to restrict use of their repositories; the very purpose of electronic institutional repositories is to make their content more widely accessible. In all

[37] With respect to digitisation of orphaned works, the Imperial War Museums was concerned that if limited to non-commercial uses of such works they would be "...unable to recoup any of the digitisation costs from licensing." (Retrieved 2014-10-02

[38] The Copyright and Rights in Performances (Quotation and Parody) Regulations 2014 (Retrieve 2014-10-02)

[39] Copyright and Rights in Database Regulations 1997 (SI 1997/3032) reg 20

cases institutional concerns with respect to text and data mining were about the rights that applied to theses *within* repositories, and respecting the wishes of the copyright holders and other stakeholders in those individual records. It seems unlikely that database rights would become an area of concern in the case of text and data mining of theses.

3.3 Risk Management

Although it is of course important to understand which text and data mining activities may be carried out in a given situation, in a practical sense the *risks* associated with infringing intellectual property rights depend entirely on the likelihood of that infringement being challenged. As only a rights holder (or exclusive licence holder) may bring a claim against an infringing party, some prospective text and data mining projects may consider moving ahead even in the face of legal uncertainty, if they believe that the risk of any rights holder bringing a claim of infringement is low. Considerations of the views of stakeholders in the creation of theses (section 4) are crucial to assessing this risk.

In the case of the British Library however, it is in the Library's best interests to avoid potentially unlawful uses of theses, and to avoid the appearance of supporting or condoning any such unlawful uses.

4 Community Views

From section 3.1 it is clear that at this point there is significant uncertainty surrounding what uses of theses may be lawful without the express permission of copyright holders. It may become necessary for some text and data mining projects to obtain permission from copyright holders for their work to be lawful.

The creator of a work is the first owner of that work's copyright with one exception: when works are created by an employee in the service of their employer. As PhD candidates are not considered employees in service of their institutions,[40] the first owner of any new content created in a thesis is the **author** of that thesis. Higher education institutions share the view of EThOS[41] that "Copyright in a thesis generally belongs to the author."[42]

So long as the author retains ownership of their work, their permission must be granted for any copying of their work that is not excluded or excepted from copyright to be lawful. For large-scale text and data mining activities, obtaining permission from the author of each and every thesis analysed is in for all practical purposes an impossible task – even if contact details are available for all authors, the sheer logistics would be beyond the resources of many proposed projects.

However as mentioned in section 3.1.2.2.1, in recent years some institutions have developed explicit licensing agreements with PhD students that permit access to and re-use of their theses. For example, permission has been given for all theses accessible through the University of Birmingham's

[40] It is *possible* that where PhD candidate's work is funded by a commercial sponsor the terms of that sponsorship may characterise the candidate as an employee and claim ownership of all copyright generated during the course of a PhD, including the thesis itself – but these are exceptionally rare cases.
[41] Several higher education institutions around the UK were contacted for their views on thesis rights and access, and without exception agreed that authors own the copyright to all work created in their theses.
[42] EThOS Toolkit (Retrieved 2014-10-02)

repository to be freely accessed and re-used for any non-commercial purpose.[43] As obtaining individual permissions is impractical, policies like these help to reduce the uncertainty associated with access to and re-use of theses, opening them up for large-scale text and data mining projects.

The policies of higher education institutions are influenced by the aims and concerns of many different stakeholders in the higher education community. These relationships among stakeholders in production, use and re-use of theses and their content can be complex, and their attitudes towards greater thesis availability varies. An understanding of these relationships and attitudes is essential to any consideration of the benefits and challenges of using theses for text and data mining for three key reasons:

- The issues of access to and re-use of PhD theses for text and data mining cannot simply be reduced to a legal or technical problem. Community concerns play a large part in driving (or hindering) progression towards greater availability of thesis content.
- The British Library is committed to respecting not only the legal rights but also the ethical, moral, and personal concerns of members of the research and higher education communities.
- Given the uncertainty still surrounding what uses of theses may be lawful without the express permission of copyright holders, understanding the risks associated with mining theses is impossible without considering the likelihood that rights holders may bring a claim against any perceived infringement.

Institutions have such a wide variety of approaches to managing rights within theses that it would be impractical to detail the entire range of possible rights workflows involved in thesis production. Instead of detailing specific processes, section 4.1.1 qualitatively addresses the various concerns that stakeholders may have so as to give an overview of the forces within the higher education and research communities that may facilitate or hinder greater thesis availability.

4.1.1 Increasing Thesis Availability

Although the benefits of greater access to research outputs are difficult to quantify, generally institutions and researchers are of the opinion that raising the visibility of their work and a freer flow of information can bring significant benefits. Furthermore several funding bodies now mandate that publications resulting from research they fund be published “open access” (free to access and re-use). Although this mandate has not yet been explicitly extended to theses – which as mentioned in section 1.1 are not always considered “published” content – this constitutes a strong incentive for institutions to help early career researchers establish good practices with regard to managing and making available their research outputs.

There is however some resistance within the higher education and research communities to making theses more widely available for access and re-use. The reasons for this resistance are varied, and involve complex and often interrelated concerns and perceptions. A visual representation of the aspects of thesis content that stakeholders may have vested interests in protecting from greater availability is given in Table 1 and this information is discussed in more detail below.

[43] eTheses Repository Policies (Retrieve 2014-10-02)

Table 1: Stakeholders in Thesis Content

Content			Examples	Stakeholders
Thesis	Copyrighted content	New work	Work created by the thesis author	• Authors • Institutions • Publishers
		Existing work	Work created by third parties included in thesis	• Third party rights holders • Authors • Institutions
	Other content (not covered by copyright)	Exploitable content	Commecially valuable knowledge and ideas	• Authors • Collaborators • Commercial sponsors • Tech Transfer • Supervisors • Publishers
		Sensitive content	Private or politically sensitive information	• Authors • Research subjects

Although some of the concerns discussed below are more relevant to some fields than others, it is worth noting that it is much easier from the perspective of higher education institutions to have a single uniform policy regarding access to and re-use of theses. So in some cases the concerns of researchers in one field may end up having a direct influence on re-use policies that apply to another field through an institution-wide policy.

4.1.1.1 Copyrighted Content

Only the owners of copyrighted content have the authority to allow re-use of that content, so the attitudes of copyright owners are fundamentally important to greater availability of theses. A thesis may include two kinds of copyrighted content.

1. **New content** created by the author in the course of producing their thesis

As discussed in section 3.1, the thesis author is the first owner of any copyright attracted by new work within a thesis and remains the copyright holder unless they choose to transfer their rights to a third party. This is an extremely rare occurrence; generally an author will choose instead to license various uses of their work to third parties.

In particular, institutions will generally require a license from thesis authors to permit them at the very least to preserve a copy of the thesis. As discussed in section 3.1.2.2.1 these agreements vary from institution to institution. Most institutional repository managers support moving towards increased permissions in these agreements and encourage students to allow greater access to their theses. Institutions generally believe that making theses as accessible as possible will help raise the profile of research at an institution, benefiting the institution's reputation.

That said, institutions are wary of alienating members of their own communities by pushing towards greater thesis accessibility without the enthusiastic involvement of students and staff, and

therefore prefer to convince researchers of the benefits of greater thesis accessibility in an inclusive way before moving to change any existing policies.

Publishers may on occasion be willing to publish all or part of a thesis as-is, and would require a license to do so. However generally publishers would expect any work based on a thesis to be significantly modified for publication, which means further publications then become a question of exploiting the ideas underpinning the work (see section 4.1.1.2) rather than the creative expression in the thesis itself. Furthermore, as publishers hope to make a profit off the work they publish they are generally not at all interested in promoting greater access to theses with their licensing agreements.

2. **Existing content** with copyright held by third parties

In some cases an author will reproduce content from previously existing works within their thesis; this content may contain copyright belonging to third parties. In theory the author should obtain all necessary permissions to re-publish third party content within their thesis before it is submitted to their institution. However in some cases this may be difficult, impractical, and extremely time-consuming, and in practice some authors simply do not bother obtaining permissions to re-use the third party content in their theses. A typical example would be a thesis in the field of art history, which may reproduce many images of artwork for discussion and analysis.

Historically theses have only been accessible as physical copies held in awarding institutions, so the chances of a third party discovering the re-use of their content and bringing a claim of infringement against the author has been very low. However greater access to theses means that in cases where third party content has been used without permission, authors and institutions are exposed to greater risk of claims of copyright infringement.

The obvious way to avoid this risk is to encourage best practices with respect to re-use of third party content. Institutions consider educating students and other researchers to be the best way to achieve this, and institutions that accept electronic deposition of theses are twice as likely to have formulated policies regarding third party copyright.[44] That is, policies for dealing with the issue of third party content are naturally being developed alongside movements towards greater thesis availability.

4.1.1.2 Other Content

As discussed in section 3.1.1, not all aspects of a thesis attract copyright protection. Furthermore although only copyright holders have the legal authority to allow re-use of their content, other stakeholders play an important role in influencing thesis authors' views about greater availability of their theses. It is therefore important to consider the vested interests in other kind of content within a thesis, and why authors and other stakeholders may wish to restrict access to such content.

1. **Exploitable content** that is potentially valuable to stakeholders

There are many ways in which content not protected by copyright may nonetheless have significant value to stakeholders in the creation of theses. Making theses more available can have an impact on stakeholders' ability to exploit that content – indeed several institutions reported that for

[44] Influencing the Deposit of Electronic Theses in UK HE (Brown, Sadler & Moyle, 2010)

this reason, commercial sponsors of PhD research and theses are the most opposed to increasing availability of theses, demanding embargos for as long as 20 years after a thesis is created.

One salient example is the case of patents. A key consideration in the assessment of a patent application is whether or not the knowledge or invention described in the patent has been made publicly available. Stakeholders such as commercial funders, collaborators, commercialisation or technology transfer departments within institutions, and of course the thesis authors themselves therefore have a compelling reason to restrict access to any potentially patentable information within a thesis until a patent application has been made. For this reason many stakeholders – in particular commercial sponsors – insist upon temporary or indefinite embargos on access to thesis content.

This was indeed one of the concerns raised by institutions asked to participate in the NCC pilot project. Commercialisation and technology transfer departments at some institutions expressed concern that by allowing access to theses for this or similar projects, they may be inadvertently releasing potentially valuable intellectual property. In this case data collectors rightly made the point that depositing a thesis in an institutional library is already considered making the information public – in fact, simply sharing the content of a thesis with a single thesis advisor may be considered making knowledge public unless there is a clear agreement of confidentiality.[45] Therefore in this context there is no reason to be any *more* concerned about making theses accessible online rather than in an institutional library – but this is a fact many researchers and institutions may not have considered closely.

A second example is the case of prior publication, particularly of monographs. In several fields the publication of a monograph is considered crucial to academic advancement, and many thesis authors go on to publish monographs based heavily on the contents of their theses. There is widespread concern among researchers that making theses widely available decreases the value of their content in the eyes of publishers, and may therefore impact a researcher's ability to publish a monograph based on a thesis. Even in scientific fields, data collectors in the NCC pilot reported that some chemistry researchers worried that if their novel molecules could be easily found through ChemSpider, their work might not be considered sufficiently novel for publication in some journals.

Although many researchers perpetuate the belief that publishers will not agree to publish work based on an easily accessible thesis, very few efforts have been made to investigate how much truth underpins this view. A survey of publishers carried out in 2011 in the USA found that the vast majority would still consider publishing work based on an open access thesis[46] – although one criticism of this study is that the formal policies of these publishers are not necessarily reflected in the decisions made by individual editors. Perhaps a more telling finding is that in UCL's 2010 survey of 144 higher education institutions, only a single unsubstantiated instance of a researcher being denied a publishing deal as a result of the availability of their thesis was reported.[47]

Assuming however that there is some risk for authors wishing to publish monographs, several institutions address this risk by educating authors about their options with respect to publication.

[45] http://www.ipo.gov.uk/o05308.pdf
[46] An Investigation of ETDs as Prior Publications: Findings from the 2011 NDLTD Publishers' Survey (Retrieved 2014-10-02)
[47] E-theses Best Practice Summaries: Impact on Future Publication (Retrieved 2014-10-02)

Chris Awre, Head of Information Management at the University of Hull reported that many early career researchers are not even aware that they have the option of negotiating the details of publishing deals. Like other institutions, the University of Hull believes that educating researchers about their options – including which publishers are more accepting of work based on open access theses – is the best way to allay fears associated with greater availability of theses.

Furthermore, in some cases publishers are explicitly supportive of greater availability of theses. Jill Russell, Head of Digital Assets at the University of Birmingham reported that some publishers will even include links on their websites back to the thesis on which a publication is based, and the survey of publishers in the USA found that some publishers consider interest in an open access thesis a good indication that there is a market for publications based on that thesis.[46]

A third example of exploitable work within a thesis is the case of ongoing research. In larger research groups topics are often passed from one person to another as researchers graduate and are replaced by incoming students. In some cases therefore the research reported within a thesis might be part of ongoing work within the thesis author's research group.

For this reason supervisors and other researchers within that research group may wish to restrict access to the knowledge within a thesis even in the absence of any clear commercial value to that knowledge, to ensure that they have the chance to fully investigate and publish any related research before revealing their results to potential competitors. Laura Broad, head of data collectors in the NCC pilot project, reported that supervisors were much more amenable to offering data collectors older theses on topics which were unlikely to be taken up by future students for further research.

2. **Sensitive content** that may pose a safety or privacy risk

The subject matter of theses varies widely, and in some cases theses address topics that would be considered politically controversial or sensitive. For example, a researcher originally from outside the UK may write a thesis that is critical of the government in their home country. Where theses are only available as print copies within an institutional library this is unlikely to cause problems – but if a thesis is accessible online from any part of the world this may put the thesis author, their friends and family, or sources consulted during the course of research in personal danger. In such cases there is no choice but to restrict access to the thesis in such a way as to protect the safety of all parties.

Similarly in theses which address topics such as terrorism it may be necessary to restrict access to theses in the interests of national security.

In other cases, theses may include information that would threaten the privacy of research or interview subjects. As with issues of third party copyright discussed in section 4.1.1.1, in theory thesis authors should employ best practices to anonymise all research data and prevent this becoming a problem. In practice however, personal information within a thesis may not be properly anonymised. Again as with third party copyright this potential concern is best addressed by better education about and implementation of best practices.

4.1.2 Working with the Community

With respect to access and re-use of theses, a key point made by Lucy Ayre (Research Online Manager at the London School of Economics) is that "Everything is connected." That is, issues of file formats, accessibility, copyright, openness, and re-usability for text and data mining are all interrelated. Like all other institutions consulted during the course of this project, it is the view of the London School of Economics that all changes made to policy in these areas should be carried out in the interests of researchers and the research community.

This means that although many researchers and institutions are keen to reap the benefits of making the information within theses more available, moving forwards on this issue will inevitably involve a gradual, stepwise process, making sure all vested interests are included and consulted along the way. Concerns and risks must be understood, clarified and addressed, but are not insurmountable. Education and best practices, coupled with the option to apply temporary or indefinite embargos to theses on a case-by-case basis, should all but eliminate the concerns associated with making theses more available online – eventually.

5 Mining EThOS

Having covered in detail the technical, legal and ethical concerns relevant to text and data mining, it is now possible to describe a framework for assessing prospective text and data mining projects in terms of the feasibility of using the EThOS database, as well as potential risks and benefits to the British Library and other stakeholders.

5.1 Assessing Potential Projects

Despite the fact that text and data mining is in many respects an inchoate technology, several requests have already been made to use theses recorded in EThOS for text and data mining projects. It seems likely that such requests will increase in frequency alongside the growing recognition of theses as an underutilised source of information, as well as the development of more advanced text and data mining techniques.

It is the goal of the British Library to respond to these requests as efficiently and as positively as possible, while respecting the interests of all stakeholders in higher education and research communities. With this in mind, whether a prospective text and data mining project can and should make use of theses recorded in EThOS can be assessed with four fundamental questions:

- Is EThOS infrastructure sufficient for the prospective project?
- How will thesis contents be used by the prospective project?
- For which theses within EThOS would this use be permitted?
- Will the intended use benefit the Library or other stakeholders in the HE community?

Each of these questions is addressed in detail below.

5.1.1 Is EThOS infrastructure sufficient?

As discussed in section 2.1, large-scale text and data mining requires that theses be accessible not just on an individual basis but for bulk access by automated tools. At this point in time EThOS is of limited usefulness in both these respects. Firstly EThOS does not hold the full text of all records within its database; many records describe print theses that have yet to be digitised, or electronic

theses held in separate institutional repositories. Secondly records which *do* contain the full text of a thesis may only be accessed one at a time, after agreeing to terms and conditions that apply to accessing the thesis. While these limitations are manageable for small-scale projects, they significantly limit the feasibility of harvesting greater numbers of theses for large-scale text and data mining.

A second concern is the comprehensiveness and accuracy of the metadata such as subject classifications associated with EThOS records. As described in section 2.1.2, metadata can be an effective way to filter a database for potentially relevant targets for text and data mining. Although great improvements have been made to the metadata attached to EThOS records, around 10% are still without even subject classifications. This could limit the ability of a text and data mining project to narrow down a dataset of relevant theses, although alternative methods of selecting a dataset do exist (see section 2.1.3).

Thirdly, as discussed in section 2.2, the vast majority of the information contained within theses is unstructured – almost all theses are held either as PDF files or TIFF files. This means it is less readily available for analysis and generation of new knowledge. However unstructured data is an extremely common problem in the field of text and data mining – it is likely that any projects intending to carry out large-scale analysis of theses will be aware of this issue and prepared to cope with unstructured data in the form of PDF files.

5.1.2 How will thesis content be used?

As discussed in section 3, several aspects of text and data mining processes require considerations of the intellectual property rights that apply to the content being analysed. Whether or not a prospective text and data mining project can be lawfully carried out on theses would depend on whether it entails making or reproducing copies of theses in part or in whole, and if so, whether those activities will be carried out in ways that fall under the scope of exclusions from or exceptions to copyright.

The main ways in which text and data mining is likely to entail making or reproducing copies of works are:

- **Preparing works for analysis:** creating a digitised or annotated version of a work generally entails making a copy of the entire work
- **Publishing outputs of analysis**: output of text and data mining projects may re-publish excerpts copied from the works being analysed

Although it would be the responsibility of a prospective text and data mining project to ensure that any copying of copyrighted works is carried out in a way that is lawful, it is in the British Library's interests to be aware of the limitations to lawful copying in these circumstances and to avoid the appearance of condoning or facilitating copyright infringement.

5.1.3 For which theses would this use be permitted?

If a prospective text and data mining project intends to make copies of theses in part or in whole, an important consideration is whether permission is required from thesis copyright owners for those activities to be lawful.

The combination of the copyright exceptions for text and data mining and quotation (discussed in sections 3.1.2.2 and 3.1.2.3, respectively) effectively mean that when text and data mining is carried out for non-commercial purposes, permission is not needed to make copies of entire works which are lawfully accessible, or to copy and re-publish parts of works so long as they constitute fair dealing.

However there is possible cause for concern due to the fact that some theses within the EThOS database may have been digitised and made available online unlawfully, as well as due to the uncertainty surrounding the definition of "non-commercial" purposes in the context of text and data mining. Again it would be the responsibility of a prospective text and data mining project to act within the law, but within the British Library's interests to be aware of these possible concerns.

If the prospective text and data mining project is not able to operate within the scope of exclusions and exceptions to copyright it would be necessary to identify theses for which specific permissions for re-use have been granted. At this point in time, no information about rights of access or re-use is attached to individual records within the EThOS system. In some cases the thesis file itself may include a statement about rights indicating the licensing situation for that thesis. In others, information about rights and licensing may be available through the awarding institution's own website. In still others there may be no indication whether permission from the copyright holder has been obtained to either digitise the thesis or make it accessible online.

This makes it difficult to quickly or easily identify theses within EThOS for which permissions have been granted for a given type of re-use. However, it may be possible to identify institutional repositories that contain collections of theses for which particular permissions have been granted, in which case a prospective text and data mining project would be able to work with this smaller subset of the EThOS record.

5.1.4 Will there be benefits to the Library or other stakeholders?

This is a difficult question to answer in general terms, as the potential end uses of text and data mining of theses are still largely unknown. However two recent cases exemplify instances of text and data mining that could bring direct and indirect benefits.

In the first case, a group from Virginia Tech approached the British Library with a request for theses on which to train an algorithm.[48] Their goal is to develop an automated method of analysing documents and assigning Library of Congress subject classifications to each. This project has a clear potential benefit to the Library: once the algorithm is deemed reliable, it could be used to assign subject classifications to the 30,000 EThOS records without subjects, thereby enhancing the metadata in EThOS and the discoverability of EThOS contents.

In the second case the British Library was approached by a project called FLAX which requested access to law theses from the EThOS database. FLAX intends to use these theses to generate language-learning tools specific to the field of law, so that students from non-English-speaking backgrounds studying law can work on their language skills with vocabulary and phrases that are

[48] "Training" an algorithm generally entails testing it against a "learning set" of documents to identify and correct flaws. In this case, the learning set would consist of theses which already have subject classifications assigned. The algorithm would be run against the learning set and the results compared to existing subject classifications; any disagreement would be investigated and used to adjust the algorithm for greater accuracy.

relevant to their studies. Although this project does not directly benefit the British Library, it can clearly be considered beneficial to members of the higher education community.

5.2 Existing Projects

5.2.1 The National Compound Collection

Unfortunately, at this point it seems unlikely that the NCC project would be able to make much use of theses as a potential source of novel molecules. Considerations of the potential barriers to using theses for the NCC project are described below, with reference to the framework set out in section 5.1.

- Is EThOS infrastructure sufficient for the prospective project?

The RSC's tool for analysing theses and ranking how similar they are to RSC publications could be effective in identifying a corpus of theses to target for the NCC project, as discussed in section 2.1.3. However as mentioned in section 5.1.1, there is currently no way to harvest theses in bulk from the EThOS database. The NCC project would be looking to analyse and extract information from hundreds of theses, which is impractical within the current infrastructure of EThOS. Furthermore, tools for automated extraction of molecular structures from unstructured data, such as images in theses stored as PDF files, do not yet offer significant advantages over manually extracting molecular structural information.

- How will thesis contents be used by the prospective project?

Although the initial goals of the project are to extract molecular information and place it online in a free database (ChemSpider), the intent of the RSC is to pursue potential partnerships with industry in order to monetise outputs of the NCC project and make it self-sustaining. This could well be construed as commercial intent, and the copyright exception for text and data mining applies only to non-commercial research. Therefore even though the outputs of the project are scientific facts (chemical structures) which do not attract copyright, if the processes used to extract those facts involve making copies of large parts of theses, they may not be lawful without permission from copyright holders.

- For which theses within EThOS would this use be permitted?

At this point although several higher education institutions allow free online access to and re-use of their theses for non-commercial purposes, no institutions are known to have policies giving blanket permission for *commercial* re-use of theses. Therefore if the intent of the NCC project is considered commercial there are no collections of theses that may currently be lawfully copied in their entirety for the project.

- Will the intended use benefit the Library or other stakeholders in the HE community?

In an area that is fast becoming more collaborative among researchers,[49] making thesis content more discoverable has the potential to drive research and innovation and benefit not only the higher

[49] In this opinion piece David Fox of the RSC describes how the need to manage increasingly complex challenges is driving researchers to collaborate in the traditionally highly competitive field of drug discovery. (Retrieved 2014-10-02)

education community but the general public, as well as helping to demonstrate the value of UK theses. If it were able to proceed, the NCC project could therefore bring significant benefits to EThOS and other stakeholders in the HE community. Unfortunately at this time these potential benefits are outweighed by technical and legal barriers to the project.

5.2.2 FLAX Interactive Language Learning

As mentioned in section 5.1.4, the British Library was recently approached with a request for law theses to aid in the development of language learning tools. This prospective project was assessed using the framework described above and it was decided that it was a project that EThOS could and would support. Below is a brief summary of considerations undertaken to come to this decision.

- Is EThOS infrastructure sufficient for the prospective project?

In this case the team from FLAX only required access to a small number of theses – fewer than fifty. The lack of bulk access to theses through EThOS was therefore not a problem as on this scale theses could simply be downloaded manually. Subject classifications within the metadata were similarly more than sufficient to identify a small group of theses created in the subject of law, and the team had already developed tools for creating language tools from unstructured data.

- How will thesis contents be used by the prospective project?

FLAX's intention was to use theses solely to develop language tools for non-commercial, educational purposes, without reproducing any significant part of the thesis.

- For which theses within EThOS would this use be permitted?

As the purpose of FLAX's project was non-commercial they were confident that their use of theses would be covered by the text and data mining exception to copyright.

- Will the intended use benefit the Library or other stakeholders in the HE community?

As discussed above, although FLAX's project brings no direct benefits to the British Library it will clearly be of benefit to parts of the higher education community. Furthermore as many stakeholders in the higher education community are keenly interested in seeing evidence of tangible benefits of text and data mining projects, if FLAX's project succeeds it could serve as a valuable "success story" to demonstrate the potential benefits of text and data mining of theses.

5.3 Looking Forward

As discussed in this report, there are several areas in which EThOS infrastructure could be improved to facilitate text and data mining projects in future. In particular improving **metadata** such as **rights information**, and developing avenues for **bulk access** to theses would be particularly useful.

The more comprehensive and accurate metadata records are for theses within EThOS, the more discoverable (and therefore useful) those theses will be. Continuing to improve metadata in the EThOS database should be an important consideration not just for facilitating text and data mining projects, but for facilitating use of EThOS records in general.

In particular, developing the infrastructure to include human- and machine-readable rights in the metadata for EThOS records would greatly simplify some of the legal uncertainty around access

to and re-use of theses. At the moment, legal uncertainty could pose a significant barrier to future text and data mining projects. Work is now underway to begin to identify what permissions may be applied to which EThOS records, with the ultimate aim of attaching rights information to as many records as possible.

Finally, bulk access to theses is crucially important to any potential large-scale analyses of theses in aggregate. The current infrastructure that requires an individual to click through the terms and conditions of access for each thesis downloaded makes downloading even a handful of theses frustratingly onerous; for larger text and data mining projects it is simply not feasible to harvest theses in this way.

Looking further into the future, several other issues are likely to become more relevant to ensuring that the EThOS database provides as valuable a service as possible to all interested parties.

One issue not addressed in this report, for example, is the possibility of preserving and making available the *data* underpinning thesis results. This is an idea that is beginning to surface among higher education institutions around the UK, as particularly in the natural sciences information such as calculations, spectral data, and other recorded measurements are being recognised both as a significant part of the work that goes into a PhD thesis and as crucial to any future investigations of the results reported in theses. In some subjects this data can already reach volumes in the gigabyte range; looking forward, as institutions begin to consider ways to preserve this data alongside theses themselves, the British Library will need to consider its approach to the question of preserving thesis data in the EThOS database.

Another concern for some higher education institutions is how file formats and storage should be addressed in the preservation of theses and data. Several institutions contacted during the course of this work reported that the main reason holding them back from moving to digital-only collection of theses was concerns about their ability to preserve theses indefinitely. Although it is not the British Library's place to dictate how institutions should handle their own records and data, these institutions expressed an interest in any guidance the Library might be able to provide with respect to best practices for preservation of theses.

Another interest reported by institutions was in consolidating usage metrics across EThOS and institutional repositories for particular theses, in order to make usage data more accurate and comprehensive. This presents challenges in the form of risks of mismatching or duplicating records across different repositories – challenges that could be solved by attaching unique object identifiers to theses, another idea that is beginning to gain traction within the higher education community.

Finally, as mentioned briefly in section 5.2.2, *many* stakeholders in higher education are actively looking for "success stories" to demonstrate the potential benefits and uses of text and data mining of research outputs. One possible idea to promote future projects and demonstrate value might be to ask a small number of institutions if they would be willing to volunteer their theses for non-commercial test cases, gather a corpus of theses from those institutions, and present those theses either to individual interested parties or as part of a "hack day" type of event to encourage developing creative uses for the information within them. Any positive results generated by such a project would have a powerful impact on the move towards greater access to and re-use of theses, which would benefit researchers and stakeholders across the higher education community.