Text Analysis and Visualisation: creating deep interfaces to read textual document collections

Jaume Nualart Vilaplana

9 December 2016

A thesis submitted for the degree
of Doctor of Philosophy in Communication
University of Canberra

||*||

Abstract

This research brings together data analysis with software engineering and visualisation, with a specific focus on text mining and large document collections. My aim is to devise new, rich, and simple visualisation interfaces, which I call deep interfaces.

With deep interfaces I introduce the idea-rich content as a product of the statistical analysis combined with human curation of labels and interpreted as a flow of subjectivity, complexity, and diversity between reader and interface and vice versa.

The focus of such interfaces is not the representation of textual document collections as in Moretti's distant reading, but to revisit traditional reading from the point of view of state of the art methods of textual analysis. Thus, the proposed interfaces can help us discover and explore text document collections by reading their contents. This is a practice-led research project that develops theoretical issues through the generation of practical artefacts. The research process is cumulative, following a reflexive methodology. The key outcomes of the project are embodied in an interface to a large collection of ANZAC war diaries: Diggers' Diaries — http://diggersdiaries.org.

UNIVERSITY OF CANBERRA

FORM B

Certificate of Authorship of Thesis

Except where clearly acknowledged in footnotes, quotations and the bibliography, I certify that I am the sole author of the thesis submitted today entitled –

Text analysis and visualisation: creating deep interfaces to read textual document collections

I further certify that to the best of my knowledge the thesis contains no material previously published or written by another person except where due reference is made in the text of the thesis.

The material in the thesis has not been the basis of an award of any other degree or diploma except where due reference is made in the text of the thesis.

The thesis complies with University requirements for a thesis as set out in the *Examination of Higher Degree by Research Theses Policy.* Refer to http://www.canberra.edu.au/current-students/current-research-students/hdr-policy-and-procedures

________________________________ ___15_/_May___/__2016____

Candidate's Signature Date

________________________________ __16__/__May___/__2016____

Primary Supervisor's Signature Date

NOTE: The wording contained in Form B must be bound into the thesis, preferably as the third page, and signed by the author of the thesis and the supervisor. However, the exact layout does not need to be duplicated.

Contents

List of Figures

List of Tables

Acknowledgements

To Jordi, Kiesse, and Amelia. To Maria Rosa and Jaume. To the rest of my family. To my best friends.

Thanks to Dr Gabriela Ferraro, for her support, collaboration, and friendship; to Dr Joelle Vandermensbrugghe for her guidance to PhD students, her advice, and her friendship; and to Dr Giulio Zambon for his advice, conversations, and friendship.

I'd like to thank my colleagues from the Machine Learning Research Group of NICTA for their help, support, and interest shown in relation to my research project. Special thanks to my former supervisors: Dr Wray Buntine, Dr Mark Reid, and Dr Hanna Suominen; and to Dr Bob Williamson, the leader of the group, for his support and trust.

I'd like to thank the Faculty of Arts and Design staff for their help and professionalism.

Thanks to Dr Rob Fitzgerald for his warm support and the interesting and challenging project we shared.

And, finally, I would like to thank Dr Mitchell Whitelaw, my thesis supervisor, who has shown and taught me how academic research is done. Mitchell has introduced me to the academic discipline of digital humanities, a territory where we have shared our passion for data visualisation.

1. Introduction

In 2015 there was a growing amount of accessible data. We know how to handle the growing speed with which the data becomes available(Cisco, 2016, Bono, 2015) and its volume (Pappas, 2016). . This allows us to plan future information infrastructure needs to store it, but what we do not know is how we are going to use such an amount of data. This thesis aims to contribute new methods we can use to access those data, focussing on textual documents that are grouped into digital collections.

This chapter briefly presents four issues that are fundamental to understanding the nature of this research project. Firstly, the context and motivation that informs this research on interfaces to textual document collections; secondly, the project's aims and goals; thirdly, an introduction to the practice-based and practice-led nature of the research and its corresponding methodology and process; and, finally, an overview of the whole project, setting out the structure and topics of the chapters that will follow.

1.1. Context and Motivation

We have moved on from information overload in the 1990s to big data. Information overload was presented as the biggest problem of the digital age, and the necessity of computational help to deal with it was pointed out ((Maes et al., 1994). Today, in the age of big data, this excess is seen as an opportunity to index, mine, analyse, and discover insights, patterns, and, ultimately, solutions to information-related problems (Mayer-Schönberger and Cukier, 2013).

This project is developed in the context of big and accessible data, and the unsolved problems we deal with due to these large amounts of data. Specifically, this research focusses on how we can deal with large collections of textual documents.

These texts can be found in major public institutions that host digital collections, usually accessible online by means of a text-based interface. This project explores the possibilities of implementing new techniques to improve the way we interact with such large collections of text.

According to that goal, this project has some necessary ”preconditions”, such as the availability of document collections in digital form. This needs to be spelled out because, while the collections are catalogued online, the contents of these collections are not always accessible. Different contents from different knowledge domains have different degrees of access and are used for different proposes. This project works with text collections from academic and cultural domains, where full text documents are increasingly available. Full text access to the contents of text collections is also required in order to apply state-of-the-art techniques of text analysis that can enrich what we know about the collections. Text analysis techniques such as topic modelling are increasingly mature and well developed and, due to new easy-to-use software, their use is spreading within digital humanities (DH) and other fields. Furthermore, the increasing power of web browsers and the establishment of solid technical standards have enabled the development of rich interactive displays.

From a more human-centred point of view, the urgent need of tools to deal with the new digitised collections is related to education, culture, and, ultimately, personal and collective freedom. In that sense, my experience over the last fifteen years working with free media, public libraries, and free knowledge is so far integrated in the project's goals as to make digital collections, and especially text collections, more accessible using a combination of analysis and visualisation techniques.

1.2. Aims and Goals

This research project seeks to bring together data/text mining, software engineering, and visualisation, with a specific focus on text mining and large document collections. Its aim is to apply text mining and analysis to the creation of powerful new interfaces to digital document collections.

As a broader goal, this project seeks to bring to life the stored knowledge that lies behind the poor, text-based interfaces of textual document collections. Its aim is to develop new interfaces to these collections in order to make them more accessible and useful as tools for knowledge discovery and transfer in education, research, and other cultural contexts. This project introduces ”deep interfaces”, which combine text analysis with visualisation elements.

The practical outcomes of this project are web applications with interfaces focussed on exploration for reading. Exploration-for-reading tools take up the challenge of reading large-scale collections that include tens of thousands of pages. By bringing the collection's content “to the surface”, deep interfaces reveal it to the user and make it accessible for reading. In contrast, many representations of digital collections concentrate on high-level overviews, thereby pushing the actual content farther from the user.

1.3. Practice and Research

According to Candy (Candy et al., 2006), there are two kinds of research projects based on practice: practice-led research and practice-based research. This project involves a combination of both types. The main focus of this research project is to advance knowledge about practice; more specifically, it seeks to learn about practice through making artefacts. The process of making combined with reflection and iteration has led to new theoretical contributions. The notion of deep interface is an example of a novel concept that has arisen as a consequence of the practice, thereby matching Candy's definition of practice-led research.

At the same time, in practice-based research the contribution to knowledge is demonstrated through the existence of created artefacts, and this also applies to this project, which has produced and published several interfaces to significant text collections (see chapter 5 Results). These artefacts embody the novel reading experience presented in this dissertation. They also show that the techniques are feasible, and how these approaches can be implemented in practice in a reusable and accessible form. All source code for the project is accessible through open repositories ((Nualart, 2016).

1.4. Dissertation Overview

This dissertation is organised in seven chapters, including this introduction. The chapters follow a traditional dissertation outline.

Chapter 2 ”Context and background” reviews relevant literature and practice in three related fields. First, it addresses digital libraries (DL) in relation to text collections. It shows how DLs are reaching a post-human scale while, at the same time, major public institutions do not use state-of-the-art tools and techniques in the interfaces offered to explore and discover their collections. Second, it reviews existing work on text analysis applied to text collections, showing that there are a variety of techniques —document clustering, topic modelling, topic labelling, and evaluation— available today, but no extended application of them in interfaces to text collections. Thirdly, and finally it reviews interfaces to text collections. This part revisits some paradigms applied when interfacing to DLs, discusses the use of data visualisation techniques to represent collections, and provides a brief review of e-reading and interfaces oriented to reading texts.

Chapter 3 describes and analyses the gaps, opportunities, and lessons learned from the literature review. On that basis, it then outlines the project's key research questions and contributions to knowledge. The lack of advanced interfaces to text collections in major institutions described in Chapter 2 becomes in Chapter 3 a well defined gap. The innovation found outside the institutions, in practitioners and research demo sites,shows that there is an opportunity to use state-of-the-art techniques of analysis and apply them to text collections. Finally, this chapter shows that there are techniques to deal with the problem that readers have no time to read the vast amounts of available texts and materials. The proposed solution is contained in the concept of crossreading, that is, that large collection of texts can be segmented into small pieces and then a sample of the pieces can be read, thereby providing a view, or at least a taste, of the collection.

Chapter 4 describes the methodologies applied during the research project, including practice-led and practice-based research and reflexive and empiricist methodologies. It then proceed to describe the methods, tools, and techniques used in the development of the artefacts. They include techniques for data gathering (i.e., APIs and spiders), text segmentation, topic model analysis, topic model labelling and evaluation, and interface development with modern Javascript libraries. The chapter concludes by presenting the methods used to evaluate the artefacts. That is, online questionnaires and semi-structured interviews.

Chapter 5 presents the results of the research project, introducing each artefact in order of creation:

- Visference, an improved interface to the papers of an academic conference;
- • Crossreads (I and II), a way to read text collections as fragmented narratives;
- • Diggers' Diaries (I and II), an interface to support the reading of a historical collection of diaries from Australian soldiers in World War I.

Each artefact is described with: a figure showing a detail of the generated interface, the artefact name and description, the dataset(s), the collaborators, the data process, the data analysis, interface development, user evaluation studies, project outcomes, and a brief narrative about the artefact.

Chapter 6 presents the discussion. It considers the significant gaps described in Chapter 4 —the text analysis, the reading task, and the interface— and, after analysing them in detail, defines the concept of deep interfaces, which refers to a group of techniques and strategies aimed at improving the creation of new interfaces to text collections. Deep interfaces encompass techniques that help to answer the main research question: How might we create interfaces to text document collections that let us explore the collection by reading its content? The idea behind deep interfaces refers to interfaces that look similar to standard web sites but offer options that take the user to an interpretative browsing of the contents. That is, to the deeper meaning and structure of the collection. The so-called depth is a product of the integration of text analysis elements into the collection interface.

Chapter 7 recapitulates the whole dissertation before discussing the limitations, applications and possible future work in relation to the proposed methods and the presented artefacts. In closing this chapter presents some final thoughts as the conclusions of the research project.

2. Context and Background

Most of the data we generate everyday is unstructured, mainly in the form of text (Grimes, 2008, Heer, 2010). Part of this text is born digital (social networks, emails, etc.) and is stored on hard disks, and part comes from printed paper and, after a digitisation process, is mostly stored in digital libraries (DL) to form collections. As the number of text documents available digitally grows every day, the urgency for better information systems to access them grows too. In the age of data, every advance in a field of knowledge, sooner rather than later, will affect every other field, especially in an interdisciplinary domain such as the study of data. As this field grows in size —research centres, publications, industry, studies —it grows in scope and diversity —biology, social networks, health, government, security, entertainment, etc.

This chapter reviews current research on and development of rich interfaces to DL, focussing on textual document collections. The related fields combined in this review are: data analysis, data visualisation, and interface design. This project works at the intersection of these three fields, in an attempt to contribute to generating an interdisciplinary space where a combination of methods and points of view can produce good results.

The first section of this chapter reviews DLs from important organisations that host large digital collections and very large numbers of digital objects. All the reviewed organisations are public institutions, with the exception of the Internet Archive, which is a San Francisco-based non-profit digital library. All provide free public access to their collections through the web. The review finds that the big institutional DLs studied are also meta-libraries. That is, aggregators. This aspect contributes nowadays to the post-human scale of digital libraries: there are so many texts that reading even a small portion of them becomes humanly impossible. It implies that in order to access such collections we need computational tools to guide and support our reading. I will show that these DLs are not using state-of-the-art interfaces to their massive amount of contents. Instead, big DLs such as Trov (Trove) and Europeana (Europeana, 2008)) are based on traditional text-based web pages. The review will also show some innovative proposals from institutions, demonstrating the gradual emergence of new features to help us reading, discovering, and visiting digital libraries.

The second section of the chapter reviews text analysis. It shows that DL catalogs based on standard metadata are unable to represent the complexity of the large collections. It reveals that text analysis is important in a list of knowledge fields related to data. Namely, information retrieval, data mining, natural language processing, and computational linguistics. In other words, text analysis seems necessary to represent the mentioned complexity. In this direction, established text analysis tasks are listed and described, such as: text categorisation, entity and concept extraction, taxonomy induction, sentiment analysis, and document summarisation. The review then proceeds by introducing text-similarity techniques, such as clustering. Included is a brief history of clustering as a technique used since the 1990s to categorise or group documents within collections. It categorises three kinds of text collections according to the kind of cataloging: raw metadata, extended metadata, and object analysis. Finally, as an approximation to the rhizomatic complexity of the data, and especially texts, the concept of Capta is reviewed —Capta was defined by Johanna Drucker as an active position to counter the non-critical attitude and position of acceptance we have to data. A review of topic modelling follows, including the labelling process and its evaluation, concludes this section.

The third and final section of this chapter presents a review of interfaces to digital libraries and collections of textual documents. It begins with a review of philosophical and conceptual paradigms from different authors encompassing the last twenty years, including Shneiderman, Marchionini, Greene, Stamen, Dörk and Whitelaw. It then reviews standard and trending practices in interfaces to digital collections, thereby showing that mainstream DL institutions tend to use traditional text-based web pages as interfaces. After reviewing the role of data visualisation and interfaces to digital collections, it shows how visualisation can be used as an element of the interface. Finally, it provides a brief review of interfaces dedicated to assist in reading text that shows how standards for reading texts on digital devices are established and offered for most devices and interfaces in contemporary systems.

Based on the content of this chapter, Chapter 3 will identify gaps and opportunities, and the contributions to knowledge that this research offers.

2.1. Digital Libraries and Text Collections

The term Digital Library can refer to a collection of digital objects, and, among other things, to the management of the collection, or to the institution or service offered by librarians. Here we use DL as a collection of digital objects. A DL can include objects such as video, images, text, and audio, all in an electronic format. DL also contains metadata, that is, data about the digital objects.

A digital collection can contain, according to Shneiderman, seven types of data, where text is a 1-dimensional data-type that includes "textual documents, program source code, and alphabetical lists of names which are all organized in a sequential manner" (Shneiderman, 1996). Textual documents can be kept together, forming a text document collection. As in other collections, textual documents can have diverse structures and contents. In this dissertation we refer to textual document collections that give access to the full text of the documents, so the texts can be read, analysed and exhibited.

When a set of textual documents that are part of one or more DL are grouped, they form a collection of textual documents. The cohesion factor that generates a collection can be broad, and, therefore, can affect the way the collection is presented and accessed. For example, the cohesion factor of a digital collection can be historical. This happens with, e.g., collections of press articles from a period of time and a defined location (Trove, 2010, library of congress, 1800). But a collection can be also a product of a curated work (American Anthropological Association, 2016). In both cases the nature and history of the collection can affect the design of the interface to access it.

Digital Libraries were born in the second half of the twentieth century, in parallel to the computer, and the digital era. The initial dream of a digital library as an integral source of knowledge, accessible from any place, is a reality today (Bush, 1989, Besser, 2004). One of the considered first DL is dated in 1976, the Oxford Text Archive (OTA, 1976). contains literature and language resources, this is, textual document collections. The term DL emerges in 1994 with the ”Digital Libraries Initiative” (Fox, 1999). For some time the terms “virtual”, and electronic are also used to refer to DL. Today “virtual libraries” usually refers to distributed libraries, known also as aggregators (Fox and Sornil, 2003).

As Borgman [1999] argues, ”research and practice in DL has exploded worldwide in the 1990s”. Borgman identifies several factors feeding the growth of DL, including the increased availability of networked computing and the availability of targeted research funding. Rehear (2004) explains that US, Europe, and Asia governments invested important amounts in DL research projects in the second half of the 1990s. Initially some humanists and librarians were sceptical about computer scientists researching DL, and not applying the results to real libraries. Hurtle in an editorial of D-Lib Magazine in 1999 says: ”Rightly or wrongly, the DLI-1 grants were frequently criticised as exercises in pure research, with few practical applications” (Hirtle, 1999). esser, about this period, says ”we will call this the experimental stage of digital library development”. In the period 1995-2000 international conferences, journals, and online news publications about DL were created in both scientific and humanities disciplines (Besser, 2004).

Today digital libraries can be found across a number of disciplines and domains. The scope and range of current DLs, includes:

- Public collections: institutional archives, public libraries (e.g., Library of Congress).
- Private not-for-profit archives and collections (e.g. The Internet Archive).
- Commercial databases, online book shops (e.g. Google Books, Amazon, etc.).
- Public scholarly collections (e.g. university repositories)
- Commercial scholarly databases and journals (e.g. JSTOR, EBSCO, etc.).

A key point of DL is the access to their collections and digital objects, that is, the copyright of their contents. Sometimes, institutions can offer only access to metadata of specific digital objects due to license restrictions. Visualisation approaches that are reviewed in this chapter refer to freely accessible online DL. An example of this is the Million Book Project (Linke, 2003). his option looks like one of the solutions to make the contents of DL accessible to the public, but there are still unsolved financial problems, since open access contents are expensive to maintain, and no clear business model is associated with them (Seadle and Arms, 2012). In any case, there are creative and revolutionary proposals in academia, like Shamos that argues: ”copyright does not protect facts, information or processes, we propose to scan works digitally to extract their intellectual content, and then generate by machine synthetic works that capture this content, and then translate the generated works automatically into multiple languages and distribute them free of copyright restriction” (Shamos, 2005).

Examples of DL that offer open access to their contents include the Internet Archive, whose collections contain digital objects that are out of copyright and donated by public libraries, as well as works under copyright but published under license from the copyright owner. Archive.org acts as a publisher (archive.org). Europeana is a metasearch engine that harvests metadata from other institutions. Another case of, also called, aggregators (of metadata) is Trove (Trove, 2009), the biggest Australian source of digital collections . Trove accepts contributions from remote collections, but it is also a ”growing full-text digital resource” where press articles, personal diaries and letters, books, journals, and biographies can be freely accessed. In the research world a notable open access repository of scientific papers, arXiv.org, was established in 1991 by Paul Ginsparg (arxiv.org). This is a repository for preprint versions of scientific papers, as Seadle says ”in essence anybody can post an unreviewed paper claiming that it is original research” (Seadle and Arms, 2012). Arxiv is a modern concept for the distributed curation of digital library content by the community of its users.

The scale of these archives is notable, especially in the context of this project, which focuses on reading. For example Trove offers over two hundred million digitised newspapers. Arxiv.org offers over a million e-prints of scientific papers. The Open Library of Archive.org offers one million free ebook titles available to read. Modern digital libraries have reached what might be termed a “posthuman” scale, where they are too large to be read in the conventional sense.

Today some DL have become meta-libraries, also referred to as virtual libraries, and as aggregators. Aggregators collect metadata from source DLs and index it in new databases. This is the case of Europeana, an aggregator that collects metadata from more than two thousand cultural institutions across Europe (Europeana, 2008). The Australian Trove ”brings together content from libraries, museums, archives and other research organisations and gives you tools to explore and build” (Trove, 2009). Other popular aggregators include public libraries, and university repositories. These repositories are also sources collections for aggregators, allowing localization of titles through meta-search (OpenDOAR, 2005).

Since 1999, digital libraries use systems to catalog their contents based on metadata ((Milstead and Feldman, 1999). Metadata practices, as a descendent of the cataloging, grew out of traditional library management —indexes, card catalogs etc. Metadata repositories can store data about physical and/or digital objects. Best practices in repositories recommend the use of standards for metadata vocabulary terms (Duval et al., 2002). A widely adopted set of standards in this direction is Dublin Core Metadata Initiative ((DCMI, 2001), that defines a vocabulary of terms to be used for physical and digital resources. Standardization of metadata allows easy and more effective sharing of resources among machines.

With the evolution of online resources, the relation between data and metadata becomes diffuse, thus, e.g., Munzner, in her recent manual ”Visualization analysis and design” (Munzner, 2015)argues that ”the line between data and metadata is not clear, especially given that the original data is often derived and transformed”. Therefore Munzner, decides to not ”distinguish between them, and refer to everything as data”. In this text we differentiate between data and metadata in the sense that data is unstructured (since we study textual documents) and metadata has the structure of a list of property-value pairs (e.g. in JSON {”Title”:”Tree of Science”,”Author”:”Ramon Llull”, ”year”:1596 } etc).

Focusing on textual collections coming from paper, and the access to their contents, there is one aspect that differentiates textual documents from other media, that is the digitization process (Seadle and Arms, 2012).The difficulty in having full text access to the documents of collections is that, once the documents are digitised, a transcription is needed, and this task requires human resources. OCR systems can help in the process of transcription but success depends, mainly, on the quality of the copies. The accuracy of OCR has a big impact on the critical study of texts (Hinneburg et al. 2012, Milligan 2013). Sometimes the digitised collections are published online, as in the collections of the University of Washington Libraries (U.Washington), and the State Library of New South Wales (SLNSW, 2010). The digitised texts are not always available as machine readable text. When the contents are readable by machines, then we can get all the advantages of information retrieval, and text analysis. Several institutions have developed strategies to make them available as text: Trove (Trove, 2010) uses "crowdsourced" corrections to an automated OCR transcription. The Bentham papers archive uses crowdsourced transcription but is now also developing a machine learning system for transcribing handwritten textuses "crowdsourced" corrections to an automated OCR transcription. The Bentham papers archive uses crowdsourced transcription but is now also developing a machine learning system for transcribing handwritten text (see e.g. Causer and Wallace, 2012, UCL, 2000, TranScriptorium, 2013) Finally the State Library of New South Wales (SLNSW) funds manual transcription of documents such as the WWI diaries collection (SLNSW, 2014)

When the full text of collections is available, then we have metadata and data, all as text. Additionally, the textual documents contain some kind of structure: paragraphs, sentences, and/or chapters, sections. This extra data can be extracted and stored as metadata too, thus, enriching the structure and adding extra dimensions to the collection.

In addition to this, in this dissertation we talk about data that comes from the analysis of the contents of the collection. Metadata is defined as the data that describes an item in a collection as a whole. However, analysis can result in data that is derived from, or is a sub-product of, the contents of an item. In a web interface that includes these two kinds of data, the format used to save them would be similar. The data is saved as property-value pairs grouped by items. Usually, a unique set of properties is defined for each item of the collection.

The main difference in both kinds of data is in its structure. In most of the cases the standard properties about items are 1-dimensional, e.g. title, author, date, category, email, URL, gender, etc. The chapters (5, 6, and 7) show that the output of data analysis for the presented artefacts include properties that are multidimensional, e.g.:

- Crossreads for each segment of text there is a list of IDs of other text segments within the collections, sorted by similarity.
- In Diggers Diaries for each segment of text there is a list of percentages of each topic
- Visference is a similar case: each paper of the collection has a property that defines the percentage of each of the ten available topics in relation to its contents.

The ”library” model for digital collections of documents brings with it specific conventions. The traditional library stores books, and makes them findable (via index or catalog). One key function for the librarian is to ”look inside” the collection, advising users on its content. With digital textual documents the collection contents can become transparent. At the same time new challenges of scale arise. Thanks to computers it is possible to efficiently analyse large amounts of text; this analysis can bring structure to collections of documents, and a deeper knowledge of the collection contents. This new information can helps with generating overviews, relationships, richer models of content, and, eventually, improved understandings of a collection and its contents. The next section presents a review of text analysis and text document collections.

This section has reviewed big and public institutional DL. Most of them show the tendency to become big aggregators, this is, concentrations of information (metadata) to access to local and remote resources (digital collections, and digital objects). This aspect makes the DL grow even faster, to what we call a posthuman scale. This scale brings urgency for new tools to visualize, explore, and to read the text documents of the DL. The review shows the contrast between this posthuman scale, and the lack of active innovation in the online versions of the institutional digital collections.

2.2. Text Analysis and Text Collections

This section introduces the concept of text analysis and its role and use in the development of DL and text collections. As shown in the previous section, the number of DLs and text collections grows everyday. The urgency for tools that help us to deal with them also grows. As text analysis seems to be an essential ingredient of these tools, this section reviews its use and the opportunities that text analysis provides when working on DLs and text collections. The review starts with an introduction where text analysis is defined and established methods are listed and described, while the following section concentrates on the role of text analysis and metadata in cataloging digital objects. Five major digital libraries are compared in their use of methods to catalog collections. Finally, this section discusses the concept of data and its complexity and subjectivity, in contrast with the simplification that often arises when using standard metadata properties to represent the rich contents of digital collections.

When first encountering a specific collection of textual documents, before any analysis takes place, the kind of metadata properties expected for a textual document could be title, author, dates, locations, etc. These fields are standard metadata fields and are the primary source for cataloging collections. Digital libraries can readily increase the number of indexed items in their databases due to the easy and inclusive metadata formats used. This is the case in all five DLs reviewed (see table1). he use of standard fields of metadata brings homogenisation, but it only provides a glimpse of the abundance and complexity of the collection. Metadata inevitably brings a simplification or summary of the content. Standard metadata simplifications were primarily designed to support search and retrieval. An example of a collection of texts represented almost exclusively by metadata is "Mapping the republic of letters", a meta-project to visualise letters from 1600 to 1800 by recognized intellectuals (Findlen et al., 2011). The visualisations listed in the project represent the metadata of the letters, such as dates, locations, senders, and recipients. However, in most cases there is no access to the letters themselves.

Beyond the metadata there is the analysis of the documents of the collection with procedural (i.e., computer-aided) text analysis. Procedural text analysis is a key element of information retrieval (Liu, 2009, Salton and McGill, 1986), data mining (Han et al., 2011, Kantardzic, 2011), natural language processing (Manning and Schütze, 1999) and computational linguistics (Grishman, 1986, Hausser and Hausser, 1999), Today, most of the methods used in those fields use techniques grouped under the label of machine learning (ML).

Established and developed text analysis tasks include:

• Text categorisation: this is the technique of assigning documents of a collection to two or more predefined categories. A typical example of text categorisation is to classify emails as spam or not-spam. This is a typical ML task. An example and study of a visualisation of document collection classification is Di Munzio's work (Di Nunzio, 2006).

- Text clustering: a generic name for a variety of techniques that deal with the task of grouping documents according to their similarity. See the following discussion for examples of text clustering.

- Entity and concept extraction: this is part of the more general field of information extraction.It includes the task of identifying and interpreting meaningful parts of a text. These parts can be names —entity recognition extraction— or concepts defined in a customised or pre-existing thesaurus (Tseng, 2002).

- Taxonomy induction: a techniques for organising terms of a document in a hierarchical way using external lexical resources (Fountain and Lapata, 2012). The most used resource in this area is WordNet (Fellbaum, 1998). Other resources such as Wikipedia (Ponzetto and Strube, 2011) have been used.

- Sentiment analysis: the analysis of the attitude of the author of a message in relation to one or more topics. It is also known as opinion mining (Leetaru et al., 2013, Pak and Paroubek, 2010).

- Document summarisation: the process of shortening a text while showing the main points of the whole text (Jones et al., 2002, Gong and Liu, 2001).

To realise the huge evolution of the field, notice that in 1999 text analysis was limited to text categorisation and clustering, information extraction, and summarisation. Clustering of documents was used for document visualisation and categorisation (Tan et al., 1999). In this dissertation, we focus on the most popular text analysis techniques applied to visualisation of collections of texts, which are text clustering techniques. Such techniques can produce multidimensional properties of the objects of the collections, and this generates a richer classification, and therefore, a wider possible range of interpretations of the collection.

Text analysis is a broad concept that includes multiple techniques, methods and fields. In order to decide which aspects of the text analysis in relation to collections of texts to review, three dependent levels of text analysis in collections can be defined:

- Level I: Raw metadata. Text collection without text analysis, that is, using only standard metadata properties.
- Level II: Extended metadata. Text collection with extended metadata. This process enriches the initial metadata generating derived properties, e.g. from the fields "age", to generate "groups of age". This case can include some analysis of the object, e.g. counting the length of the text of each document.
- Level III: Object analysis. Text collection with text analysis of the textual documents of the collection. This process generates new properties, e.g. a list of topics for each document, a classification of documents according to contents, authorship analysis, recommendation system according to text contents or style, etc.

In the cases studied, and in general, in institutional DL there is a lack of application of techniques from the trending field of big data and data analysis, and in particular text analysis. To find projects that use text analysis as a technique that can help to understand and get to know a collection, and to build tools to explore and discover the collection, we need to go to independent data visualization practitioners and researchers. Section 2.3.3. Reviews data visualization and interfaces to text collections, and includes a list of cases that meet the criteria for level IIIn this dissertation, I focus on the most popular text analysis techniques applied to visualisation of collections of texts, which are text clustering techniques. Such techniques can produce multidimensional properties of the objects of the collections, thereby generating a richer classification, which in turn results in a wider range of possible interpretations of the collection.

Text analysis is a broad concept that includes multiple techniques, methods, and fields. In order to decide which aspects of the text analysis in relation to collections of texts are to be reviewed, three dependent levels of text analysis in collections can be defined:

- Level I: Raw metadata. Text collection without text analysis. I.e., only using standard metadata properties.
- Level II: Extended metadata. Text collection with extended metadata. This process enriches the initial metadata by generating derived properties. E.g. ”groups of age” from the field ”age”. Extended metadata can include some analysis of the object like, for example, counting the length of the text of each document.
- Level III: Object analysis. Text collection with text analysis of the textual documents of the collection. This process generates new properties like a list of topics for each document, a classification of documents according to contents, authorship analysis, recommendation system according to text contents or style, etc.

In the cases studied, and in general, in institutional DLs there is a lack of application of techniques from the trending field of big data and data analysis, and in particular regarding text analysis. To find projects that use text analysis as a technique that can help to understand and get to know a collection, and to build tools to explore and discover the collection, it is necessary to refer to independent data visualisation practitioners and researchers. Section 2.3.3. Reviews data visualisation and interfaces to text collections and includes a list of cases that meet the criteria for level III. .

Reviewed site	Level	Comments	URL
Europeana	Aggregator: level depend on the source	An aggregator that indexes millions of online resources.	-
State Library of New South Wales	I	Interface for reading It offers crowd-transcription of documents	http://transcripts.sl.nsw.gov.au/ project/World%20War%201%20Diaries
Internet Archive	I and II	Complete interface for reading	https://archive.org/details/MBLWHOI
Trove	I and II	Indexes millions of online and offline resources. newspapers collection is full text searchable. It offers crowdsourced transcription of documents.	http://trove.nla.gov.au/newspaper/article/ 13279967?searchTerm=&searchLimits=l-australian=y
Library of Congress (USA)	I and II + transcriptions	Newspapers collection ise fulltext searchable. Complete interface for reading	http://chroniclingamerica.loc.gov/lccn/ sn83045389/1916-04-03/ed-1/seq-1/
Non institutional projects	I, II and III	There is a gap on the use of Text Analysis techniques in institutional digital collections.	See 2.3.3 Data visualization and interfaces

Table 1. Comparison of five major digital libraries according to how they use text analysis. None of compared cases meet the criteria for level III.

Traditional natural sciences define data as a representation of reality with a quantified, and usually controlled, error of measurement. The results, with natural sciences methods, are expected to be objective. The concept of data, however, is less deterministic when considered from a humanities point of view. Accordingly, Drucker (2011) differentiates between data and capta. Since data, from the Latin ”given”, assumes a passive position of the observer studying the real phenomenon, capta, from the Latin ”taken”, places the observer in an active position, being able to interpret the measurement and, therefore, critically extract conclusions. Drucker says: ”From this distinction, a world of differences arises. Humanistic inquiry acknowledges the situated, partial, and constitutive character of knowledge production, the recognition that knowledge is constructed, taken, not simply given as a natural representation of preexisting fact". Text analysis treats the document as data and attempts to use computational techniques to reveal features of the texts. The complexity of the data, its subjectivity, instead of being an obstacle in this research project, is a motivation for the review of cases that use procedural methods, as well as curated ones, for the text analysis of the collection. Rather than treat the results of text analysis as "data", which objectively represent the analysed text, Drucker's concept allows us to frame them as "capta" for critical human interpretation.

2.2.1. Text Clustering and Interfaces

Text clustering is a text-analysis technique that groups documents according to their similarity, whereby the degree of similarity is estimated on the basis of some defined properties. Representations of collections of texts with clusters are very common, in 2D (Weskamp, 2004, Paulovich and Minghim, 2008, Masad and Nayar, 1011, Lagus et al., 2004), and in 3D as well (Chalmers and Chitson, 1992, Wise et al., 1995, Carter and Capretz, 2003, Andrews et al., 2002).

The most popular technique to measure the similarity among documents is cosine similarity. Every document belonging to a collection of textual documents can be represented by a vector using methods based on term frequency-inverse document frequency (TF-IDF). TF-IDF is a statistical value intended to score the importance of a word within a single document in a collection. In essence, it provides a quantitative measure of the occurrence of specific words (relative to the document as a whole). By comparing significant words from different documents it is possible to generate a measure of document similarity (Lee et al., 2005). This analysis generates a network of documents in which the cosine of the angle between two document-vectors is a measure of the similarity of the two documents.

The similarity among the documents of the collection generates a network of relationships, and visualisations of sch networks include Andrews's work InfoSky (Andrews et al., 2002), which proposes a space travel metaphor whereby planets are documents and proximity among them is proportional to their similarity. Today the use of TF-IDF techniques is related to big data processes (Leskovec et al., 2014). Another example of similarity broadly understood, is the networks of chapters in

Grimm's Fairy Tale Network, by Jeff Clark (2009), . Similarity among documents can sometimes be seen in comparing representations of the documents. This is the case of Word Storm by Castella and Sutton (Castella and Sutton, 2014), in which each paper from a conference is represented by a word cloud. The algorithm that places most frequent words in a 2D area has been modified in order to place recurring words in the same location, making it easier to compare word clouds. Methods to calculate document similarity are used in information retrieval systems for the task of term weighting, which, according to Zhang et al, ”is the job to assign the weight for each term, which measures the importance of a term in a document” (Zhang et al., 2011). Examples of the visualisation of search results similarity through 2D network diagrams include Kartoo (Baleydier, 2001), and Touchgraph (TouchGraph, 2001). Since Kartoo and Touchgraph were commercial projects, the techniques used to measure similarity have not been published.

2.2.2. Topic Models

Topic modelling is a group of statistical methods that analyse collections of textual documents by extracting the thematic composition of each piece of the corpora. A topic model is represented by a set of words that have high probability of appearing together in a document within a corpus of textual documents. These words, usually called terms, ”look like topics because terms that frequently occur together tend to be about the same subject” (B(Blei, 2012). The number of topics that the methods generate is usually set as a parameter to the analysis process. Each document gets a normalised score for every topic. Topic model algorithms are not based on semantic analysis, but purely statistical co-occurrence. In machine learning terms: ”topic models offer an unsupervised, data-driven means of capturing the themes discussed within document collections” (Aletras et al., 2014).

Topic models were introduced by Blei, Ng, and Jordan in 2003 as a ”generative probabilistic model for collections of discrete data such as text corpora” (Blei et al., 2003). LDA developed from previous works include the early work on Latent Semantic Indexing (LSI) (Deerwester et al., 1990) and the probabilistic Latent Semantic Indexing (pLSI) approach by Hofmann (1999).

Topic models can be used in many ways, and documents can be grouped on the basis on combination of their topics. Part of the collection can be filtered and selected for further reading according to specific topics. One example of grouping documents according to their similar topics is the Stanford Dissertation Browser ((Ramage and Chuang, 2012) where an interactive visual browser allows the exploration of theses, showing similarities among different domains of knowledge.

The relationship between topic models and digital humanities (DH) is an example of interdisciplinary collaboration between two relatively young research communities. An influencer of this collaboration is Franco Moretti and his distant reading tools. As a general strategy, Moretti recommends that literature researchers read less and, instead, represent more (Moretti, 2005). In this way, the texts can be studied from overview representations, timelines, graphic comparisons, and story diagrams, among other possibilities. Topic model methods fit very well into Moretti's idea, and have been used in the humanities extensively, as described in the following paragraphs.

Literature and scientific texts have been a perfect terrain for experimentation with topic models in recent years. There are currently tools (e.g., Mallet) easy to use for non-experts in math and statistics particularly suitable for DH researchers, who have criticised the complexity, the interpretation, and in some cases the value of topic models. For example Lisa M. Rhody, who analyses the behaviour of topic model analysis in poetry, finds that the reduction of the complexity of figurative texts (such as poetry) caused by topic model analyses demonstrates the very complexity of the source texts. She suggest that ”topic modelling poetry works, in part, because of its failures”, and talks about ”an interpretive space (...) between the literary possibility held in a corpus of thousands of English-language poems and the computational rigour of Latent Dirichlet Allocation (LDA)” (Rhody, 2012). Schmidt in is ”Words Alone: Dismantling Topic Models in the Humanities” (Schmidt, 2012) was quite sceptical of the use of topic models by humanists, and argued: ”simplifying topic models for humanists who will not (and should not) study the underlying algorithms creates an enormous potential for groundless — or even misleading — insights.” (Schmidt, 2012).Goldstone and Underwood have independently applied topic models to the study of the history of literary studies. They used different software, stop-word lists, and numbers of topics. The results have overlapped and diverged in different places, demonstrating the non-universality of the methods. Despite the results that repeatedly indicate the complexity of the interpretation of topic model analyses, they ”reached a shared sense that topic modelling can enrich the history of literary scholarship by revealing trends that are presently invisible” (Goldstone and Underwood, 2012).

Today, topic model applications in DH, and to text collections is supported and recommended as a powerful tool to analyse large collections of documents. Recent works, like Jockers's Macroanalysis: Digital methods and literary history (Jockers, 2013) are optimistic about the use of computing analysis, and advocate the revolutionary potential of large-scale literary analysis. Two cases of topic model analysis applied to text collections are discussed in 2.2.3: ”Mining the dispatch”, by Nelson (Nelson, 2010a), and ”Topic Modeling Martha Ballard's Diary” by Cameron Blevins (Blevins, 2010)

One more remarkable case of topic model exploration and management is MetaToMAT (Metadata and Topic Model Analysis Toolkit), by Snyder et al. In their words, this approach is ”a visualization tool that combines both metadata and topic models in a single faceted browsing paradigm for exploration and analysis of document collections” (Snyder et al., 2013).

The efforts to make topic modelling available to a community of non-experts is a key point in the growing use of these techniques by the humanities community. There are several free and commercial software tools that calculate, represent, and manage topic models. To mention the most popular ones: Mallet (McCallum, 2002), LDAvis (Sievert, 2015), dfr-browser (Goldst, 2014), Termite (Chuang et al., 2012), Serendipity (Alexander et al., 2014). Most of these tools also offer representations of the topic models. They represent the relative scores of each topic, their distribution in time, and the comparison of terms in topics.

2.2.3. Topic Model Labelling

Topic model labelling is a method that is used to complement topic model analyses, making topics human-readable. As mentioned, a topic model is an abstract statistical construct that may or may not be equivalent to a “topic” or a theme for a human reader. Therefore, in order to be interpreted and used by humans, topic models need effective representations, adapted to the nature of the collection and oriented to specific tasks. One way to represent a topic model is through the number of terms that co-occur frequently along the documents of the collection. As shown in Figure1 the list of topics that a topic model software such as Mallet outputs by default, have no label and simply identified as topic1, topic2, topic3, etc. Efforts in improving this basic representation of topic models have been published and research on the issue is currently being actively pursued.

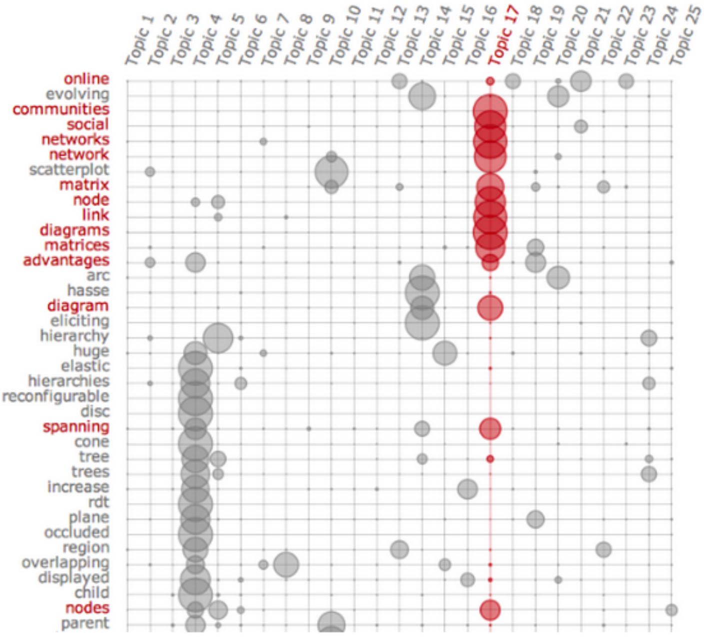
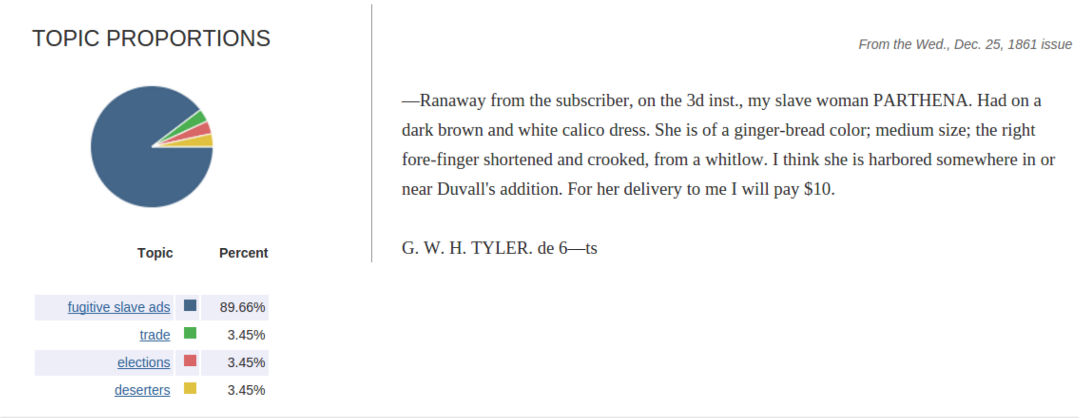


Figure 1. Topic visualisation with “Termite”. By default the generated topics are named: topic1, topic2, topic3,. . .

Aletras et al (2014), show in their evaluation studies that text based labels are easily interpreted by humans. Text labels, which are more effective than image labels, can consist of a list of terms, a single top term, a phrase, or a sentence (Mei et al., 2007).

Several methods for text labelling of topic models have been published, including procedural and curated methods. Procedural methods, seeking to assign labels to topics automatically, are used in cases where the number of topics generated is large, and where the whole process is automated. Some methods recombine the n-terms and then re-apply the model, getting a ranked list of terms within a topic; then the highest ranked term can be used as label for the respective topic (Lau et al., 2010). Other methods involve external data sources, like in the case of Lau et al (Lau et al., 2011)), who obtain text labels from the n-terms, from titles of Wikipedia articles containing the terms, and from sub-phrases extracted from the Wikipedia article titles. Since topics can overlap with one another, some works produce labels from topic hierarchies based on parent-child relationship between topics (Mao et al., 2012). The mathematical review of these procedural methods is beyond the scope of this dissertation.

Curated methods of labelling topics (i.e., human generated labels) are common in DH, where experts can label according to a defined goal (Schmidt, 2012). Some authors from science consider the subjectivity of curated labelling as a problem that ”can easily be biased towards the user’s personal opinions” (Mei et al., 2007). Meeks (2011)) comes to the conclusion that small corpora of texts produce broad topics. Since he is interested in topic networks, a bigger corpus creates networks too complex to be visualized in 2D. He finally assigned topic labels asking a group of experts for ”a comparison of labelling of topics (...) based on their interpretation of the topic’s connection to various words, as well as to various papers”.



TOPIC PROPORTIONS

Topic		Percent
fugitive slave ads	🟦	89.66%
trade	🟩	3.45%
elections	🟥	3.45%
deserters	🟨	3.45%

From the Wed., Dec. 25, 1861 issue

—Ranaway from the subscriber, on the 3d inst., my slave woman PARTHENA. Had on a dark brown and white calico dress. She is of a ginger-bread color; medium size; the right fore-finger shortened and crooked, from a whitlow. I think she is harbored somewhere in or near Duvall's addition. For her delivery to me I will pay $10.

G. W. H. TYLER. de 6—ts

Figure 2. “Mining the dispatch” by Robert K. Nelson (2010). Detail of one news text of the collection and its topics pie chart.

There are two DH projects that inspired my research: ”Mining the dispatch” and ”Topic Modeling Martha Ballard Diary”. ”Mining the dispatch”, by Nelson

(Nelson, 2010a) There are two DH projects that inspired my research: ”Mining the dispatch” and ”Topic Modeling Martha Ballard Diary”. ”Mining the dispatch”, by Nelson (Nelson, 2010a), is a topic model analysis of a collection of news texts from the U.S. Civil War during the years 1860-65 in Richmond, the capital of the Confederacy’s newspaper of record (see Figure 2). This project shows line charts graphing the change in topics over time. Nelson explains how the labelling of topics was done, and his impressions about the results: ”I have given each topic a label based upon my reading of pieces drawn from that category; these labels are informed judgement calls and are imperfect”. For each topic, the interface offers a numbered list of excerpts of news texts ordered by relevance. One more click on each excerpt shows the full text of the article.

.

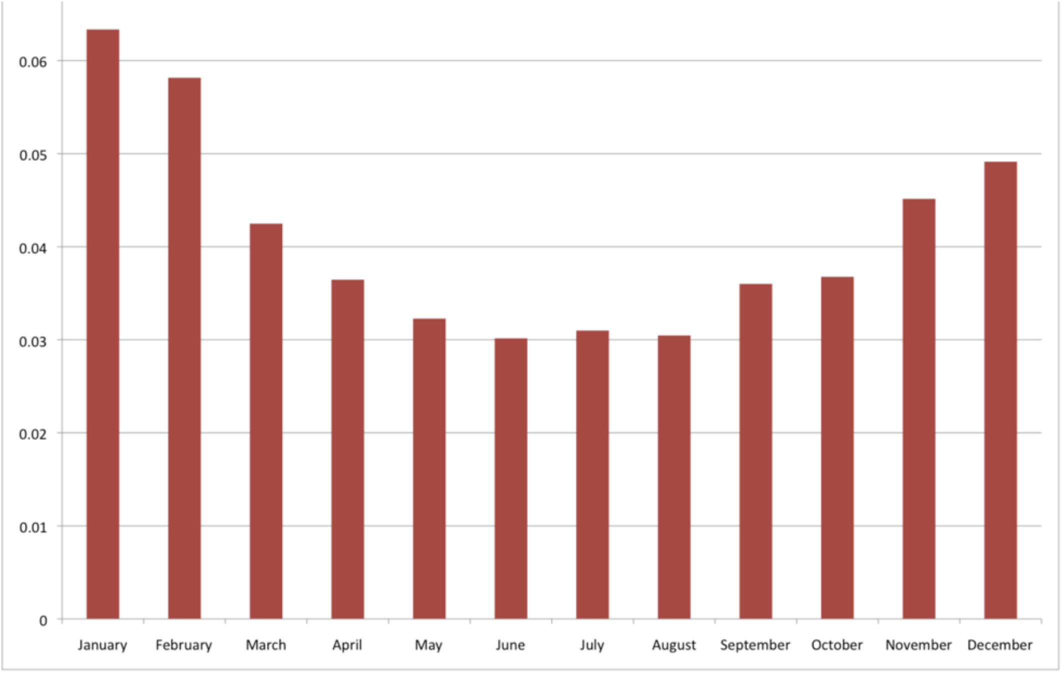


Figure 3. Scores for topic “cold weather” grouped by month, from “Topic Modeling Martha Ballard’s Diary”, by Cameron Blevins (2010).

”Topic Modeling Martha Ballard’s Diary”, by Cameron Blevins (Blevins, 2010) analyses the diaries of Martha Ballard, a New England midwife who kept a daily diary for over twenty-seven years beginning from 1785. The collection consists of one thousand four hundred handwritten pages. The topics generated are manually labelled with “descriptive titles” such as: midwifery, church, death, gardening, shopping, illness, cold weather, etc. This project outputs charts with topic evolution, and provides striking evidence of the effectiveness of topic modeling in representing document content (see Figures 3, and 4).

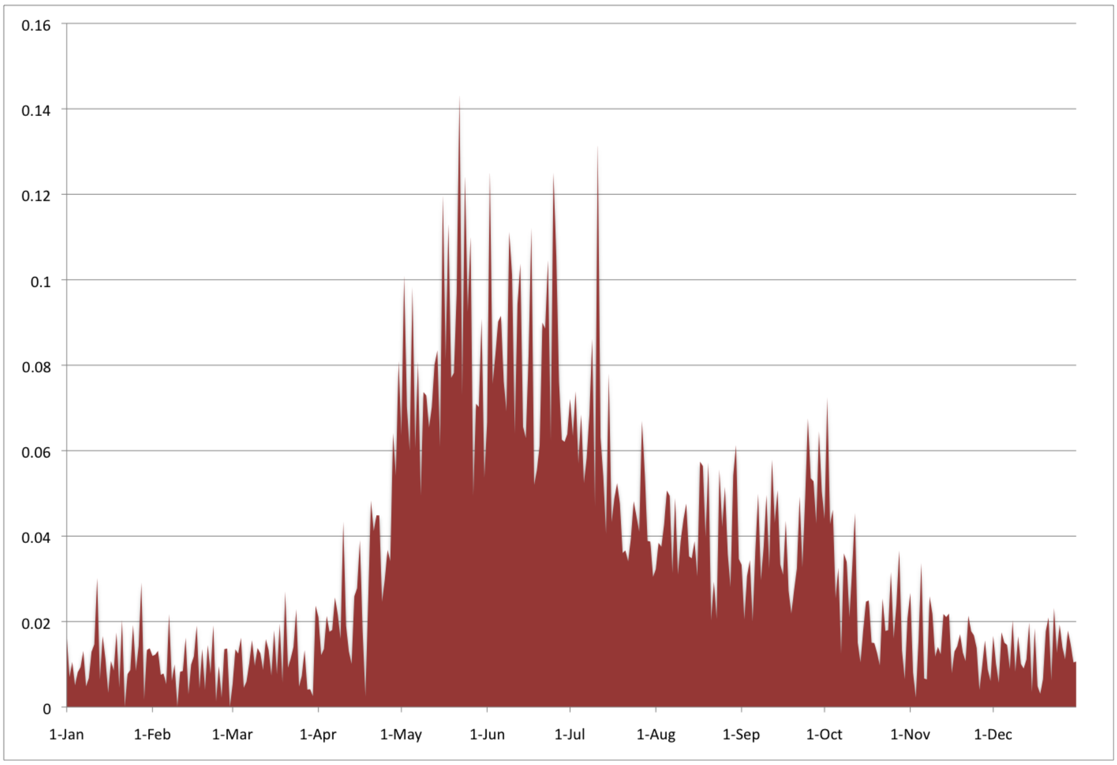


Figure 4. Scores for Topic gardening by day, from ”Topic Modeling Martha Ballard’s Diary”, by Cameron Blevins (2010).

2.2.4. Topic Model Evaluation

Topic model interpretation is a not well-defined task. According to Schmidt (Schmidt, 2012), topic models are ”like words, they are messy, ambiguous, and elusive”. Therefore it is difficult to establish methods to evaluate and compare the quality of topic model analysis. A number of questions arise initially: How good are the results of topic model analysis? How might we compare different analyses? How might we measure their consistency from a human point of view? All these questions guide research in topic model evaluation.

In DH we find that evaluation of the results of topic model analysis is undertaken by experts in the contents of the analysed documents. This is the case of Goldstone and Underwood, who compare their independent analysis of the same corpora and validate, as experts, which topic models diverge and converge in a more appropriate way (Goldstone and Underwood, 2012). There are also cases where evaluation consists of finding evidence in the corpus; this is the case of Ballard's diary, where entries related to cold weather are higher in winter, and, complementarily, entries in the diary talking about gardening are higher in summer (see Figures 3 and 4). These charts validate the use of Mallet for topic model analysis, as the author states (Blevins, 2010).

Computer-generated topic model labels can be evaluated by humans. This is the case of Newman et al, who in 2010 published a scoring model that predicts human scores. The experiment produces a ranked list of terms of a topic, where the best word is used as a label for representing the topic. This list is evaluated by by asking humans do rankings that are then compared with those automatically generated. The comparison is based on point-wise mutual information (PMI) of word-pairs —PMI measures the association between two words according to large text corpora. The model uses external sources such as Wikipedia, Google n-gram data set, and Medline to adjust the initial calculated ranking of terms (Newman et al., 2010).

Visualisation tools are used to overview and evaluate topic models, usually by visual comparison, like Termite (Chuang et al., 2012), or with a dashboard that allows multi-comparison and editing of the topics on-the-fly, like Serendipity (Alexander et al. 2014). Despite these and other works, the subjectivity of topic models when evaluated by humans makes it difficult to establish reliable measures of topic model quality.

This section has shown that while text collections are increasingly available as full text documents, current DLs use only basic metadata to represent and provide access to them. The text analysis techniques reviewed here can be applied to characterise large collections. The main text analysis methods with examples were mentioned, with particular attention to text clustering, as a general technique to group similar documents, followed by a more in depth description of topic modelling, which is the prominent modern technique for text analysis of large collections of textual document in the humanities. The success of topic models is linked to the process of labelling the topics, thus making them human-readable. This has encouraged innovation in evaluation of topic model labelling methods. In some reviewed experiments, procedural labelling produced results comparable to those of curated labelling, although it must be noted that in DH the reviewed cases were curated by experts. The application of topic models showed a gap: most of the reviewed cases use the output of the analysis to practise "distant reading"; that is, to represent the collection as a whole. The use of topic models in textual document collection interfaces to help exploration and reading is not common in the reviewed cases.

The next section presents a review of interfaces to text collections. It includes a review of paradigms in interface design strategies, a review of standards and trends in in-production interfaces to text collections, a review of the role of data visualisation in interfaces to text collections, and a brief review of interfaces where the main aim is to assist in reading textual documents.

2.3. Interfaces and Text Collections

The development of systems to store, catalogue, and interact with digital collections has outstripped the development of user interfaces. In recent decades, the interfaces to text collections in the domains of cultural heritage and scientific publication have not developed much, especially when compared to commercial interfaces, such as online shopping, social network, and even online banking sites (Mario Perez-Montoro and Jaume Nualart, 2015).A broad definition of interfaces to text collections includes any software and hardware that connects humans with digital data. For the sake of this research project, the expression "interfaces to text Collections" refers to software mostly developed as web applications accessible online. Since this is a very big field, this section presents reviews of four interrelated topics: interface paradigms, standard and trending practices in text collection interfaces, data visualization, and interfaces for reading.

Firstly, several interface paradigms or metaphors that initiated strategies for innovation in interface design for the last twenty years are reviewed. This review will help the reader understand the approach of the artefacts and their innovation value. Secondly, standard practices in contemporary interfaces to text collections are reviewed. Several examples of image collections are included, as inspiration and a source of applicable ideas for text collection interfaces. Thirdly, , data visualisation approaches used as interfaces to text collections is reviewed, to help in defining its role in modern interface design. Lastly, due to the importance of textual documents in this dissertation, this section concludes with the review of what is called “interfaces for reading”. That is, interfaces that bring textual collections closer to their potential readers.

2.3.1. Interface Paradigms

This section reviews general strategies and concepts to advance the design of interfaces to text collections and digital libraries. It includes selected works that are relevant to the topic and have influenced this research project, especially during the conceptual development of the artefacts.

Probably, the most influential work in interface development and evolution for the last twenty years is ”The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations” by Ben Shneiderman —which has more than 3,500 citations in the literature (Shneiderman, 1996), is Visual-Information-Seeking Mantra was: “Overview first, zoom and filter, then details-on-demand”. In Schneiderman’s words: ”To sort out the prototypes and guide researchers to new opportunities”. He proposed ”a type by task taxonomy of information visualizations” for collections of items, with multiple attributes each. These tasks are: overview the collection, zoom it, filter it, get an item’s details-on-demand –included in his mantra–, get related items, provide access to the exploration history, and make possible the extraction of subsets of the collections.The mantra was presented as a design principle based on Schneiderman’s experience in ”information collections” and libraries.

Following the direction taken by Schneiderman, in 1998 Marchionini et al (Marchionini et al., 1998) analysed early interface developments in a collaboration between University of Maryland designers and staff from the Library of Congress. The goal was: “users should maximise their interactions with information resources and minimise their attention to the system itself”. Three kinds of tools were developed: “Overviews of collections, object previewers, and object gatherers”. The interface design was based on simplicity and clearness: “A standard toolbar on the left of the screen lists the functions available”. The justification for the necessity of interfaces was defined in relation to the size of the collections: “very large amounts of materials in many different formats with varying levels of descriptive metadata makes searching difficult and browsing more important"’. In these early years some structural elements were introduced: hierarchical table of contents in sidebars, navigation bars, and quick and advanced search forms.

In 2000, Greene, in collaboration with Schneiderman and others, (Greene et al., 2000) established a methodology to apply the idea behind Schneiderman’s Mantra. With the aim of aiding ”designers of digital library interfaces”, they ”present[ed] a framework for the design of information representations in terms of previews and overviews”, as well-delimited elements to be used as pieces and features of the interface. Previews were defined as an analogy to bibliographic records, to act ”as a surrogate for, a single object of interest”. A bibliographic record suggests text —but a preview, rather than textual, can be a thumbnail image or other representation of the object: ”An effective preview is an information surrogate that communicates to the user, at the appropriate time, sufficient information about the primary object it represents to support users in making a correct judgement about the relevance of that object to the user’s information need”. Overviews, as opposed to previews, were seen as analogous to catalogues. In that sense Greene et al, supported by Marchionini’s idea that information seekers ”engage to change their knowledge state” until they have satisfied their information needs, considered previews and overviews as representations that supported browsing and scanning for exploration. “Previews and overviews” from this point of view, can be seen as a formalisation and as an elaborated development of Schneiderman’s previous works, such as his Information Seeking Mantra.

In 2009 a new paradigm was introduced by Stamen Studio, an influential team of practitioners, who argued that an overview could include all the items of the collection. This idea is known by the expression “show everything”. The items are not hidden any more and the visitor gets an idea of the whole collection, as opposed to a tiny search box that initially hides the collection. Despite the assertion contained in “show everything”, when the collection is too big to fit technically and/or perceptually on the screen, alternatives are required (such as zoom, and filters), which introduce intermediate states between Greene's previews and overviews. Stamen's slogan is seen as a combination of pragmatism (Whitelaw, 2015b) and provocation (Whitelaw, 2015a). Beyond what Shneiderman and Greene provide, ”show everything” is significant in that it introduces this idea into a contemporary web context and begins to show how rich collection interfaces can be within a modern web browser.

The combined Shneiderman-Stamen paradigm could be seen metaphorically as a walker who arrives to the top of a hill (the home page) and from there can see a valley (the overview) and decide where to go and which paths to take (the details).

In 2011, Dörk et al introduced the paradigm ”Information Flaneur” (Doerk et al., 2011).They presented ”explorability as a new guiding principle for design and raise research challenges regarding the representation of information abstractions and details”. The information flaneur proposes to break the rigid axes of information seeking systems based on the concept of offices and corporate buildings by elevating the traditional hierarchy of files and folders to a new perspective using the city as the metaphor for the information seeker. The seeker —the flaneur— is motivated by curiosity and explores the city following emotions and personal experience. This model reframes the users by giving them different motivations, rather than tasks and functions. The Information Flaneur emphasises individual interpretation and subjective experience, and changes the criteria for ”success” in interface design by recognising that there is more than ”task” and ”information retrieval” when accessing a collection. This idea moves complexity, individual interpretation, and subjectivity to the centre of the design.

From these ideas evolved the concept of Generous Interfaces (Whitelaw, 2015), which refers to interfaces that show the abundance of digital collections, in contrast with the classical interface that “is ungenerous”, “withholds information, and demands a query”. A Generous Interface emphasises browsing and visual exploration, and supports experimentation with visual elements. Most of the cases presented in Whitelaw's paper are applied to digital collections of images. Nevertheless, the ideas and artefacts Whitelaw introduces are applicable to any collection of digital objects, including text collections.

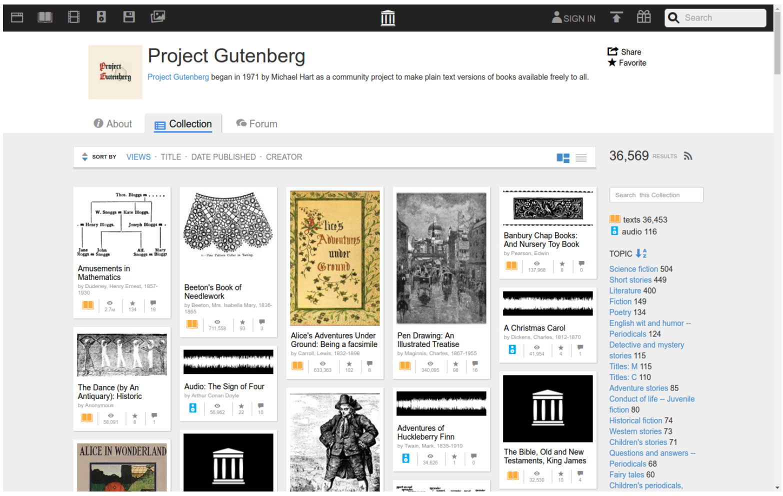
This review shows the conceptual development of trends and strategies in interface design that, throughout the works of the reviewed authors, seems to follow the same path, from Schneiderman's "Mantra" to Greene's "Overview and previews". These efforts shed light onto the beginning of the digital era and the evolution of information displays, which began twenty years ago, and provide initial proposals that function guidelines for researchers and designers of interfaces to DLs. The paradigms of Stamen, Dörk, and Whitelaw incorporate subjectivity and challenges and are less explicit, but more flexible, in what they propose. Far from eluding complexity, this brief journey through paradigms and metaphors —Schneiderman, Greene, Stamen, Dörk, Whitelaw— lead us to a point I will develop in chapter 6.Discussion. Namely, that too big, too complex, or too undefined does not necessarily imply a complex interface, with associated problems in explaining how to use a new feature or, ultimately, a whole new interaction concept. Simplicity in the interface does not require simplification of collection contents, but rather a simplification of visual language and interaction features. The complexity and diversity of cultural collections, as defined in generous interfaces, is what should be revealed through interfaces.

2.3.2. Standard and Trending Practice in Text Collection Interfaces

This section discusses a list of what can be considered standard practices in DL and, more specifically, in text collection interfaces. Since approximately 1990, interfaces to digital collections have developed in parallel with digital libraries. Therefore, to review interfaces to textual document collections, it is necessary to review interfaces to DL in general because, as it will be shown, there are no significant differences between DL and text collections when traditional, text-based interfaces are used.

The DLs maintained by institutions, archives, and libraries that are reviewed in this section mostly provide a text-based interface. Text-based means that the interfaces to access the collections are standard text pages that always include (Tedd and Large, 2004) navigation and search systems (Mario Perez-Montoro and Jaume Nualart, 2015). . Navigation is a simple tree of categories and subcategories, and the search system is a traditional full text query with the option of advanced search that includes a set of filters. The results are flat lists of items with a short description of each item and sorted according to the categories (that is, by date, name, etc.) or to search system ranking, which are usually not communicated to the user (Mario Perez-Montoro and Jaume Nualart, 2015).

Archive.org, one of the referents for online non-profit DLs (archive.org), has introduced facetted lists of collections (for textual documents and other multimedia collections alike) and modern graphic design to prepare for searches. The facetted lists 5) are rudimentary: sizes depend on to the length of the titles but not on the size of each collection or section of collection. Only a basic by-type (video, image, text, audio) filter is generalised. No overviews are offered, neither metadata visualisations or visual analysis. A facetted view is a kind of ”poor” overview which reveals the contents of the collection but in a very limited way.



Project Gutenberg

Project Gutenberg began in 1971 by Michael Hart as a community project to make plain text versions of books available freely to all.

36,569 RESULTS

Amusements in Mathematics
by Dudeney, Henry Ernest, 1857-1930
2.7M 134 18

Beeton's Book of Needlework
by Beeton, Mrs. Isabella Mary, 1836-1865
711,558 93 3

Alice's Adventures Under Ground: Being a facsimile
by Carroll, Lewis, 1832-1898
633,363 102 8

Pen Drawing: An Illustrated Treatise
by Maginnis, Charles, 1867-1955
340,095 98 16

Banbury Chap Books: And Nursery Toy Book
by Pearson, Edwin
137,968 8 0

A Christmas Carol
by Dickens, Charles, 1812-1870
41,954 4 1

The Dance (by An Antiquary): Historic
by Anonymous
58,091 8 1

Audio: The Sign of Four
by Arthur Conan Doyle
56,962 22 10

Adventures of Huckleberry Finn
by Twain, Mark, 1835-1910
34,626 1 0

The Bible, Old and New Testaments, King James
32,530 10 4

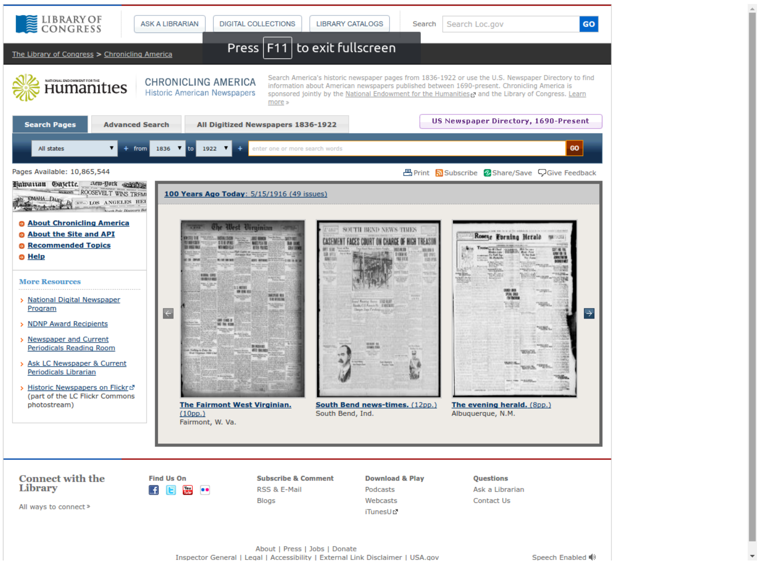
texts 36,453
audio 116

TOPIC

Science fiction 504
Short stories 449
Literature 400
Fiction 149
Poetry 134
English wit and humor -- Periodicals 124
Detective and mystery stories 115
Titles: M 115
Titles: C 110
Adventure stories 85
Conduct of life -- Juvenile fiction 80
Historical fiction 74
Western stories 73
Children's stories 71
Questions and answers -- Periodicals 68
Fairy tales 60
Children's periodicals,

Figure 5. Screenshot of Project Gutenberg interface at of archive.org.]

The Library of Congress offers several digital collections that range from performing arts to legislative texts. The collections that include textual documents are: Historic Newspapers, United States Legislative Information, and Web Site Archiving. All of them offer a search box with filters, and a glimpse of the collection. For example, in the case of historical newspapers, the home page of the collection includes images of the newspapers of the last one hundred years (see Figure 6). The search results are a facetted list that includes images of the newspaper page with the query highlighted in it.



LIBRARY OF CONGRESS
ASK A LIBRARIAN
DIGITAL COLLECTIONS
LIBRARY CATALOGS
Search
Search Loc.gov
GO
Press F11 to exit fullscreen
The Library of Congress > Chronicling America
NATIONAL ENDOWMENT FOR THE Humanities
CHRONICLING AMERICA
Historic American Newspapers
Search Pages
Advanced Search
All Digitized Newspapers 1836-1922
US Newspaper Directory, 1690-Present
All states
+ from 1836 to 1922 +
enter one or more search words
GO
Print
Subscribe
Share/Save
Give Feedback
About Chronicling America
About the Site and API
Recommended Topics
Help
More Resources
National Digital Newspaper Program
NDNP Award Recipients
Newspaper and Current Periodicals Reading Room
Ask LC Newspaper & Current Periodicals Librarian
Historic Newspapers on Flickr
Connect with the Library
All ways to connect »
Find Us On
Subscribe & Comment
RSS & E-Mail
Blogs
Download & Play
Podcasts
Webcasts
iTunesU
Questions
Ask a Librarian
Contact Us
About | Press | Jobs | Donate
Inspector General | Legal | Accessibility | External Link Disclaimer | USA.gov
Speech Enabled

Search America's historic newspaper pages from 1836-1922 or use the U.S. Newspaper Directory to find information about American newspapers published between 1690-present. Chronicling America is sponsored jointly by the National Endowment for the Humanities and the Library of Congress. Learn more »

Pages Available: 10,865,544

100 Years Ago Today: 5/15/1916 (49 issues)

The Fairmont West Virginian. (10pp.)
Fairmont, W. Va.

South Bend news-times. (12pp.)
South Bend, Ind.

The evening herald. (8pp.)
Albuquerque, N.M.

(part of the LC Flickr Commons photostream)

Figure 6. Screenshot of the "Women's history" collection interface, at the Library of Congress (USA)]

The Europeana site (Europeana, 2008). adopts a different strategy. While the site itself uses conventional approaches, they are innovating via other platforms mainly through data sharing. Europeana is an aggregator and a meta-search engine that incorporates external resources in order to add features to its contents. The strategy to offer not only text-based navigation and search in Europeana has two main directions. On the one hand, Europeana uses external services like exhibitions at Google Cultural Institute and has active accounts in Pinterest, Facebook, Google+, and Twitter. On the other hand, Europeana maintains and promotes a rich API (Application Programming Interface) to encourage others to work with their materials. Europeana implements interface standards in facetted lists of digital collections and also promotes the reuse of content, as mentioned above, in commercial free services.

This year, 2016, the New York Public Library (NYPL) integrates standard and innovative ideas, for example by publishing its public domain collections in a revolutionary way: it created a Git repository that published the digital objects of their public domain collection of images. Furthermore, the NYPL regularly publishes a snapshot of the collection. That encourages people to work on the collections and encourages innovation. Like other institutions, the NYPL innovates through an internal group called NYPL Labs. To present the collection of multimedia objects, (more than 186,000 in the initial release), the interface presents the collections and the items of each collection as a standard facetted list (NYPL, 2016a). The NYPL also presents several experimental interfaces to the collections, like a long mosaic of small images representing 180,000 public domain images (NYPL, 2016b).

The NYPL has published several works regarding collections of textual documents. For example, ”Navigating The Green Book” (Foo, 2013) and ”Gutenberg authors” (NYPLlabs, 2015), which offers visual ways to explore the text and the authors of Gutenberg books. The “Green book” has a derived interface that shows the contents of the book by means of a map, as the book were a travel guide. With a more general scope, “The Networked Catalogue” is an interactive space-metaphor exploration tool of the NYPL catalogue by topics and similarities (REF). It uses only Library of Congress Subject Headings (standard descriptive metadata) but looks for co-occurrence of subjects in metadata items, which, combined with the number of items per category, creates a graph of relatedness between subject terms. Another remarkable application provided by the NYPL is the “NYPL Archives and manuscripts” network generator, which allows exploration by dynamically building a network of relationships (REF). Its interactive view offers access to each item. These NYPL initiatives offer promising ideas and prototypes that in future could be widely adopted.

This subsection has presented a brief review of what big DLs such as Internet Archive, Library of Congress, Europeana, and New York Public Library offer as interfaces to their digital collections, and especially to their collections of textual documents. The practices of these institutions are standard for text collection interfaces. The review also mentions more innovative proposals from the NYPL. The following subsection focusses on data visualisation approaches used for collections representation and exploration interfaces.

2.3.3. Data Visualisation and Interfaces to Text Collections

The simplest definition of data visualisation could be ”the science of visual representation of data” (Friendly and Denis, 2001). Indeed, any interface is a concrete representation of an abstract data structure (including, for example, a list of search results). As shown below, when building interfaces to digital collections, data visualisation can be used either as an element of an interface or as a full interface.

This review shows that visualisation techniques (designed and systematic relations between data features and visual features) are increasingly integrated into text collection interfaces in non institutional sites.

The case of the German digital library visualisation provides a remarkable example of data visualisation. The library, which hosts the cultural and scientific heritage of Germany in digital form, contains a constantly growing list of collections of all kinds of digital objects. At the time of writing, the collections include more than eighteen million objects. While the official collection site uses a standard text-based interface (DDB, 2016), researchers from Potsdam University of Applied Sciences have developed an experimental visualisation: ”Deutsche Digitale Bibliothek Visualized” (Bernhardt, 2015).This visualisation interface offers four options to explore the collection: Periods & Sectors, Keywords, Places & sectors, and Persons & Organizations, respectively based on timeline, tags, map, and network representations. The visualisation is based on the facetted metadata exposed by the DDB API (see ttp://infovis.fh-potsdam.de/ddb/). The proposed visualisation dynamically shows changes in the facets, which helps to show the contents of the collection. The ”navigation” aspect of this visualisation involves linking to the official collection and triggering a facetted search —returning a list of documents. As well as showing visualisation techniques, this example illustrates how the relationship between DL and interface is now flexible —that we can have multiple interfaces, and both official and experimental interfaces can co-exist online.

The above mentioned four options to browse the collections of the German digital library (DDB 2009) combined with the type of objects (Archive, Library, Media, Research, Museum, Monument protection, other), offer interactive data visualisation tools that are integrated in a standard interface—standard in the sense defined in section 2.3.2. The interface of this DDB visualisation does not need special directions or tips to be used for a first-time visitor because of the clear use of established conventions like timeline, word cloud, use of interactive widgets, and tool tips that help to explain the visualisation as the interaction with the system proceeds.The visualisation element actually occupies 90% of the page, but it clearly sit within a conventional web page structure (e.g. header navigation, introduction page, etc). It is also worth noting that the interface is delivered using web standards and does not require any special software at the user's end.

This practice of integration of visualisation elements can also be found in "Explore Australian Prints + Printmaking" (Ennis and Whitelaw, 2014). This project provides a representation of the prints collection of the National Gallery of Australia, which includes "54,158 works, 26,612 images, 19,949 artists, 3,081 galleries, 8,726 exhibitions and 9,090 references". As in the DDB visualisation, the site offers on the home page, in addition to the conventional search interface, several ways to discover the collection: by Subject, Timeline, Works and Networks, Decade Summary, and All Artists. As with the German digital library, elements and techniques of data visualisation, mostly based on metadata representation, are integrated into a web interface.

Another significant example of data visualisation to be found in the Eugenics Archives by the Social Sciences and Humanities Research Council of Canada (Collective, 2016). This rich archive collects information from Canada related to the practice of surgical sterilisation for those deemed "mental defectives", in order to improve the genetics of future generation. This practice was in effect until 1972. The home page of the archive offers "twelve interactive tools to explore this archive" (Encyc, World, Game, Connections, Our Stories, Timeline, Players, Institutions, Interviews, Pathways, Media, and Database) that "reflect the collaboration of scholars, survivors, students, and community partners in challenging eugenics". The range of options goes from pure narrative, to charts, videos, and network relationships.

According to the reviewed cases, the use of the home page of DLs of collections as a portal of interfaces is growing. The traditional catalog-based view of items is present in all cases, but next to it a list of possibilities is offered. In all cases an effort is made to avoid the necessity of an initial explanation of how to use and interact with the interface by using established and well-defined techniques. Keeping in mind this idea of data visualisation elements integrated within an interface, this section continues with a review of techniques for the visualisation of collections of digital objects, and in particular textual document collections.

Spot by Jeff Clark (2009) is a real-time representation of tweets grouped by

Banner (i.e., by similarity), Timeline, User, Word (word within related tweets), software used, and groups (tweets with the same words). This example again uses data visualisation elements integrated to provide several options to discover the collection. In this case the groupings come from the analysis of tweets (such as similarity among tweets and tweets related to a specific word) combined with their metadata (author, date, software client). This same idea of multi-representation can also be found in static, almost infographic representations of large collections of items, like the case of "X by Y" by Moritz Stefaner (2010), which presents almost 40,000 project submissions to the Prix Ars Electronica, from the early beginnings in 1987 up to 2009. The items are grouped in several ways (by country, topic, year, prize, and category of the prize), so that the overview of the collection is multidimensional at a glance.

Other strategies to represent collections of texts are not based on a list of data visualisation elements integrated within an interface, but a full proposal that becomes an interface by itself. This is the case of Newsmap by Marcos Weskamp (2004), which offers a dynamic treemap with the latest Google News ordered by categories (colours) and weighted by area sizes according to the number of related articles that the Google News aggregator includes as relevant. This representation of the collection of latest news is dynamic and helps to show the changes in the stream of news.

Another example of data visualisation as an interface is the case of Grimm's Fairy Tale Metrics by Jeff Clark (2013). Clark represents the sixty-two stories of the "Grimms's Fairy Tales" by designing a multi-sortable table that allows comparison of the tales in several ways: physically (by length and lexical diversity) and by topics (analysing frequencies of keywords). Each cell gives a graphic score (represented with a bar) of the respective property. This is a powerful tool to compare documents, in this case from the domain of literature.

In other cases the strategy is not innovative in the presentation of the visualisation, so the users can certainly understand the interface with no need of introductory explanations. This is the case of Word Storm by Quim Castella and Charles Sutton (2014),which represents a collection of journal papers from a conference as a special facetted list. Each paper is represented as a word cloud, but the authors modified the algorithm that generates the word cloud in a way that the words appear always in the same position for all the documents of the collection. This admirable trick makes word clouds visually comparable, and, as a result, the traditional static list of documents becomes substantially enriched.

The examples reviewed in this section show that data visualisation can be used as element(s) integrated within standard interfaces to text collections and that sometimes the data visualisation elements can represent the whole interface. The traditional tasks of search and detail-on-demand via a catalogue entry are always present in the reviewed cases of institutional DLs. But, at the same time, a trend seen across the reviewed cases is the introduction of a home page that offers several overviews and/or options to interact with the collection. Additionally, it has transpired that the reviewed cases try to avoid features that require an initial explanation on how to use the interface.

The next part of this review will focus on data visualisation projects specifically conducted to represent textual documents. When compared with the already reviewed institutional DL interfaces, the data visualisation interfaces for text —usually done outside the official institutional sites— show state-of-the-art techniques, experimentation, and innovation, including the use of text analysis. For example, Spot is a twitter-like dashboard to follow trends, people, keywords. The data are dynamic and the input tweets are constantly analysed, indexed, and visualised. By contrast, there is a lack of text analysis and data visualisation tools in institutional DL interfaces. Since the aims of this research project are to create interfaces that support and invite to read text collections, the next subsection reviews the concept of e-reading and the standards of reading interfaces.

2.3.4. Interfaces and Reading

Since this research project intends to propose practical techniques of interface design and develop of text visualisation and exploration techniques oriented towards reading, this subsection reviews the designs of interfaces whose primary goal is to support text reading by integrating texts from collections and prioritise their reading. This subsection presents a brief review of the concept of active reading and briefly considers the wider concept of readability. A review of interfaces to text collections that allow online reading will then conclude this background and context chapter.

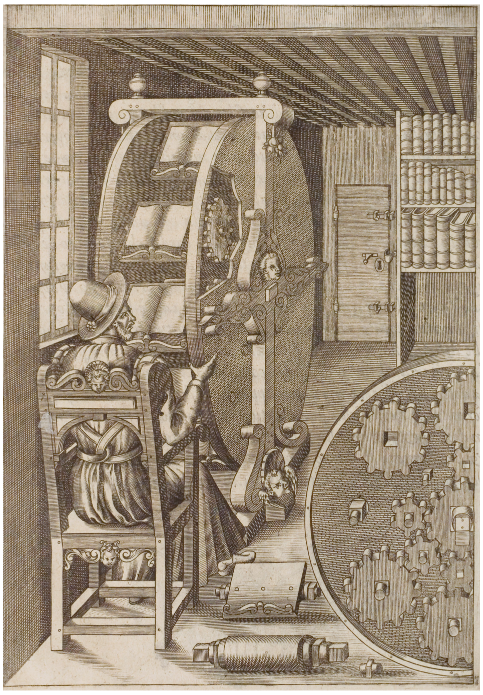


Figure 7. Bookwheel, a mechanical reader designed by Agostino Ramelli from his bookLe diverse et artifiose machine, 1588.

The study of text reading on screens and on paper has been a productive and rapidly evolving field of research since the electronic display of texts became possible, as demonstrated by the highly cited paper of O'Hara and Sellen (O'Hara and Sellen, 1997), which presents the multiple advantages of paper versus screens when reading texts. The advantages of paper mentioned in that paper include: supporting annotation while reading, quick navigation, and flexibility of spatial layout. These advantages today look outdated since modern reading interfaces provided by applications for e-readers on tablets, mobiles, phablets, etc., include them by default. The action of underlining, highlighting, and commenting the text while reading was defined as "active reading" in 1972 (Adler and Van Doren, 1972). Today, a variety of applications with these features can be installed on any digital device, thereby potentially making any device used as an e-reader ready for active reading.

In this context, it is necessary to introduce some concepts associated with text readability. Text readability has been measured in a variety of ways for a several different purposes. A number of methods to measure and calculate readability (Tinker, 1963, Fry, 2006) to take into account elements of graphic design, typography, human perception and psychology, legibility, vocabulary, syntax, etc. have been published.

One of the aims of this research project is to increase readability (i.e., how to make a text more accessible through reading). Independently of the method used to calculate readability, there are two distinct ways to do it. The first possibility is to modify the contents of the text by using text analysis techniques that make a specific text easier for people to understand, including non-native speakers and those with special needs, like people affected by poor literacy, aphasia, dyslexia, or other language deficits (Ferraro et al., 2014). The second possibility to improve readability is to modify the visual aspect of the text and how the readers interact with it by means of specific hardware-software interfaces. This is possible because adding graphic elements to a text can improve comprehension of the text (Ferraro et al., 2014, Suominen et al., 2014).

At the moment of writing there are many online services that are accessible via the readers embedded in web browsers. Some of those services are commercial SAAS (software as a service) that allow online management of documents with annotation, text-highlighting, share button, and bookmarks. This is the case of Issuu, Google docs, Scribd, PDF.js, and others. Other publishers offer software implementations of e-readers in addition to e-reader devices; examples of this second type of tools include the Amazon Kindle app and web page, Barnes & Noble's NOOK book app, and Google play.

Table 2 summarises the features some institutions' online readers. Two DLs listed in the table support powerful online readers with multiple features: the Australian Trove and the Internet Archive. Both offer a book metaphor, share, read-this (English), presentation mode, zoom, 1-2 pages viewer, and print. Trove has extra features such as: tools for social tagging, commenting and transcription, transcription reader, download, and citation formats. The Library of Congress offers multiple formats for the texts, mainly as HTML pages. Europeana, due to the fact that it is an aggregator, redirects the reader to the referenced institution page.

Reviewed site	Reader features
Internet Archive	Book metaphor, share, read-this (English), presentation mode, zoom, 1-2 pages viewer, print.
Library of Congress (USA)	Several formats: from HTML to PDF. When available it offers read-this (English).
Trove	For newspapers: tools for social tagging, commenting and transcribe, transcription reader, scanned image, zoom, print, download, citation formats,
Europeana	Redirects to the reader of each institution

Table 2. Reader features of some major institutional DL

Another aspect related to the way digital text is read is communication fragmentation. The number of messages received per day has increased, as well as the number of channels delivering short multimedia messages. In the past decades several works have explored the possibilities of breaking the linearity of a text. The philosophers Deleuze and Guattari have described the rhizomatic structure of knowledge, which inspires this project too: ”In a book, as in all things, there are lines of articulation, segmentarity, strata and territories; but also lines of flight, movement, deterritorialisation and de-stratification” (Deleuze and Guattari, 1987). In general, the complexity of the knowledge produces a multiplicity of narratives. In the novel Hopscotch by Cortázar (1966), the author proposes two reading orders for the chapters; the text starts with: ”In its own way this book is many books, but mostly it’s two books”. Another relevant work is the Project Xanadu from 1960 (Ted Nelson, et al, 1960), considered the first hypertext project in the digital era and presenting a visionary definition of standards for the WWW that were mostly excluded from the standard protocols we currently use. One of Xanadu’s rules states: ”Every document can consist of any number of parts each of which may be of any data type”. The open Xanadu project encourages non-linear navigation of text. The aim of Xanadu’s demo is to show the possibilities of hypertext. It is examples like these that have motivated me to investigate the effects of alternative ways of reading in combination with normal reading.

The examples that allow the reading of narratives in more than one way form the basis of the idea of segmenting long texts as a strategy for reading. This idea of narrative multiplicity is developed in detail in chapter 6.

2.4. Summary

This chapter presented three distinct and complementary brief reviews. Many big public DL institutions are becoming meta-libraries that catalog digital objects not only hosted locally, but but also held in other institutions, making it possible to search across multiple collections. The most extreme example of such a strategy is Europeana, which only works as an aggregator, not hosting any objects of its own.DL collections are now far too large to be readable in any traditional sense, and this posthuman scale reinforces the urgent need for tools to deal with huge volumes of materials. At the same time, there is a contrast between this posthuman scale of DLs and the limitations of the traditional interfaces based on lists, facets, and standard metadata that we find in the official pages of the reviewed institutions.

The availability of large digital collections in DLs offers significant opportunities for computational text analysis. The examples reviewed here show that techniques such as topic modelling can help represent document content. A simple classification of levels of data in indexing digital objects of a collection is proposed: raw metadata (which only uses metadata), extended metadata (which enriches the raw metadata by means of simple operations), and object analysis (which includes new information about the collection through analysis of the text of the documents of the collection). The DLs reviewed do not use object analysis, but examples of text analysis in text collections can be found in practitioners and researchers working outside institutional DLs. In 2.2.3 several cases of data visualisation that include text analysis are reviewed. This third level that includes text analysis can add subjectivity in the interpretation of the contents of the collection but, at the same time, it can improve the representation of the complexity and abundance of the collection —as seen in 2.2.3 ”Mining the dispatch” and ”Topic Modeling Martha Ballard’s Diary”. In this direction, Johanna Drücker has influenced this research with her concept of capta, as opposed to data, which proposes an active attitude in interpreting data and metadata of a collection. The section ends with a review of topic models, its relation to DH, and its labelling (which is what makes topic models human readable) and evaluation. Methods for topic labelling go from pure procedural (computer-aided, common in sciences and used when the number of topics is high), to pure curated (manually done, usually by experts, common in DH). This variability in labelling has resulted in extensive literature on the evaluation of topic labelling methods. The reviewed cases show that procedural labelling can in some cases be comparable to curated labelling. In any case, the review shows that the use of topic models in text collection interfaces to help exploration and reading is not common.

The third part of the review covered the concepts and practice of interfaces to DL and text collections taking into consideration theoretical frameworks and philosophies that support innovation in interface design developed over the last twenty years. This review shows design concepts that go beyond the traditional search task. Interface design can bring freedom to the user and improve explorability of collection contents. Concerning the interface standards and trends in text collections, the review shows that interfaces to major institutional DLs are conventional, text-based web pages, but also that opportunities to enrich these interfaces exist and are clear. The examples discussed show how data visualisation can be included in collection interfaces, either as a whole or as an element of the interface. This review shows that more innovative interfaces to text collections are found use data visualisation techniques, although, In most cases, these progressive works are not integrated into the official sites that host the collections. Finally, a short review of interfaces for reading texts on screens and displays shows that the main features for e-reading seem established. Additionally, several experiments with non-linear reading point towards possible solutions for reading parts of collections too big to be read from beginning to end. The following chapter analyses the reviews and identifies some gaps and opportunities for research and innovation in the development text collection interfaces.

3. Gaps, Research Questions and Contribution to Knowledge

This chapter starts by describing the gaps and opportunities identified in the previous chapter. It then introduces three questions that my research intends to answer. Finally, it presents the contribution to knowledge that this research project has generated. Since this is a practice-led research project, it has produced specific digital artefacts in addition to contributions to theory. Detailed information about these creative works and an analysis of the overall outcomes of the project can be found, respectively, in Section5.1 and Section 6.1.

3.1. Gaps

The reviews presented in the previous chapter (2 Context and background) resulted in a list of gaps and opportunities for contributions to knowledge. Due to the diversity of the three areas reviewed —digital libraries, text analysis, and digital collection interfaces— the findings are diverse and complementary. The process of developing the artefacts involved experimentation in all three fields (see chapter 5).

The subjects of this chapter follow the same order as those of the previous one (i.e., Chapter 2). First, it reviews the major DLs that are considered mainstream.They represent the current standard of contemporary interfaces to DL in general and, more in particular, to collections of texts. This first review shows the contrast between the posthuman scale that DLs have reached in their continuous growth and the poor innovation in the interfaces officially offered by big public institutions. These institutions host DLs that grow everyday, but they use text-based interfaces that are ineffective at representing collections. This contrast reveals an opportunity in the development of interfaces to large collections of documents that fits with the aim and motivation of this research project.

The second review concentrates on text analysis. The main gap found here is that the output of text analysis, in particular the topic modelling of text collections, is generally applied at a collection level. As with Moretti's other "distant reading" methods, topic model analysis is commonly used to show overviews of the collection, distribution of topics ("Martha Ballard's diary"), comparison of topics and/or documents, and similarities among documents ("Mining the Dispatch"). In contrast, no cases are to be found dedicated to representing single documents or parts of documents. The generated representations of the collections are presented as diagrams, charts, maps, but rarely are the results of the analysis integrated into the interface. The cases found that integrate the outputs of the analysis of the collection in exploration systems are for image collections, as in "Discover the Queenslander" (Whitelaw, 2014), where the images are analysed to extract a palette of colours, thereby allowing to browse the collection of images by colour. The application of text analysis to text collection and the integration of its results into interfaces presents a clear opportunity to create features that support exploration and reading.

As already discussed, topic modelling is a powerful tool that it is still not widely applied. Most of the reviewed applications of topic model analysis are justified on the basis of one or more of the following reasons: to compare different topic, to evaluate individual topic models, or to represent a collection as a whole for overview, evolution, and comparison. The support for reading the content of the collections is limited. For example, "Mining the Dispatch" shows "exemplary articles" for specific topics, but the interface is non centred on the documents, in the sense that the documents are exemplars of the topic, rather than the topics being a guide to the documents. "Martha Ballard's Diary" is not an interface. Therefore, it does not encourage exploration, but again the large scale analysis emphasises topics across the diary, instead of its contents. Obviously, all authors present topic models related to collections of documents, but rarely are the documents central in their approach. In particular, topic model analysis is rarely used to support reading documents, and this is a significant gap.

More investigation about the appropriate number of topic labels for each collection and context is necessary. Literature shows that the list of topics that the algorithm produces can be modified by a curation process (Nelson, 2010a). There are tools that help in that process: for example, MetaToMATo (Metadata and Topic Model Analysis Toolkit) (Snyder et al., 2013) is a web-based application that allows combining metadata and topic models of collections of texts. It can filter documents by topic and summarise views with metadata and topic graphs. Another related approach is provided by the prototype of TopicExplorer (Hinneburg et al., 2012), which combines topic modelling, keyword search, and faceted lists to explore a large collection of Wikipedia documents and other collections. In the process of reviewing topics by curation, some topics can be removed because they are considered too general, or semantically unusable; some can be merged with other topics, in cases where topics overlap or are included semantically in others, or are, in fact, synonyms. The literature and the reviewed examples point to opportunities for experimentation in the process of labelling topic models.

In topic model labelling the literature recommends choosing inclusive labels. For example, in Nelson's ”Mining the Dispatch”, one of the labels is ”fugitive slave ads” (Nelson, 2010b) but, as some of the texts are not ”ads”, Nelson concludes that topic models are better suited to representing the prevalence of a topic in the collection, rather than strictly and accurately classifying every text.

According to the literature, there is no clear definition of how to optimise topic model analysis parameters, and in particular the number of topics to generate in relation to the size of the segments of text that make up the collection. In determining the size of the segments, researchers try to find any pre-existing segmentation, such as pages of a diary, as in ”Martha Ballard's diary (Blevins, 2010). Some research opportunities exist in experimenting with these parameters through the development of real cases. See chapter 6 for a detailed explanation.

When reviewing the examples of interfaces that apply topic modelling a generalisation emerges: while experts in analysis, usually mathematicians, statisticians, or computer scientists, apply computer-based processes with no curated methods other than dataset annotation for training (i.e., they conduct a qualitative evaluation of the machine-based method), researchers in the humanities can find computational methods too opaque due to knowledge barriers. The review shows that from the point of view of the humanities, it looks acceptable and even desirable to mix curated and procedural methods in the same process. This relates to the mentioned Drucker's concept of capta: data is "taken" for interpretation, and curated labelling is just another form of interpretation.

Concerning interfaces to digital collections, the common strategy is to follows the schema of Schneiderman's Mantra: an initial overview allows contextualisation and/or understanding of the collection from a faceted lists to a geo-map and from a network visualisation to a treemap. Following the overview, an interaction shows collection details, thereby allowing to reach individual items. Some approaches only follow one or two of these three schematic steps. When building interfaces for reading, this approach does not take the shortest path to a piece of text in the collection. Following the three steps strictly, the visitor needs several clicks (i.e., decisions) to reach (and read) an individual item in the collection.

This suggests to revisit the "overview first" strategy and see what can be done to put the contents of the collection in the centre of the scene. Even though the overview is necessary, useful, and perhaps inevitable, sometimes it can act as a substitute for the content. This observation influenced how to develop the communication channels when creating artefacts for this research project. Lessons learned when answering this question are analysed in chapter 6.

The interface to a collection of texts spans from the data items to the design of the smallest graphic elements that the user sees. Its development is therefore a process that can encompass many steps. Some of those steps are perhaps manually done by experts, others will consists of procedural methods performed by computers. This review has shown a variety of these two types of methods used to perform similar tasks, but neither of the two types is a priori better than the other. The reasons for choosing a particular approach depends on the goals and the context of each individual case.

When reviewing interfaces dedicated to reading text, it transpired that standards for reading texts on digital devices are already well established and offered for most devices and interfaces in contemporary systems. For this reason, when developing interfaces for reading, it is appropriate to follow the main features that can be considered standard according to the classic definition of "active reading" made by Adler and van Doren (1972)./

The following subsection introduces the two research questions at the basis of this research project that cover the two closely linked aspects of the research: its practice-led nature and its theoretical findings.

3.2. Research Questions and Contribution to Knowledge

The central question that this research project aims to answer refers to the practical case of interfaces to text collections and drives both the theoretical methodology and the creative work: how can we create interfaces to text document collections that let us explore the collection by reading its contents? This question is answered with the creation of real artefacts as well as by discussing the theoretical issues presented in this dissertation, which are documented and described as new methods for creating text collection interfaces focussed on reading.

As a contribution to knowledge, a group of digital artefacts are presented in the form of visualisation interfaces to text collections developed with a focus on exploration-by-reading and on the possibility to freely move from inner view to overview and vice-versa. The term ”Inner view” is used here in opposition to overview, to indicate a perspective in which the collection is viewed from within, looking at a sampling of the details, instead of from the outside, looking at the content as a whole. The proposed approaches use relationships among collection items to guide the browsing of the collection, instead of the ”details-on-demand” approach that traditionally ends an information-seeking session. With the approach presented here, the reader can move between related detailed views or back to a collection overview to dive again into the contents of the collection.

The collections chosen to work with in the presented artefacts include a range of contents that range from scientific papers to art and historical collections. As a direct consequence of this research project, these collections have gained accessibility (i.e., readability) and presence, since the created interfaces are universally accessible online. Therefore, this project contributes to expanding resources for the public acquisition of knowledge.

The methods used in the analysis of the texts are novel, in that they use state of the art techniques such as topic models, entity recognition, and text similarity. Besides in the one-to-one categorisation of collection objects, topic model analysis is also used in this project to create rich and complex multi-dimensional relationships between segments of texts. This provides a more advanced way to characterise the collection contents and enables navigable relationships between collection elements. This technique is an essential element of the ”deep interfaces” approach developed in detail in chapter 6.

Since the developed interfaces support and give priority to reading the documents of text collections, it is important to consider how to deal with collections that are too big to be fully read. The answer to this question is the introduction of the concept of crossreading as the non-linear reading of large collections. To adapt a collection of texts to crossreading, a two-step process is necessary: first, the texts that belong to the digital collection are divided into small segments, which become the objects of a new collection; then, once the new collection has been analysed with topic modelling, new multidimensional metadata is associated with each object. As a result of the process, new ways of exploring and reading the collection emerge.

As a contribution to theory, the process of generating “deep interfaces” to text collections is described in detail, so that it can be reproduced with other datasets and in other contexts. Every step of the process has been validated and improved through a cumulative development strategy that, following a reflexive methodology, has produced the set of individual artefacts presented in this dissertation.

All documents and their sources related to this thesis, including the details of theresearch project, the literature review, source code, source files of images and documents, and publications can be found in the Github repository at https://github.com/jaumet/myacademydata.

In summary, this chapter has described the gaps and opportunities for new research revealed by reviews presented in Chapter 2. After introducing the two research questions that this project intends to answer, this chapter has also identified the contributions to knowledge that this project provides. The following chapter introduces the methodologies followed and the methods applied.

4. Methodologies and Methods

The list of methodologies introduced in this chapter is structured to best illustrate my research, which is based on the production of interfaces to real cases of text collections, and, more in general, to describe the research process with a list of good practices that could be applied for other projects and by other researchers. Accordingly, this section starts with introducing the methodologies and continues with presenting a list of methods and techniques used in the research project.

4.1. Methodologies

Practice-led and Practice-based Research:

1. This is a project based on practice, and in particular on the practice of software creation. According to Linda Candy (Candy and Studios, 2006), there are two types of practice-related research: practice-based and practice-led. In her words:If a creative artefact is the basis of the contribution to knowledge, the research is practice-based.

2. If the research leads primarily to new understandings about practice, it is practice-led.

Following Candy's definitions, this project is presented as a combination of these two practices. According to the mentioned aims of the project,

1. The project has created a number of software artefacts that have contributed to knowledge, making the presented collections more accessible. In that sense this project is practice-based.

2. At the same time, the cumulative process of creation of artefacts has produced theoretical insights and a methodology to build interfaces to text collections. In that sense the project is practice-led.

This double-practice combination of research methodologies follows Greame Sullivan's braid metaphor, where creative practice cannot be separated from research theory. In other words, according to Sullivan, the complexity of that relationship, as in a braid of fibres, reveals all kinds of structures within practical and theoretical research (Sullivan, 2005).Sullivan's metaphor explains the richness, the complexity, and the fruitful relationships between practice and theory in this research project.

Reflexive Methodology The symbiosis between practice and theory in this research project is embodied in the process of iterative development of the interfaces, learning from experience, developing new understandings, and reapplying these to the production process. The reflective practice produces improvements during development due to the continuous learning and the questioning that the process itself generates. I was inspired to apply this reflexive methodology by its tradition in the humanities and creative arts. Sullivan talks about four kinds of reflexive practices: self-reflexive, reflexive, dialogue, and questioning. Even though all four practices are related and can be used simultaneously, this project is in essence a reflexive dialogue, as clearly explained by Sullivan: "the plausibility of an interpretation of research findings will be determined in part by the capacity of the reflexive researcher to openly dialogue with the information" (Sullivan, 2005).

From a more pragmatic point of view, this reflexive idea of continuous inter-practice inquiry could be understood as "learning by doing" as defined by Roger C. Schank (Schank, 1995): since this research project produces artefacts that act as “doing-devices” it demonstrates that learning by doing is feasible practice.

The rationale of applying the reflexive methodology to this research project is that it allows experience and insight to be exported from one artefact to the next in a cumulative process. Indeed, the final artefact (Diggers Diaries) embodies the knowledge and techniques developed during the creation of the preceding artefacts.

Empiricist Methodology Empirical methodology is based on sensory experience and the evidence that can be extracted from it. Applied to this research project, sensory experience comes with the creative production of digital artefacts. The evidence emerges from the validation of the adopted processes through two forms of evaluations:

- qualitative, by means of online questionnaires;
- quantitative: through semi-structured interviews, public exhibition, and feedback from users.

4.2. Methods

This section introduces the methods used during the research project. The methods included in first subsection refer to tools and techniques that have allowed the studies and experiments to be successfully conducted, while the methods discussed in the second subsections were used for evaluation purposes, to validate the usability and goals of the generated artefacts.

4.2.1. Tools and Techniques

Curated —conducted by humans— and procedural —conducted by computers— methods are used across the multiple stages of the development of each of the presented artefacts. In the following list the methods are grouped by task and subtasks.

Data Gathering: Data for the projects was gathered through a combination of procedural (APIs, spiders, and scripting for scraping) and curated methods (manual download, annotation, first classification for storing).

Text Analysis: choices and evolution:

- Text document segmentation:

Several of the interfaces rely on the segmentation of source texts to enable a cross-reading approach. This makes the step of segmenting texts a necessary key method. Three different methods were tested: a curated segmentation (Crossreads II, see

5.3), a procedural segmentation (Crossreads I, see 5.2), and a segmentation by diary's pages (Diggers Diaries I and II).

In a curated segmentation process, one or more editors perform their own segmentations of each document of the collection. In a procedural process, an application splits documents into pieces based on length, sentences, paragraphs, sections, etc. The reason for these two approaches is that the curated segmentation task is very subjective, in the sense that a human expert is likely to add a personal view to the segmentation (Crossreads II). A procedural segmentation (Crossreads I) can accomplish well this task in terms of size of each segment, but it cannot be expected to have the richness of a segmentation curated by an expert, who can rely on the semantics of the text to draw the boundaries between segments.

The collection studied for the final project, the Diggers Diaries (I and II), was segmented into handwritten pages.

- Topic model analysis

I used manual algorithms and automated tools like MALLET, which McCallum describes as "a Java-based package for statistical natural language processing, document classification, clustering, topic modelling, information extraction, and other machine learning applications to text." (McCallum, 2002).

- Topic model labelling

I used curated labelling plus labels grouping. This generated a two-level hierarchical menu of topics. The topic grouping and labelling process is discussed in depth in Section 6.

The web application was implemented by means of client-side Javascript relying on the libraries angularJS, bootstrap, and JQuery.

4.2.2. Evaluation and Validation

In order to evaluate the generated artefacts, it was necessary to determine how users accepted and used these interfaces. In the case of Visference, the evaluation focussed on new features introduced through text visualisation by comparing it with existing or conventional presentations of text. In the case of Crossreads there was no existing interface to compare it with the tool provides a novel exploratory presentation of specific text. For this reason, Crossreads was evaluated using semi-structured interviews. The results of these two evaluations informed the design of the last artefact, Diggers Diaries, which relied on the conclusions of the first two studies.

The evaluation addressed the following questions:

1. Do users detect the new features?
2. Do users prefer or require access to the presentation that existed before the artefact being evaluated?
3. Do users understand the new features?
4. Do users feel confident and positive about using the new features?

Accordingly, the evaluation was based on the established Technology Acceptance Model (TAM) (Davis et al., 1989), and task technology fit (TTF) (Goodhue, 1995). TAM attempts to understand why people accept or reject information technologies, while TTF claims that technologies will be used if, and only if, their functionality supports the user's activities. Consequently, its focus is on the match between what the uesrs need to accomplish their tasks and what functionality of a given technology is available. The questions 1 and 2 are directly related to TAM, while 3 and 4 are related to TTF. The questions were designed following (Taylor-Powell, 1998).

The evaluation was in the form of a mixed-method approach consisting of:

a) Online questionnaires for standard users. Here ”standard users” refers to potential visitors of each real scenario. For Visference, these are readers of academic journals.

Online questionnaires are the most popular method to gather from users quantitative data for statistical analysis. They allow participation from an unlimited number of people and can collect data on knowledge, beliefs, attitudes, and behaviours (Taylor-Powell, 1998). Online questionnaires also make it easy to protect the privacy of participants.

b) Semi-structured interviews with expert users. In order to complement the online questionnaire and provide open-ended evaluation and feedback on each tool from experts, a small number of semi-structured interviews with experts in each artefact's scenario or domain were conducted.

A semi-structured interview is a qualitative method. It is open, in the sense that participants express their ideas freely following pre-defined questions, thereby allowing new ideas to be expressed during the interview (Kitchin and Tate, 2013).

The results of the online questionnaire can be found in the Appendix. The analysis of the questionnaire results and the questions and opinions from the semi-structured interviews are discussed in chapter 6 Discussion.

5. Results: the Artefacts

This chapter presents the results of the research project. Due to its practice-led nature,this chapter focusses on the created artefacts. Analysis of results, lessons learned, and general discussion will take place in chapters 6 Discussion and 7 Conclusions and Future.

All the artefacts are digital and are accessible online. Nevertheless, there is a website that lists and introduces each one of them as part of this research process. The title of the site is "Web of artefacts", with URL `http://research.nualart.cat/woa`.

The description of each artefact presented in this chapter follows the same outline:

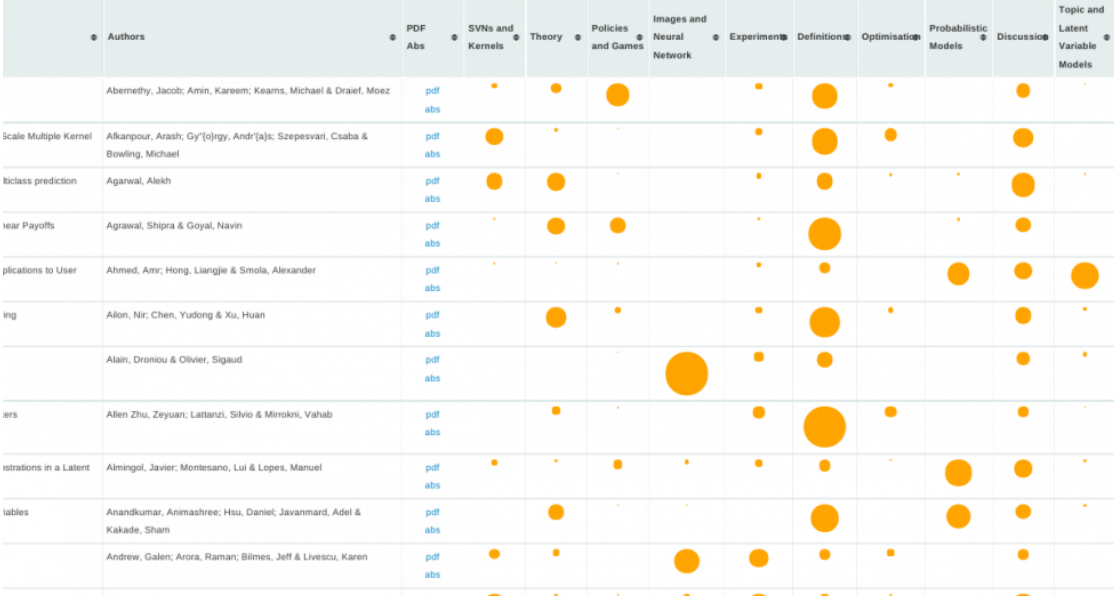
- A figure showing a detail of the generated interface.
- Name, version, and short description of the artefact.
- Narrative: a description of the artefact motivation and aims, as well as of its contribution to the project evolution.
- Dataset: description of the primary dataset to which the artefact interfaces and, when applicable, of complementary datasets generated during the research or taken as external sources.
- Collaborators: a list of direct collaborators and their field of expertise in relation of each artefact.
- Data process: a description of the process in terms of data gathering, reformatting, and conversions, analysis, final formatting, storage, and access.
- Data analysis: detailed description of the analysis done and the tools used.

- Interface development: description of challenges and techniques, as well as of external libraries used.

- User evaluation studies: results of the conducted user studies. The detailed results in tabular form and the associated charts can be found in Appendix A.

- Project outcomes: list of all project outputs, including publications, exhibitions, presentations, and all the code.

5.1. Visference: Journal of Machine Learning Research



	Authors	PDF Abs	SVNs and Kernels	Theory	Policies and Games	Images and Neural Network	Experiments	Definitions	Optimisation	Probabilistic Models	Discussion	Topic and Latent Variable Models
	Abernethy, Jacob; Amin, Kareem; Kearns, Michael & Draief, Moez	pdf abs	●	●	●		●	●	●		●	●
Scale Multiple Kernel	Afkanpour, Arash; Gy"{o}rgy, Andr'{a}s; Szepesvari, Csaba & Bowling, Michael	pdf abs	●	●	●		●	●	●		●	
lticlass prediction	Agarwal, Alekh	pdf abs	●	●	●		●	●	●	●	●	●
near Payoffs	Agrawal, Shipra & Goyal, Navin	pdf abs	●	●	●		●	●		●	●	
plications to User	Ahmed, Amr; Hong, Liangjie & Smola, Alexander	pdf abs	●	●	●		●	●		●	●	●
ing	Ailon, Nir; Chen, Yudong & Xu, Huan	pdf abs		●	●		●	●	●		●	●
	Alain, Droniou & Olivier, Sigaud	pdf abs			●	●	●	●			●	●
ers	Allen Zhu, Zeyuan; Lattanzi, Silvio & Mirrokni, Vahab	pdf abs		●	●		●	●	●		●	●
strations in a Latent	Almingol, Javier; Montesano, Lui & Lopes, Manuel	pdf abs	●	●	●	●	●	●	●	●	●	●
iables	Anandkumar, Animashree; Hsu, Daniel; Javanmard, Adel & Kakade, Sham	pdf abs		●	●	●		●		●	●	●
	Andrew, Galen; Arora, Raman; Bilmes, Jeff & Livescu, Karen	pdf abs	●	●		●	●	●	●		●	

Figure 8. Detail of Visference

Artefact name and description

Visference is a visualisation tool for the exploration of conference papers. It uses topic model analysis and sorting table techniques.

Datasets This project has an initial dataset (dataset-1) and a created one (dataset-2)

- Dataset-1: collection of 282 accepted papers from the Journal of Machine Learning Research (JMLR) Workshop and Conference Proceedings Volume 28: Proceedings of the 30th ICML (International Conference on Machine Learning).
- Dataset-2: a golden dataset created as a representation of a machine learning domain. A golden dataset, also called golden standard, refers to a list of texts that can be considered representative of a domain, in this case machine learning. This dataset was expanded with:

  - A list of classic books about statistics, including: Alppaydin's "Machine Learning 2010", Mackay's & Barber's "Machine Learning", Tom Mitchell's "Data Mining-Practical Machine Learning Tools and Techniques", C.M. Bishop's Pattern Recognition and Machine Learning, in The Element of Statistical Learning 2nd Ed. (ESLII), and R.W. Smola's "Learning Kernel Classifiers".
  - Every papes listed with "stats" on arXiv for the period 2010 to 2013

Narrative

Motivation: most conference proceedings present their content as a one-dimensional, non-interactive list of papers on a web page. However, the user of this kind of presentation might not know the ordering criteria used for the list, does not get an overview of the contents or relations between the papers, and has very limited search and filtering functionality available. More in detail, such a presentation has the following limitations: a flat and non-interactive list, no sorting options, no overview of the dataset, no relationships among items, only CTRL+F (or COMMAND+F) for searching, and no filtering.

Aims of the artefact: to explore more effective interfaces to represent the contents of conference proceedings. That notwithstanding, this prototype remains visually conservative, in the sense that it tries to avoid scaring the habitual visitors with unfamiliar visualisation by incorporating known interactions, like table sorting. The tool has been evaluated positively by users and experts.

The experimentation with topic models showed how significantly topic labelling can improve human readability of topics. The creation of a golden dataset as the basis for finding a set of representative topics demonstrated that the dataset must be big and representative of a knowledge domain or field. Finally, using non innovative interface elements and interactions, makes it easy to introduce new interfaces because no initial knowledge is required.

Collaborators

- Dr Wray Buntine, Machine Learning Research Group, National ICT Australia (NICTA). Specialist in topic models.

- Dr Mark Reid. Machine Learning Research Group, NICTA. Editor of the JMLR.

Data process

- Dataset-1 gathered from a bibtex file.
- Dataset-1 transformation: papers in PDF were converted to text, and then all papers were split into pages, thereby producing about three thousand segments of text.
- Dataset-2 transformation: books and papers in PDF were converted to text, and then split into pages, thereby producing approximately fifty thousand segments of text.

• The final topic model output was parsed and transformed to JSON (JavaScript Object Notation) format.

Data analysis

- The topics were generated from dataset-2.
- The topic models for the smaller dataset-1 were inferred by using the model trained on dataset-2.
- Topic model labelling: experts from the Machine Learning Research Group at NICTA suggested labels for each topic. Topic model results and labels can be found online (Buntine and Nualart, 2013).

Interface development

The proposed interface includes a simple HTML table with the 282 papers as rows and topic models as columns. The table is sortable by columns. Each cell shows a dot proportional in size to topic model score.

User evaluation studies An online questionnaire and semi-structured interviews were conducted to evaluate the acceptance of Visference as a new tool for exploring conference papers when compared to the existing static conference page. On average, 80% of the thirty-three participants preferred Visference over the existing interface for typical tasks. Task sizes and percentage of positive answers for Visference were as follows:

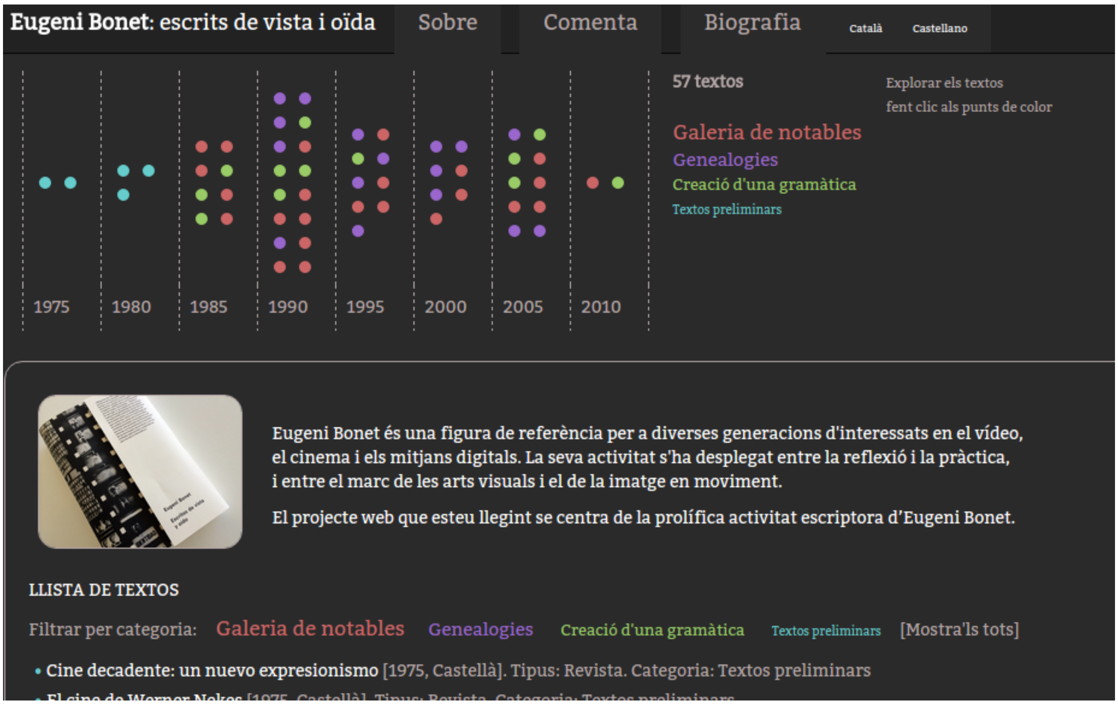
- How many papers were accepted in the conference? 91.18%
- Which papers are related to machine learning theory? 70.59%
- How many papers talk about optimisation: 85.29%
- Understanding the topics and themes of the conference: 85.29% of positive answers
- Finding papers related to your personal interests: 82.35% of positive answers
- Exploring new topics and discovering new research in this field: 88.24% of positive answers

See Appendix A for detailed results of the questionnaire. Semi-structured interview results are integrated into Chapter 6.

Project outcomes

- A demo site: `http://research.nualart.cat/visference`
- Publication: poster and a presentation at several NICTA retreats and conferences http://research.nualart.cat/?p=54.
- Ten topics that can represent the machine learning knowledge domain: SVNs and Kernels, Theory, Policies and Games, Images and Neural Network, Experiments, Definitions, Optimisation, Probabilistic Models, Discussion, Topic and Latent Variable Models
- The code is accessible under GPL licence in github (Nualart, 2014).
- Development charts are accessible in github (Nualart, 2014)

5.2. Crossreads I: Eugeni Bonet exhibition



Eugeni Bonet és una figura de referència per a diverses generacions d'interessats en el vídeo, el cinema i els mitjans digitals. La seva activitat s'ha desplegat entre la reflexió i la pràctica, i entre el marc de les arts visuals i el de la imatge en moviment.

El projecte web que esteu llegint se centra de la prolífica activitat escriptora d'Eugeni Bonet.

LLISTA DE TEXTOS

Filtrar per categoria: Galeria de notables Genealogies Creació d'una gramàtica Textos preliminars [Mostra'ls tots]

- Cine decadente: un nuevo expresionismo [1975, Castellà]. Tipus: Revista. Categoria: Textos preliminars
- El cine de Werner Nekes [1975, Castellà]. Tipus: Revista. Categoria: Textos preliminars

Figure 9. Detail of Crossreads I

Artefact name and description

Crossreads I, a way of deconstructing linear narrative text in order to read text fragments in multiple orders.

Datasets

Fifty-seven articles by Eugeni Bonet (Barcelona, 1954), video and cinema artist and writer.

Narrative

This project introduces the concept of Crossreads. It is derived from a proposal for the Museum of Contemporary Art of Barcelona (MACBA). The museum sought ways to represent a collection of texts through a web interface that could be used online and from within their exhibition space. The proposal was to build a section of the exhibition "The Listening Eye - EUGENI BONET: SCREENS, PROJECTIONS, WRITINGS" (May to August 2014). Eugeni Bonet (Barcelona, 1954) is an important figure for the Catalan and European art in video, cinema and digital media in general.

Crossreads represent an idea in the domain of information seeking and that offers an experimental way for reading texts as an alternative to traditional linear reading. It proposes to break the initial narrative line of a text by segmenting it into smaller parts. Then, the text is reordered according to similarity scores, which offer the reader multiple paths to read the text. The aim of this project is to explore and study how a reader processes fragmented information, to analyse user activity, and to support reader's exploration with visualisation techniques.

The syntactic similarity used in this artefact is weak. The reasons for this could be: the too-small size of the dataset and the technique to calculate similarity. This artefact experiment focuses primarily on the reader of the dataset, while the visualization exploration tools play a secondary role.

Collaborators

Dr Gabriela Ferraro (Machine Learning Research Group, NICTA). Natural language processing specialist.

Data process

- Documents compilation, conversion to text format.
- Procedural segmentation of texts into a total of 710 segments. segment length was about seven hundred characters in total, which equates to an average of one minute of reading for an adult (Williams, 1998).

Data analysis

The similarity between segments was calculated with off-the-shelf Natural Language Processing tools and techniques (Nualart et al., 2014). The analysis steps were:

- Tokenisation: words in the segments were separated by whitespace and punctuation characters.

- Stop-word removal: standard stop-word removal.
- Named Entity Recognition: identification and classification of Named Entities (NE) in each segment. This was done by applying the OpenNLP Named Entity recogniser [2] (where NLP stands for Natural Language Processing), which distilled four types of entities: Person, Location, Organisation and Others.
- Similarity calculation between segments.

Browsing of the segments was constrained according to the segmented dataset by imposing the limitation that each segment was linked to the most similar segment to which it was not already linked. The drawback of this approach is that the links will have a wide range of similarity scores because, in each iteration, the number of segments to compare with becomes smaller, which in turn causes the possibility of finding a segment with a high similarity score to decrease. For this reason the limitation of this method is that the similarity between segments is smaller on every new step. A very large collection of texts would eventually solve this problem. However, this method has the benefit that there will not be any orphan segments. That is, all segments link to other segments, so that the reader always has the possibility of some crossreading.

Interface development

This artefact was intended to be accessed in the exhibition space of the Museum of Contemporary Art of Barcelona (MACBA). The interface was therefore designed to make the contents easy to read. In the exhibition space several tablets were accessible to browse the collection of texts through the Crossreads interface, and one of the tablets was projected on a wall.

From a technical point of view, Crossreads I (i.e., the original version of Crossreads) was a client-side web app, written in Javascript with the libraries Jquery and Bootstrap. The data was stored in online accessible JSON files.

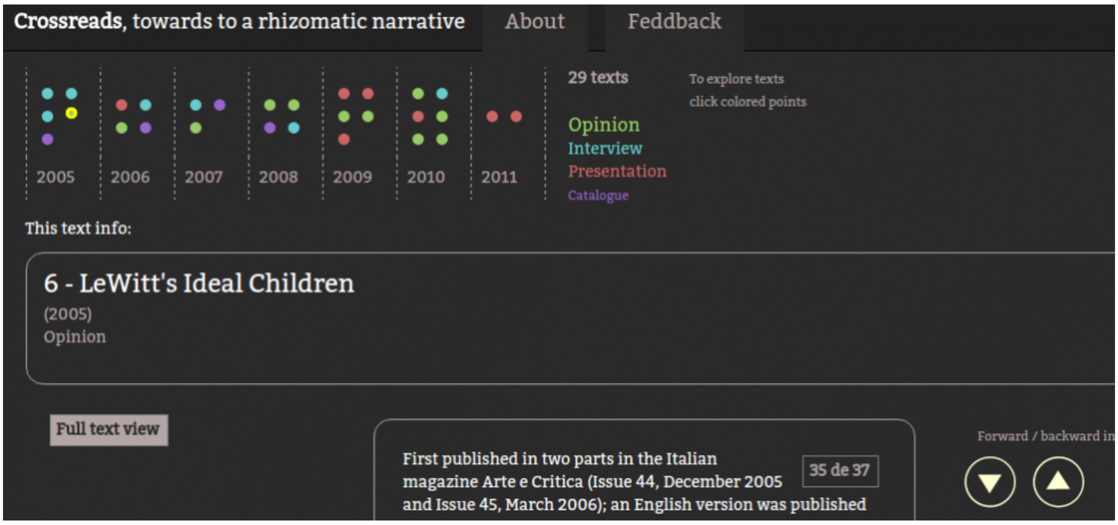
User evaluation studies

The interface development process was evaluated through presentations and reports to a team composed of art historians, librarians, and curators of the museum. The feedback received from the organisers was positive. No more evaluations were done after the closure of the exhibition in August 2014.

Project outcomes

- A website http://research.nualart.cat/crossreads/I.
- This project is part of the exhibition ”The Listening Eye EUGENI BONET: SCREENS, PROJECTIONS, WRITINGS” at the Museum of Contemporary Art of Barcelona (MACBA) 2014 May to August 2014.
- The exhibition catalog: EUGENI BONET: ESCRITOS DE VISTA Y ODIO 2014 (In Catalan and Spanish) MACBA ed, 338 pages. ISBN:978-84-92505-69-2 (http://www.macba.cat/en/publi-eugeni-bonet).
- Code is accessible under GPL licence in github.
- Development charts are accessible in github.

5.3. Crossreads II: "In your computer", by D Quaranta



6 - LeWitt's Ideal Children
(2005)
Opinion

First published in two parts in the Italian magazine Arte e Critica (Issue 44, December 2005 and Issue 45, March 2006); an English version was published

Figure 10. Detail of Crossreads II

Artefact name and description

Crossreads II, a way of deconstructing linear narrative text in order to read text fragments in multiple orders.

Datasets

The book "In your computer", by D. Quaranta

Narrative

This project is very similar to Crossreads I, but the first version was a project where both the contents and the artefact were in Catalan and Spanish. For the project to be shown internationally at the DL2014 conference, it needed to be in English. Furthermore, a new version provided the opportunity to work with a larger number of segments of ext and try different criteria when crossreading. The second version also introduced the random button, which allows a jump to a random segment of the collection, equivalent to opening a book randomly. In this second version, the code from the first one was cleaned and improved.

In Crossreads II the dataset is twice as large as that of Crossreads I. That notwithstanding, the similarity is still weak for the cosine similarity technique applied.

In the evaluation, the users confirmed the validity of placing the dataset reader in the main position of the interface.

Collaborators

Dr Gabriela Ferraro (Machine Learning Research Group, NICTA). Natural language processing specialist.

Data process

- The book is accessible online under free licences. Once downloaded, the text was converted to text format.
- Unlike when developing Crossreads I, in Crossreads II the segmentation was curated by dividing the text in 1500 segments, twice the segment size of the first version.

Data analysis

The analysis was similar to that for Crossreads I (see previous section). Version II differs from version I in defining browsing constraints according to the segmented dataset. In version II each segment is linked to its most similar pairing. To avoid repetition of pairs, segments that have already been set as most similar are skipped for ten iterations, after which the skipped segments are used again for the calculation of similarity.

Interface development

The principles and technical choices are identical to those of Crossreads I. But the interface of version II has evolved to offer link nuances. For example, by selecting the right link, the reader can immediately reach the segment that is most similar to the current one.

User evaluation studies

No user studies were done for this second version as it follows the same principles as version I, which was evaluated and approved by experts during its development.

Project outcomes

- A website `http://research.nualart.cat/crossreads/II`.
- A poster and a short paper in the workshop ”The search is over” as part of the Conference Digital Libraries 2014 (DL2014), London (University College London).

5.4. Diggers Diaries I: WWI Diaries

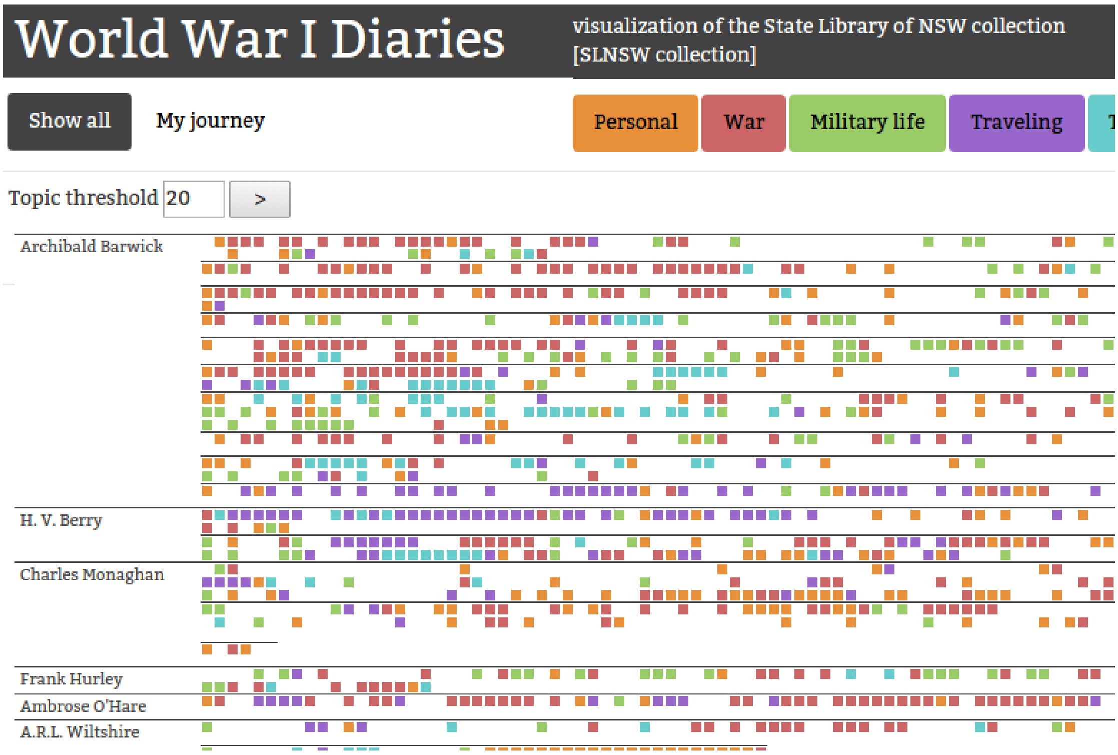


Figure 11. Detail of Diggers Diaries I. This interface has only one way to overview the diaries, this one: by diaries/pages in a vertical list.

Artefact name and description

Diggers Diaries I, an interface to explore a portion of the SLNSW (State Library of New South Wales, Australia) WWI Diaries Collection through general topics

Datasets

126 diaries from WWI as part of the WWI Diaries collection hosted by SLNSW using the handwritten 11944 pages of the diaries as pre-existing segmentation.

Narrative

This project presents part of the collection ”Word War I Diaries” held at the SLNSW. The artefact is an interface to that collection that learns from the experience acquired with Visference and Crossreads: from Visference it takes topic model analysis and curated labelling and from Crossreads the narrative multiplicity and the fragmented reading. Diggers Diaries was developed in two stages. For version I a smaller corpus of diaries was selected and the research was focussed on the analysis and its consequences for the contents. In this first version, a topic model was used to classify each page and topics were manually grouped and labelled . The process is explained in detail in section 6.1.

The dataset of this artefact is larger. Additionally, the techniques for scoring similarities among segments of the dataset are different, more appropriate for the task of crossreading, similarly based on topic model analysis

Collaborators

None.

Data process

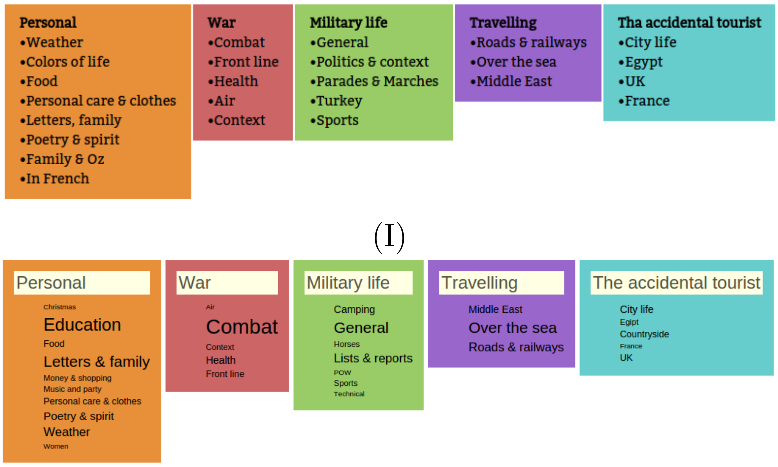
- Data was downloaded from the SLNSW's transcriptions site. This task required web spiders and scrapers since the API did not support the downloading of what was needed.
- • The dataset is a collection of handwritten notebooks. For the segmentation of texts, the handwritten pages were taken, which eliminated the need for a segmentation method. The size of each page is similar to the size already tested in Visference.
- • The downloaded data was then stored into a data model that included text and metadata of the diaries as well as the outputs of the text analysis. The newly formatted data was saved in JSON files freely accessible online.

Data analysis

The strategy to analyse the collection was modelled on that adopted for Visference, but the topic model analysis is revisited to incorporate the experiences from the previous artefact development. Several different sets of topics were generated and compared by running the curated labelling process on each individual output. They were done with 5, 10, 20, 30, 40, 50, and 100 topics.

The process of curated labelling showed that the tests with 10, 20, 30, 40, and 50 did add more semantical topics as the number of topics grew. However, for the test with 100 topics, there was no remarkable increase of differentiated semantical topics, but repetitions and overlapping of them. Furthermore, from the experience of Diggers Diaries I, the goal was to have a menu of topics that would not make topic selection difficult for the user.

The number of topics that created better topics and levels was 50. These 50 topics were assigned to 25 labels grouped into 5 categories: personal, war, military life, travelling, and the accidental tourist. See the two-level labels in Figure 12.



(I)

(II)

Figure 12. Two-level curated labels for Diggers Diaries versions I and II. Version II incorporates a tag cloud relative to the number of pages under each label.

Topic model analysis was conducted with the open source toolkit MALLET, a Java-based package for statistical natural language processing (McCallum, 2002).

Interface development

This interface is based in two layouts: the overview, and the reader.

The overview offers a list of the 126 diaries by authors' names in which every page of each diary is represented with a square. The squares are coloured according to the five categories of labels when the topic with the top score is over a threshold set with a dropdown menu. The default value for the threshold is 20%. The pages can be filtered through topic groups, coloured menus, and submenus.

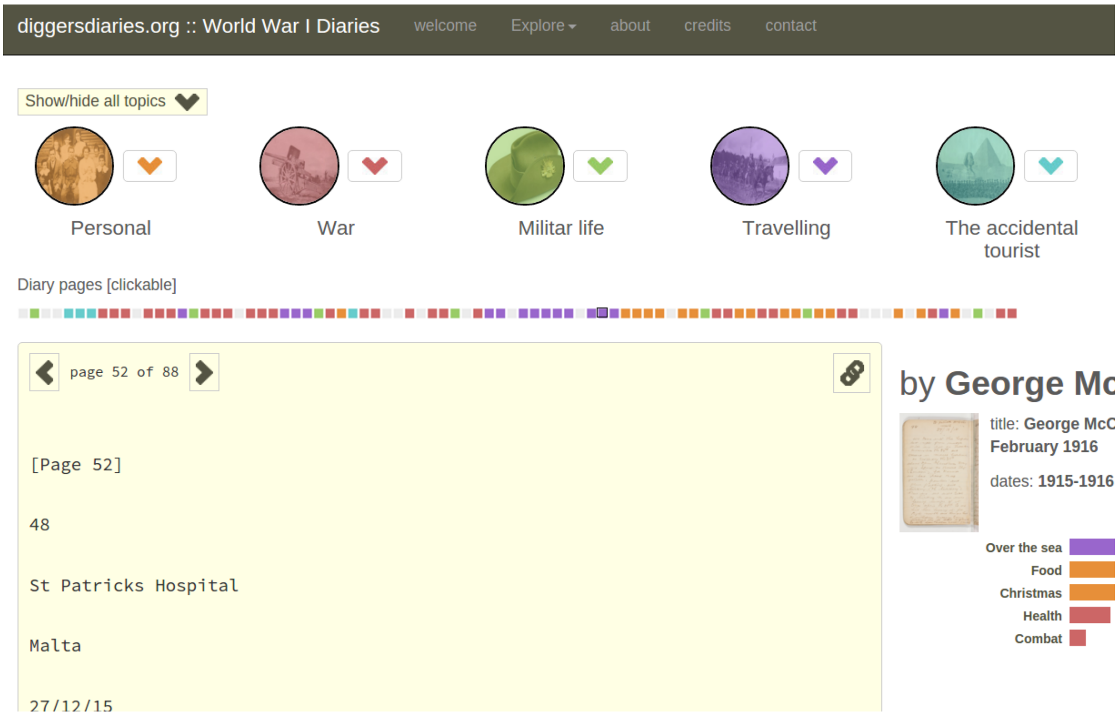
Two ways of interaction and navigation are supported: clicking on a page takes the user to the reader, and clicking on a label from the two-level menu filters the pages related to that label.

The reader shows the text of the current page, the metadata of the diary (author, title, dates), a topic chart for the page, the image of the cover of the diary, and the navigation options previous and next page within the diary.

Project outcomes

- The demo site: `http://diggersdiaries.org/1`.
- The datasets generated after the collection and transformation of the texts is accessible for reuse as JSON files. All the links are in `http://diggersdiaries.org`.

5.5. Diggers Diaries II: WWI Diaries



[Page 52]

48

St Patricks Hospital

Malta

27/12/15

by George Mc

title: George McC February 1916

dates: 1915-1916

Figure 13. Screenshot of the reader of Diggers Diaries II interface.

Artefact name and description

Diggers Diaries II, an interface to explore crossreading of large collections of documents through topics.

Datasets

685 diaries from WWI as part of the WWI Diaries collection hosted by the SLNSW using the handwritten 81763 pages of the diaries.

Narrative

As in version I, Diggers Diaries II focusses on the interface development. Since in version II the collection is about seven times larger, it became necessary to introduce some constraints that were not needed in Diggers Diaries I. To cope with the number of diaries (688) and in particular with the number of pages (81763), in Diggers Diaries II the pages are grouped into coloured squares that represent whole individual diaries. When a page is clicked on, the reader rolls up and a bar chart with the five most relevant topics for that page is shown.

Furthermore, when the user clicks on a new topic, the system jumps to a new page in the chosen topic. This mechanism adds crossreading to the Diggers Diaries by means of a two-level category menu constructed with topic model analysis and labelling. The demo site only shows crossreading applied to the collection of WWI Diaries; a production version would include more features to save lists of pages in a user's account, build playlists, and other features.

The size of the dataset is even larger than in Diggers' Diaries I (approximately by a factor of five). As a result, more combinations of possible reading paths are possible. The interface is primarily oriented to reading, and only secondarily to exploring and browsing the collection.

Collaborators

None.

Data process

The steps for the data process are similar to the ones described for the previous artefact. Some non-conceptual changes were necessary because the SLNSW system changed after Diggers Diaries I was completed.

In addition to the page transcriptions, it was also possible to download the scans of the original pages. The 81763 images added 45GB to the 77MB needed to store the rest of the data and the code. Given the amount of data available, it is appropriate to point out that through the Diggers Diaries II interface only the result from a query is downloaded to the user's computer.

Data analysis

The main difference from Diggers Diaries I is that the number of topics was increased from 50 to 100. After removing and joining the topics identified by Mallet, they were manually grouped under 30 labels. The labels were then grouped into the same five categories defined for version I (i.e., personal, war, military life, travelling, and the accidental tourist).

Interface development As already mentioned, Diggers Diaries II needed a redesign of the interface due to the large dataset and the large number of pages represented. But the concept of exploration for reading was maintained.

There are three overview formats:

- Facetted: Diaries/authors (685 diaries)
- Timeline: Dates/duration (685 lines)
- Pixel/matrix: Diaries/pages (81763 pages)

The structure is reading-oriented because the text reader, embedded in the overviews, is accessible with a single click from any of the three overviews.

The reader is also accessible directly, as the exploration menu includes the reader in addition to the three overviews. Two ways of text reading are available:

- Linear reading: natural reading of diaries, page by page.
- Crossreading: jumping from one page to another according to the two levels of labels. When a label is clicked on, the reader displays a new page that includes among its main topics one belonging to the selected label.

In order to reach every page with the smallest possible number of clicks, static images representing all the pages in each diary were created. The images were subdivided into small areas mapped to the individual pages, so that a click on an area would display the corresponding page in the reader. This image-grid technique allows for efficient interaction with a large amount of items in a standard HTML page.

Project outcomes

- The site in production: `http://diggersdiaries.org`
- On July 26th, 2015, version 1 of the Diggers' Diaries was presented to Richard Neville, Mitchell Librarian and Director, Education & Scholarship (SLNSW). His feedback generated ideas that were implemented in version II (e.g., to offer representations of the diaries by time and author. He also said that SLNSW could perhaps adopt the application in their site.

6. Discussion

This chapter describes the lessons learned from the creation of the collection interfaces described in the previous chapter and from the literature and practice outlined in Chapter 2. It offers a reflection on the making process and hands-on working with the data and in relation to existing concepts, practices, and conventions in the field. These lessons and the research outputs have generated concepts and methodologies to create interfaces to text collections oriented to support reading and the exploration of the documents within digital collections. These are the interfaces that this dissertation has named “deep interfaces”.

The chapter starts with three sections about three interrelated components: text-analysis, text-reading, and text-collection interfaces, developed with a reflexive methodology in order to create a group of related artefacts. The final, fourth section defines the concept of deep interfaces, which integrates the lessons learned through the three components.

The first component discusses how the apparently ”miraculous” effectiveness of text analysis has been applied to the creation of the interfaces presented here. Despite the incredible advances of text analysis, the results of applying analysis techniques to improve the interfaces to a collection of texts are not widely used, as shown in Chapter [2. This section shows the creative process that includes the use of topic model analysis to classify documents of a collection. The importance of topic model labelling in the integration of analysis output as elements of the interface is also shown.

The second component considers the task of reading at a time when it is difficult to find even a moment for it. How can an interface to a large text-document collection take this into account and offer features dedicated and oriented to support reading? How can we read, or at least have a glimpse of, such vast collections of texts? Is there a way to discover and explore a text collection mainly through reading the texts rather than through viewing representations of the collection as a whole? As seen in Chapter 2, several authors have experimented with fragmented narratives and data multiplicity (i.e., the multiple narrative lines that a text can contain). This idea of fragmentation, in combination with the nonlinear reading of crossreading, is presented as a strategy to deal with large text collections.

The third component, ”The Interface and its Codes”, discusses the lessons learned from building the interfaces in relation to conceptual interface design. The interface style that has been developed seeks simplicity in the sense that new users of the interface do not need any explanation to be able to start using it.

The final section, ”Deep Interfaces”, draws together all findings of the previous sections. It proposes deep interfaces as a conceptual umbrella that advocates the development of specific methods and strategies when designing interfaces to large text collections.

In summary, this chapter answers to the research question (see Section [sec3.2-:Research-questions-and]) 3.2):

How do we create interfaces to text document collections that let us explore the collection by reading its contents?

The short answer to this question is: “by creating deep interfaces”. A deep interface entails a set of strategies that make the interface-data binary deeper in meaning and richer in data structures and relationships. The discussed strategies include: simplicity of the interface, the use of text analysis outputs integrated in the interface as visual elements, and a way to read texts that are too long to read in their entirety.

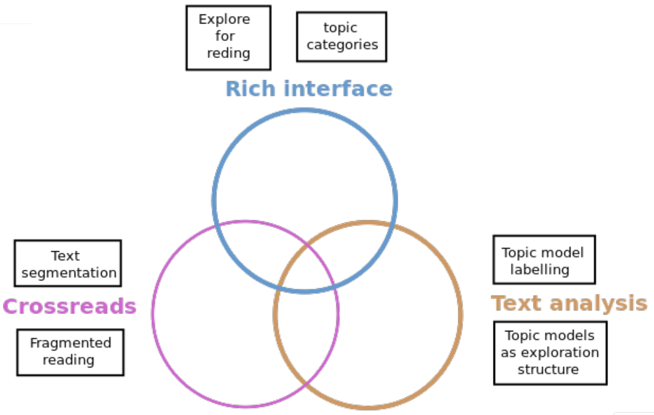


Figure 14. The three components of deep interfaces

6.1. The miraculous analysis

Visference: topic models and topic models labelling.

In the development of Visference — a sortable table for the listing of scientific conference papers — a topic model analysis was applied to a collection of 282 papers published in the JMLR proceedings (Journal of Machine Learning Research). This analysis was done in two phases. In the first phase, a golden list of ten topics was created, in order to represent key concepts and fields of enquiry in machine learning. In the second phase, the collection of conference papers was classified according to this golden list.

In more detail, in the first phase a topic model analysis was conducted on a large set of scientific texts related to machine learning and statistics. Over ten thousand papers and about twenty books were included in the corpus; the papers were downloaded from the open access repository Arxiv.org, and the books were hand-selected by experts (researchers from the Machine Learning Research Group, at NICTA, Australia). This corpus was used as a ”golden” or authoritative dataset of texts representing the machine learning discipline. From the topic model analysis of this corpus, ten main topics were determined. The purpose of this analysis was to find the ten main topics in the field of machine learning research. The ten topics were then labelled by experts. The final choice of labels and the books in the corpus are listed in Section 5.1 Visference.

In the second phase of the process, the scientific papers from the "JMLR Workshop and Conference Proceedings" were segmented by sections and classified according to the ten golden topics. Since these topics aim to define the main categories that represent all content related to machine learning, the goal of classifying papers by topic was to help researchers interested in the conference to easily find papers that could be in their areas of interest. Using the trained model (i.e., the model obtained with the mentioned big corpus, a topic inference was conducted for the almost three hundred papers to be analysed. Each paper was segmented by section, and the final topic model that represented the paper was an average of the topics contained in their sections. Visference shows two aspects in relation to text analysis:

- There is a significant opportunity to improve standard lists of papers in academic conferences. It seems surprising that scientific conferences, especially those about data analysis, still use very traditional and non-interactive layouts for long lists of documents.
- More specifically concerning the topic-model analysis conducted in Visference, it is not very appropriate to use topics to represent a whole scientific paper. A scientific paper has a length that requires segmentation in order to apply classification by topics.

Crossreads: text segmentation, and text similarity

Crossreads shows a collection of articles in the context of an exhibition (see 5.2 and 5.3). Every exhibition has its particular conditions, but they have at least one aspect in common, which is the limited time that visitors spend in front of a piece. Translated to a collection of texts, this time dependency is even more relevant, since the time needed to read a text is quantifiable depending on a minimum value of read words per minute. In order to solve this first question, I opted to segment the texts into small pieces that could be read in approximately one minute each. After reading one of the pieces, a new related piece of text from the collection appears in front of the reader. This idea of breaking the unity of the collection texts is discussed in more detail in the following section 6.2, which addresses the criteria for text segmentation and for determining the similarity of text segments.

The text segmentation is explained in the conference paper published after the development of Crossreads (Nualart and Ferraro, 2014).Two segmentation approaches were tested, each with different benefits. In both versions of Crossreads, each document was divided into segments consisting of one or more paragraphs. A segment length was about seven hundred characters in total, which is what an average adult reads in one minute (Williams, 1998). Segmentation was procedural (i.e., using computers) in Version I and manual in Version II. While the method used in Version I was fast and capable of processing large collections, the method applied in Version II allowed for greater quality of segmentation according to a more complex understanding of the context.

The reason for these two approaches is that the segmentation task is very subjective and, in Version II, a human expert could add a personal view to the segmentation. A machine produced segmentation like that employed in Version I can accomplish its task according to the size of each segment, but cannot be expected to interpret the content of the text like an expert human reader would. Both methods are compared here as part of the process of practice-led experimentation.

To develop the similarity calculus between segments necessary to create the Crossreads network, the following off-the-shelf Natural Language Processing tools and techniques were used:

- Tokenisation: words in the segments are separated by whitespace and punctuation characters.
- Stop word removal: standard stop word removal.
- Named Entity Recognition: identification and classification of Named Entities in each segment. The OpenNLP Named Entity recogniser applied [2] distilled four types of entities: Person, Location, Organisation, and Others.
- Similarity Calculus between segments.

The similarity between pairs of segments was calculated as the sum of the following factors,

$$Sim(i,j)\text{= TokSim + EntitySim + NESim/3}$$

where TokSim is the token cosine similarity between segments, a common vector-based similarity measure, whereby the tokens of each segment are transformed into vectors and then the Euclidean cosine is calculated to determine the similarity between pairs of vectors; EntitySim is the sum of the Named Entities in each segment, normalised by the number of tokens in both segments; and NESim is the cosine similarity between the Named Entities. During this process, the similarity between different NE types (Person, Location, Organisation) was calculated separately and its average is calculated. The path among similar segments was calculated as follows: first, the similarity between each segment of the entire segment collection and an arbitrarily chosen segment i was calculated; second, the segment with the highest similarity value score was set as the most similar segment to ii. Since linear reading of a document is enabled in each iteration, it was decided to skip links to segments of the same document as segment i. Finally, different constraints to each version were applied:

- Version I: In the Crossreads network, each segment is linked to its most similar segment. The drawback of this approach is that links will have a wide range of similarity scores, since in each iteration, the number of segments to be compared with is smaller, and the possibility of finding a segment with a high similarity score decreases. However, the benefit is that there will not be any orphan segments. That is, all segments link to other segments, so that the reader always has the possibility of some crossreading.

- Version II: Each segment is linked to its most similar pairing. To avoid repetition of pairs, segments that have already been set as most similar segments during ten iterations are skipped. After ten iterations, the skipped segments are used again in the similarity calculus.

Crossreads teaches two important lessons that have been applied to the final artefact (the Diggers Diaries). Firstly, the positive conclusion that crossreading is feasible. During the semi-structured interviews to evaluate the interface, the experts saw behind the idea of crossreading texts great possibilities to develop in future. Secondly, concerning text analysis, the experience shows that syntactic similarity is not as strong as semantic similarity. This is the reason that, after comparing the Visference and Crossreads experiences, in the final development of Diggers Diaries the chosen analysis was semantic (i.e., topic model analysis), as shown in the following paragraphs.

The syntactic structure of a text, or a segment of a text, refers to the parts of speech (noun, pronoun, adjective, determiner, verb, adverb, preposition, conjunction, and interjection) and the kind of sentence (declarative, imperative, interrogative, and exclamatory). In the case of Crossreads I and II the similarity score obtained with a quite small corpus (700 and 1500 text segments respectively) were very low. Similarity was calculated as described in 6.1 [Add Reference]. At the same time, segments with the same first topic, as in the Diggers' Diaries, were easily identified as similar. I intend to measure this point in a future user study.

Diggers Diaries: natural text segmentation and topic model labelling

Diggers Diaries is the last artefact and, following the reflexive methodology, incorporates the lessons learned in creating the previous artefacts. Diggers Diaries is an interface for a collection of letters and diaries from Australian World War I soldiers, digitised and held by the State Library of New South Wales)see 5.4 and 5.5). The collection of diaries contains over eighty thousand pages, grouped into over seven hundred volumes and envelopes. The diaries have been analysed using topic models with specific methods for segmentation and topic labelling.

The nature of the collection, which consists of mostly handwritten pages, brought a pre-existing segmentation that accomplished the conditions pointed out when discussing the previous artefacts Visference and Crossreads, both in terms of text length and of content fragmentation. A handwritten page, sized in most of the diaries as an A5 page, can indeed be read in about a minute. The reader has access to the scanned image of the original page and additional context information, and the controls easily allow navigation to the previous and next pages. The segmentation is intrinsic to the documents, and the reader can comfortably bring up pages and move from page to page. Diggers Diaries provides this functionality and enables the segments to be joined back together. In other words, the page provides a useful fragmentation, not only because it has the right size for analysis, but also because, corresponding to physical pages, it is well readable and readily joined.

The main difference of the two methods is in the freedom afforded to the curator. While in a procedural method a curator follows a set of rules, here the curator has the freedom to break rules, to make a segment longer than the average when s/he considers it to be more appropriate for the contents and the curation aims.

Initially the topic model analysis was tested with various numbers of topics. As explained in Chapter 2, the algorithm used to calculate topic models of a collection of texts allows setting parameters for the number of topics to be generated. In Visference this parameter was set to ten, while in Diggers Diaries, the analysis was done to obtain ten, thirty, fifty, and one hundred topics. The list of one hundred topics was selected to work with because it showed more variability of topics. The second step was to classify each topic as either removable, ill-defined, or a synonym of an already-defined topic. Classifying topics involved a subjective process of reading and interpretation to validate and analyse them. The pieces of text (i.e., the pages of diaries) were read to validate the topics (following the example of Blevins (2010)). This shows how reading and subjective interpretation is involved in the process of constructing the topics. The removable topics were those deemed to be irrelevant (e.g., ”Monday”, “Tuesday”, “Wednesday”, etc.) , nonsense (e.g., “time days leave day good back months home france england week work long weeks great place ago australia camp things”), or topics that related to transcribers notes (e.g., ”diary letter written pages australian note page letters notes book copy war printed august records Australia transcriber’s paper april read”). Since the focus of this interface was to show the collection to library visitors, these notes were not included, although, for other more specific reasons, they could be added as part of the accepted topics. Thirty-eight topics were directly removed. Of the remaining ones, some topics were found to be synonymous, almost-synonymous, or strongly related. For example, the following two topics were merged under the label “Over the sea”: “port board sydney melbourne ship boat left wharf leave ashore arrived bay harbour p.m troops fremantle town sea cape colombo” and “boat ship boats ashore board harbour water wharf aboard ships shore men port alongside small deck beach-side troops coal”. Most of the remaining topics were considered synonymous with others, which resulted in a final list of thirty topics.

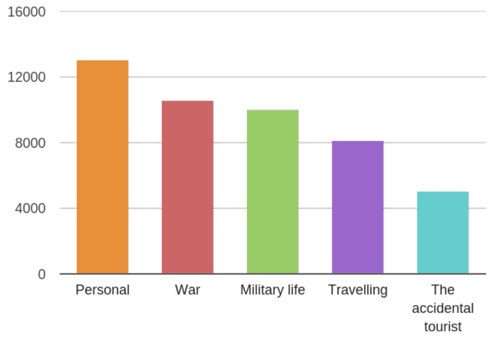


Figure 15. Most common topics by group for the collection of WWI diaries,

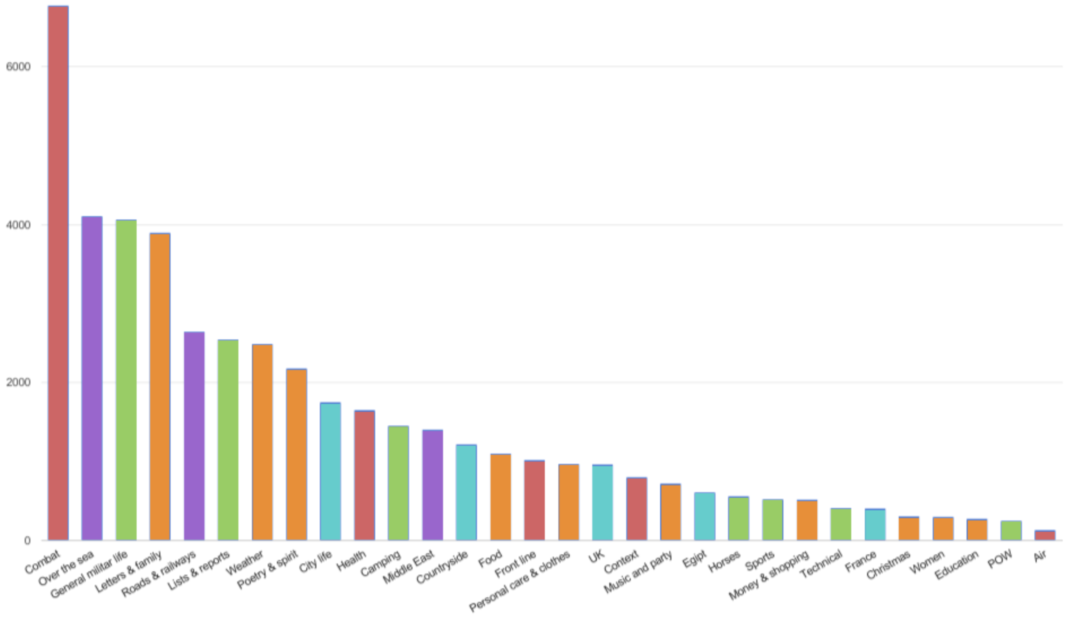


Figure 16. Most common topics for the collection of WWI diaries,

The process of labelling these thirty topics was undertaken in two phases: firstly, the topics were grouped into five general groups. The labels for these groups aimed to be simple and direct: Personal, War, Military life, Travelling, and The accidental tourist. This last label could be simply Tourism, but the chosen label gave a better definition of this topic, which comprises pages where the soldiers talk about impressions they had when visiting European capitals (Paris, London) and Egypt.

The soldiers, many of them young men coming from Australian countryside, had significant experiences in visiting these places, and they shared these with their families through their families through the letters and diaries. ”The accidental tourist” also refers to the 1985 movie of the same name (Tyle, 1985). This shows that labelling does not happen in a cultural vacuum, as labels can include jokes, references, etc. Finally, for each topic some descriptive labels were chosen. The final categories and topics are shown in Table 3. The most frequent topic categories and topics are also represented as two bar charts in Figures 15 and 16.

Groups of topics	Topic labels
Personal	Christmas, Education, Food, Letters & family, Money & shopping, Music and party, Personal care & clothes, Poetry & spirit, Weather Women
War	Air, Combat, Context, Health, Front line
Military Life	Camping, General, Horses, Lists & reports, POW, Sports, Technical
Travelling	Middle East, Over the sea, Roads & railways
The Accidental Tourist	City life, Egypt, Countryside, France, UK

Table 3. Groups of topics and topics labels for the analysis of a collection of diaries, as part of the project Diggers Diaries

The process of labelling described above was developed and evolved through three artefacts with increasing sizes of datasets and number of calculated topics: Visference, Diggers Diaries I, and Diggers Diaries II. In contrast to the described labelling process in Diggers Diaries I and II, in Visference the labelling process was simple: only ten topics were created, and these were labelled according to the opinion of several experts. The results listed in table 3 refer to Diggers Diaries II, which is the last artefact and usually simply referred as Diggers Diaries.

Entity extraction combined with text similarity used in Crossreads brought syntactic text similarity. In contrast, topic model analysis does not reveal anything about the syntactic structure of the texts, but about its semantics. in Crossreads, the entity recognition analysis needed an external source (Baldridge, 2005). In contrast, the topic model analysis conducted in Diggers Diaries is purely based on statistics. There are no external sources that result in any semantic context, as the analysis is purely based on counting words in segments of text.

The whole topic model analysis conducted for the final artefact depends on number of topics, segmentation unit, segmentation method, segment length (which impacts on the reader but also on the analysis because a longer text has more topics), and topic labelling, which includes the following curated tasks: validate, join, group, label groups, and label final topic models. See Table 4 for a comparatison of some of the parameters.

Parameter \ artefact	*Visference*	*Crossreads I*	*Crossreads II*	*Diggers Diaries I*	*Diggers Diaries II*
Dataset size	*papers ˜1800 paper sections*	*700 pieces of text*	*1170 segments of text*	*126 diaries 11944 pages*	*688 diaries 81763 pages*
Analysis	*Topic model*	*Entity extraction + text similarity*		*Topic model*	
Number of topics: initial / final	*10 / 10*	*N/A*		*50 / 25*	*100 / 30*
Segmentation unit	*A section of a scientific paper*	*One or more paragraph of a text*		*One or more paragraph of a text*	*A page of a diary or a page of a letter (most handwritten)*
Segmentation method	*pre-existing (by paper)*	*Procedural*	*Curated*	*pre-existing (by page)*	
Segment length	*Variable, but longer than one minute*	*All segments are close to 700 words*	*Variable. Depending on curation*	*Variable, Around 700 words*	
Labelling method	*Curated simple*	*N/A*		*Curated advanced*	

Table 4. Comparison of some parameters of the text analysis conducted in this research project.

Through the reflexive methodology, the process of text analysis and data enrichment from the analysis was improved across the development of artefacts. The type of analysis was chosen to be topic analysis (i.e., a semantic analysis). The preparation of the text to analyse was changed until, with Diggers Diaries, the pre-existing size of the text segments was more appropriate. Finally, the topic model labelling was improved from the very simple process described for Visference to the more elaborate process of topic selection, removal, and joining of topics conducted in the two versions of Diggers Diaries.

6.2. The fascination for the data

Several strategies have been reviewed with the aim of studying ways to deal with the enormous volume of texts available in DL: from better and deeper analysis that can bring more knowledge about the texts to improvements in the design of the interface to those texts. These two fields, text analysis and interface design, are analysed in this dissertation because they seem to be key to improving our relationship with text collections. However, there is one more field that needs to be mentioned: reading. Since it is impossible to read all texts available in collections, even those of a single big collection, the reading task can be revisited and reinterpreted.

Traditionally, reading is a linear task that goes from the beginning to the end of a text. This approach is simply not applicable when the amount of text is too great. One possible solution to this problem is to divide the text into pieces. That is, text segmentation. Obviously, when a text is segmented, we actually fragment a narrative that was written to be read altogether and sequentially. Nevertheless, in these first decades of the digital age, we find multiple examples of the fragmentation of narrative texts. We find narrative fragmentation in the number of short texts we are exposed to every day through short-text messaging (SMS, Whatsapp-like, etc.), microblogging (Twitter, Tumblr, etc.), and social networks (Facebook, Google Plus, etc.).

Reading and fragmented reality

Three aspects of the digital age are relevant here: our growing capacity to communicate in fragments, the scarce time we have to read, and the overwhelming amount of available text potentially of interest. The question is: could the task of reading be approached as a fragmented task? Several works in the past have explored the possibilities of breaking the linearity of a text. The philosophers Deleuze and Guattari have described the rhizomatic structure of knowledge: ”In a book, as in all things, there are lines of articulation, segmentarity, strata and territories; but also lines of flight, movement, deterritorialization and destratification”. In the novel Hopscotch by J. Cortázar (Cortázar, 1966), the author proposes two reading orders for the chapters; the text starts With: ”In its own way this book is many books, but mostly it’s two books”. Ted Nelson’s Project Xanadu from 1960 (Ted Nelson, et al, 1960) tandards for hyperlinked information that were mostly not included in the standard protocols of the WWW. One of the Xanadu’s features is transclusion, that Nelson defines as ”the same content knowably in more than one place”. One of Xanadu’s seventeen rules states: ”Every document can consist of any number of parts each of which may be of any data type”.

These examples bring confidence and evidence that the traditional idea of reading as a linear task can be developed to include more complex options, like multiple narratives. This idea is the basis for the development of Crossreads, which is analysed in the following paragraphs.

Crossreading: the project and its process

The first version of Crossreads (see Section 5.2) was developed with the idea of text segmentation in mind as a possible strategy to read long texts. Crossreads was developed to visualise, present, and interface a collection in the context of a museum exhibition (”The Listening Eye” MACBA 2014 (Barcelona, Catalonia) (MACBA, 2014). of a collection of fifty-seven articles by Eugeni Bonet, an artist working in cinema and video art since the 1970s, mainly in Catalonia and around Europe.

The peculiarities of that collection include:

- One author: all articles have the same author.
- The topics span across video, cinema, and art, since most of the articles were originally published in program brochures of video sessions.
- The range of published dates is thirty-eight years. This implies that the techniques, and therefore the language of video technology, has changed significantly over the covered period. Also the point of view and discourse has

changed for the past forty years.

Peculiarities of the exhibition context includes:

- *Visitors have limited time to engage with the texts.*
- *Use of projections of the texts on the walls.*
- *Use of tablets for visitors with Crossreads interface to read and browse the collection.*

As in Crossreads I the collection was in Catalan and Spanish, as already said, In order to share the work with more people, a second version with similar texts but in English was developed. In Your Computer (Quaranta, 2011) is a collection of texts written by Domenico Quaranta between 2005 and 2010 for exhibition catalogs, printed magazines, and online reviews. The book is published under Creative Commons licences that allow reuse. In this case the author was contacted as a courtesy and he agreed to have his text used in this research project.

Crossreads is defined as a way of deconstructing linear narrative text in order to be able to follow different paths through it. This project studies data multiplicity and textual visualisation interfaces. The preparation of the text for crossreading starts with segmenting it into small blocks. It then continue by calculating the textual similarity among the segments. A web interface allows exploration of the segments (Nualart and Ferraro, 2014).

As soon as a long text is segmented into pieces and these pieces are read in a different order according to, for example, similarity, a different transfer of information for each chosen order of pieces takes place. This fact seems formally logical since the list of segments and their reading order are different. Then formally, the initial linear data, the text creates an explosion of possible combination of elements. Somehow, this represents a phenomenon of data multiplicity. To study the difference among different sets of segments is beyond the scope of this research. In short, segmentation creates an exponentially increasing set of new permutations of segments.

After Crossreads I and II, a fact related to the length of the segments emerged: some segments had no other similar segment in the collection. The reason for that seemed to be that the dataset was not big enough. A much longer corpus resulting in a much larger number of segments would create more possible similarities among segments. If the number of segments is small, segments with less common content will have few similar segments or none at all.

Diggers Diaries: when crossreading becomes natural

Diggers Diaries is an interface to a part of the collection ”Word War I Diaries” from the State Library of New South Wales (Australia). The collection items are dated from 1914 to 1920. It contains texts from 337 authors (all ANZAC, Australian soldiers), 688 diaries, letters and military reports. So far, a total of 81763 pages, that is segments of text. All handwritten pages were transcribed to text by professionals and users of SLNSW.

This project proposes a way of exploring and reading the collection based on grouping pages according to the topics they talk about.

One of the aims of Diggers Diaries is to help reading text collections that you are not able to read for one or more of these reasons: time rush, text length, low accessibility and usability of the document (small font size or inappropriate colour palette), and legibility of the documents (low quality in the representation of the original document and/or poor conservation of the original).

From the previous artefacts developments, a list of lessons learned has resulted.The first regards the kind of analysis used to find relations between segments of text that belong to a collection. This issue is discussed in 6.2. In short, Diggers Diaries uses a semantic analysis of topics instead of the similarity analysis conducted in Crossreads. To solve the problem of low diversity in text segments due to the small number of segments, a much bigger collection of texts was chosen. Crossreads II had 1115 segments of text, while Diggers Diaries has over eighty thousand segments.

A notable difference with Crossreads is the segmentation process. In Diggers Diaries the since of the texts consist of pages, each segment can be taken to be a page. There is a wide range of text length in the collection of pages. The text in a page is accompanied by the image of the original page, and that strengthens the argument for using the pages as segments. As in Crossreads, in Diggers Diaries the reader can stay in the current diary and go back and forwards or jump to another page. The similar life situation of the authors (mostly young Australian men travelling to the other side of the globe to fight in WWI) generates a common narrative landscape that smoothens the crossreading experience.

A note on the limitations of the presented methods in relation to the size of the datasets. The perceived size of datasets is different in different knowledge contexts. That is, what is big data in one context is small in others. A classic definition says that a data is big when you need special software to deal with it. On the top of that, and related to the software performance, the size of the data that software can manage is also growing everyday. In any case, big or not, the size of a datasets is, in most situations, directly related to the design of the interface to the dataset. This can be questioned in the case of deep interfaces.

Although, in the case of Diggers Diaries, 81,000 diary pages is a lot of text, it is small in comparison to something like Trove's 200+ million newspaper articles (Trove 2009). At this point, a question arises: is the deep interfaces approach practical in these sorts of really big contexts? A short answer would be: it depends on the task that the user intends to do. In the case of Crossreads, the curation task refers to the manual segmentation of reading, with a collection that is too big to be fully read. For Trove a solution similar to Diggers Diaries would be possible, since the interface and the idea of crossreading are independent from the size of the dataset. A big dataset would need a powerful machine to analyse the text

Finally, as a reflection of the relationship between the human and computational parts of the process, that is curated versus procedural respectively, this dissertation has used curated as version I (see 2.3.4. Interfaces and reading). In the case of Visference and Diggers Diaries, the curation task refers to the creation of the topics and groups of topics as a result of an intellectual and subjective decision (see 2.2.3. Topic model labelling and 2.2.4. Topic model evaluation).

6.3. The Interface and its Codes

The interface to a digital collection is the tool that allows us to interact with the collection, but it also sets the limits to all we can see of the collection. From the point of view of an observer using interface, the interface “is” the collection. In fact, the interface is a simplification of the collection that highlights some aspects of it and can hide, in some contexts, its rich complexity. In those cases, the interface plays the role of the surface to the collection and using it means diving into the collection from the outside, towards the treasures hidden inside.

This section follows a chronological narrative that shows the cumulative process of interface development in the artefacts presented. It starts with Visference (see Section 5.1), which introduces new features in text collection interfaces using standard web interface elements. Then Crossreads introduces crossreading as a practice for reading across a text collection. Finally, Diggers Diariesis a text collection interface oriented to support and suggest reading that incorporates all the lessons learned from the previous artefacts, and includes some new features.

Visference: conservative design

Visference is an interface proposed for an academic conference. Academic conferences commonly list accepted papers in a flat, text-based, non-interactive web page or a printed program. The dataset chosen for Visference was a collection of 282 accepted papers from the JMLR Workshop and Conference Proceedings Volume 2: Proceedings of the 30th ICML (ICML, 2013).

The motivation of Visference is to improve the way we list scientific papers in conference websites. A relevant work mentioned is ”Word storm”, by Castella and Sutto (Castella and Sutton, 2014), that lists accepted papers of a conference adding a word cloud with a modification of the algorithm that makes word clouds easily comparable (see 2.2.1).

Visference had a design compromise from the very beginning: there is much room to innovate in conference paper lists but, at the same time, it is necessary to use web conventions that will improve the exploration of the list without the need for an initial tutorial on the interface. This philosophy has been applied to the interface design for all the created artefacts. The documents are presented in a HTML table with sortable columns. The columns are: paper title, authors, PDF link, abstract/reference link, and the ten topics. The sortable columns allow users to see the most relevant papers for each topic, as well as the related topics.

In terms of interface design, Visference is not innovative because of this low-risk design compromise. The innovation in Visference is the presentation of a list of documents that is traditionally flat, non interactive and not sortable. Visference tries to push for innovation in the conservative websites of scientific conferences and, within the development of this research project, represents the philosophy that has been applied to the rest of the interfaces: the use of standard visual elements in the interfaces and making unnecessary introductory tutorials to explain the use of the interface. The following paragraphs discuss the role of interfaces when the task of reading is of priority.

Crossreads: fragmented narratives

Crossreads as a proposal for reading text in fragments has been presented and discussed in 6.2. In relation to interface development, Crossreads is a key component of this research project because it is the first reading-oriented interface. Version I was designed for an exhibition space of interaction and, as such, implied some design considerations: the need to be a responsive design adapted to the devices used in the exhibition space (10-inch tablets); a length of the exhibited texts adapted to a short time visit; and an interface design that uses standard interface elements as in Visference because visitors do not have time to learn how to use an unfamiliar interface.

Another central part of Crossreads in relation to its interface is the design of the reader, which is the frame where the current piece of text is presented to the visitor. The reader is placed in the centre of the screen and highlighted by a border. The interactions that characterise Crossreads are the possibilities to jump to a similar page and to any other random page of the collection. The elements that can be considered ”standard” are: next/previous page of the current document and the timeline that allows browsing documents of the collections by kind of document. These two groups of elements are placed at the same visual level.

Diggers Diaries: a lot of details build an overview

Diggers Diaries is an interface to a part of the collection of diaries, letters and military reports from Australian soldiers —also called diggers— in World War I.

The project was developed in two stages: version I and version II (see Sections 5.4 and 5.5). . Both versions draw on the same collection of WWI diaries, but version I uses a smaller corpus of diaries. In version I the research focused on the analysis and its consequences in the contents of the collection. The development of Diggers Diaries version I in relation to text analysis is discussed in detail in section 5.4. In version II the dataset increases by over six fold in the number of segments of texts, therefore version II development focused on the interface because it uses the same kind of analysis process developed, for the first time, in version I (topic model analysis, and manual topic grouping and labelling). Version II focuses on elements that eventually will help visitors in reading and exploring the collection. Among these elements there are data visualization elements.

Three data visualisation elements are the tools to explore and overview the collections: by pages, by diaries, and by by date:

• By-pages overview is a page-grid visualisation device that represents the smallest unit of content (page) at the whole collection level. It draws on "show everything" type interfaces but, once more, the innovation is in using analysed text content to fill in this view. This by-page overview interface offers several display modes, as well as a colouring scheme. The grid of pages is coloured according to the five categories of topics. The pages can be filtered by category. Since each topic score is a percentage of that topic for that page, the pages are coloured according to the biggest topic score, and only scores bigger than 20% are coloured.

• By-diaries overview is a facetted view of all diaries that can be sorted by date and alphabetically by author names and topic names. When a diary is clicked on, the diary metadata is shown, and a page browser shows the principal group of topics for each page in five colours. Then, when the page is clicked on, the reader appears integrated in this diary view. Other diaries can be expanded at the same time, so that multiple diaries can be read in parallel.

• By-date overview shows the start and end dates of each diary, and, consequently, also its duration. This overview allows selection of contents according to the historical moments of the collection (i.e. the reader can easily select texts from the beginning of the war, from the end, etc.). Diggers Diaries gives several options to explore the collection from the home page like it was done in ”Explore Australian Prints + Printmaking" (Ennis and Whitelaw, 2014), and the Eugenics

Archives (Collective, 2016). See 2.3.3 for more details. The home page includes the reader options and, at the same visual level, the three mentioned exploration options. The reader options take the visitor to a page of the collection and invites him/her to crossread the collection by jumping to pages via a two level menu of topics (see Section [sec 6.1).

The reader interface in Diggers Diaries introduces a topic to be developed in future research projects (i.e., the inside-out exploration of textual document collections) and the idea that the construction of an overview of a collection of texts (and therefore, of its contents and context) can be created from samples of the collection and not necessarily from its overviews —or distant reading—.

This section has shown how elements and concepts related to text collection interfaces evolved though the creation of artefacts for real collections. From Visference and its compromise to create a simple interface based on text analysis to Crossreads and its fragmented narratives as a strategy to read unreadably large collections of texts, as well as its way to build multiple narratives upon one text. Finally, Diggers Diaries, presents the first deep interface in which standard visual elements used in traditional interfaces are at the same level as elements that come directly from text analysis outputs (i.e., buttons to select pages within the collection that semantically corresponds to topics). The next subsection defines and discusses the concept of deep interfaces that resulted from the lessons learned when developing the presented artefacts.

6.4. Deep Interfaces

This section presents the diversity of findings that this research project has produced and integrated into an umbrella concept: deep interfaces. The construction of the definition of deep interfaces is a cumulative process through the experience of developing the presented artefacts (see Chapter 5).

Deep interfaces is a wide concept used to compile the experiences of developing interfaces to textual document collections. It refers to interfaces that on their surface resemble a standard web page while allowing deeper interpretation of the meaning and the structure of the contents of the collection. Deep interfaces achieve this result by integrating text analysis elements into the collection interface.

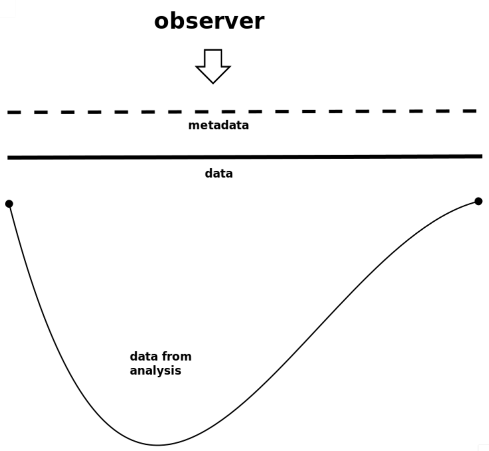


Figure 17. Deep interfaces from the point of view of data:: metadata, data, and data from analysis.

In the case of Diggers' diaries, "data" refers to the diaries themself; while "metadata" includes the diaries' authors, titles, and start and end dates. "Data from analysis" refers to the generated data that classifies each diary page with the score of each computed topic model. the curving away line is actually an indication of the increased depths that the user can explore.

Whilst text analysis is widely applied in more fields every day, there is a gap in its use for digital collection interfaces. In the digital humanities, text analysis is used for so-called "distant reading", mostly dedicated to overview and explore the collections rather rather than to improve the reading of their content (see Section 2.3). Deep interfaces fill this gap, incorporating text analysis into the interface to text collections as visual elements.

Text analysis of every text of the collection adds a new layer of metadata that enriches the information structure of the collection, which is usually only represented by standard metadata fields. New structures within the documents of collections offer new relationships among them that result in new meanings and ways to interpret the collection. In this context, "deep" is about the volume of data exposed to the visitor via the interface, and text analysis is what enables document content to be represented at the interface.

In order to get to know big textual collections —those too big to be fully read— the strategy of crossreading has been practised during this research project. That is, to cut the documents of the collection into pieces according to some criteria, and then read part of the collection with the help of links among those pieces. Crossreading is presented as a fragmented narrative reader. According to this idea, collections of texts can be overviewed reading small parts of them. That is, a reader can construct her/his own very subjective idea of the collection based on details of the collection. To generate an overview of the collection is not the goal of crossreading but, in a sense, crossreading the collection eventually generates a view —or opinion— of the collection.

Crossreading is a concept that complements the idea of building interfaces for reading, or oriented to read, a text collection. In a reading interface it is not necessary to explore the collection, nor to overview it, before starting to read texts the collection contains. Deep interfaces are oriented to take the visitor of the collection directly to one of its items, no matter how big the collection is. Deep interfaces can be explored by diving into the collection in this way and then jumping to other items according to topics or similarities among the items (or parts of the items)d. The texts of the collection are effectively accessible from the home page, or only a single click away. At the same time, this proposed system or strategy to build interfaces to document collections is not incompatible with other standard techniques for exploration, like data visualisation techniques, to offer the items of the collection according to some standard metadata fields such as: timelines, geo-maps, faceted lists by authors, etc.

This practice-led research has developed several interfaces to textual document collections with the technical aim of bringing together textual collections with the power of text analysis. Its broader goal is to generate ideas to bring to life knowledge that lies stored in institutions in the form of textual document collections. Since this research project methodology follows a reflexive—cumulative—process, the final compilation of lessons learned and implemented is presented under the all-encompassing label of deep interfaces. The idea of deep interfaces is a concept that rests on three pillars: the way we analyse textual documents and use its results, the way we read long texts, and the way we build interfaces for textual collections.

The following last chapter recaps the whole research project and points out ideas to develop in the future related to interfaces to digital collections.

7. Conclusions and Future

This practice-led research project has produced accessible interfaces to real textual document collections that offer new ways to interact with these collections. The experience of creating those interfaces has led to the concept of deep interfaces. This approach assembles the findings of this research into a set of recommendations for building new interfaces to text collections in which features are derived from state-of-the-art text analysis and dedicated to support the reading of texts within the collections.

This dissertation has explained the rationale, the results, and the overall experience of creating interfaces to text collections. As the introduction showed, the context, motivation, and aims for this research are linked to the growing amount of text available —some coming from a digitisation process, some directly born in digital form— and the urgency for better means to interact with that quantity of information and its posthuman scale.

A review of current practices in major public digital libraries revealed consistent features, such as the aggregation of contents from other institutions, the dominance of traditional text-based interfaces, and the sometimes limited access to the objects of the collections. This review also considered cases that offer innovative ways to reuse their collections, like the New York Public Library, which offers a snapshot of the public contents for download and reuse. As a general view, while institutions are beginning to innovate through these created "labs" —departments of the institutions that experiment with its contents— the official sites remain very conventional.

A brief review of the field of text analysis introduced its standard methods for finding similarities among texts, such as document clustering, topic model analysis, and related work in topic labelling and evaluation. While text analysis techniques like topic models have been applied to text collections in the DH, they are not used to support interfaces aimed at reading, but rather to do ”distant reading” analysis. During the research project, most of these techniques were used to manipulate the original collections of texts and to create interfaces to interact with them. This project uses semantic similarities calculated with topic model analysis supported by a curated process of topic clustering and labelling to make the statistical analysis human-readable and usable.

The third literature review considered the development of interfaces to digital libraries with a specific focus on text collections. It reviewed the theoretical framework as well as new paradigms and metaphors, proposed by various authors, that have influenced interface development for the last twenty years, starting with the popular “Overview first, zoom and filter, then details-on-demand” and “Previews and Overview” by Schneiderman and their collaborators, moving on to “Show everything” by Stamen, and, finally, more recent proposals of “Information Flaneur” by Dörk, and “Generous interfaces” by Whitelaw. These metaphors brought fresh air to the design of new interfaces and new features such as the opportunities beyond the search box and support for serendipity facing task oriented design. These sort of features contribute to the freedom of the user when interacting with the data represented in the interface.

With this idea of innovation with simplicity, a review of standard practices in interfaces to text collections and the use of data visualisation as interfaces to these collections showed that most innovative proposals come from practitioners and researchers that work outside major institutions and from work in data visualisation rather than digital libraries. Finally, a brief review of the concepts of reading texts on screens (i.e., e-reading) found that the standards for e-readers established three decades ago have today become accepted and, therefore, are included as common features in all kinds of digital readers.

From these multiple literature reviews several gaps and opportunities for new research were identified and introduced in Chapter 3. One key gap is the lack of the use of interfaces not based on a traditional search box in official sites of major public institutional DL. Innovation in these interfaces is found in projects from external practitioners and researchers. With a promising tendency, institutions are beginning to include some of these external projects in their official sites. See for example the works of Whitelaw (`http://mtchl.net/`). Following this tendency, Crossreads I was developed outside of the institution but included in its official site (see Section 5.2).

While computational text analysis has been applied in fields such as information retrieval for decades and, more recently, in literature, history, and the digital humanities, text analysis is not used in current interfaces to text collections. As a contribution to knowledge, this research project uses text analysis outputs as visual elements integrated into large-scale collection interfaces. The study of reading as a task shows that it is impossible to read text collections at the scale considered here. In response to this issue, this project proposes the concept of crossreading, which is a method to limit ourselves to a sample of an unreadably large collection of texts. The idea is to divide long texts into segments and, aided by a computer recommendation, jump from segment to segment according to topics or other relevant features.

In order to explain the process followed in these experiments and make them repeatable with other collections, Chapter 4 explained the project's methodologies and methods. The main methodologies include reflexive research, which makes the whole research a cumulative process building on the lessons learned and transferred from artefact to artefact. In that sense, all the project's key advances have been applied to Diggers Diaries, which is the artefact that was developed last. The practice-led and a practice-based nature of this research means that theory can arise from practice. Concerning the methods and tools used during the research project, a software list and links to all source code and data are provided, in order to encourage other researchers and practitioners to reproduce and improve the methods presented here.

Chapter 5 described the results of the project in detail. Five practical projects were introduced: Visference, Crossreads (I and II), and Diggers Diaries (I and II). The chapter also described the process of creation, including data gathering, transformation, analysis, interface development —libraries and languages used— evaluation studies, and project outcomes. A brief narrative explained the experience and highlighted issues related to each artefact.

Drawing on these artefacts and the lessons learned from the creative experience, the discussion chapter outlined a narrated timeline from three key points of view: text analysis, reading, and the interface. These components are the three "pillars" that support the project's central proposal for deep interfaces. Firstl lef, the "miraculous" text analysis, offers rich opportunities for experimenting with text analysis outputs to be integrated into user interfaces to text collections. The second leg responds to the question of how we can encompass the posthuman scale of digital text collections now accessible online. This question is answered through the creation of Crossreads (I and II) artefacts and the concept of crossreading, which the final project, Diggers Diaries, builds upon. It can be said that the melting pot of experiences during the practical creation of the artefacts has been used in Diggers Diaries (the final artefact) as all the lessons learned at a theoretical level are used there to define the concept of deep interfaces.

This chapter continues with a description, with examples, of the limitations, potential applications, and future work prompted by this research. The chapter ends by briefly drawing together some thoughts to conclude this dissertation.

This project presents significant contributions to the creation of new interfaces to digital text collections. At the same time, its limits should be acknowledged. Each interface works with a real collection of digital texts. Using real collections demonstrates the validity of these techniques, but it also means that the design of each of the interfaces responded to specific constraints and contexts of development. These constraints also introduce specific limitations related to the design and features of the interfaces.

While they offer important new approaches to text collection interfaces, the artefacts created in this research project are rather conservatives when it comes the actual interface design. This choice was dictated to avoid the need for introductory tutorials to explain new features and interactions.

Furthermore, the features offered in the interfaces are quite limited. For example, none of the interfaces offers a full text search box for the collection. In part, this is done for pragmatic reasons, to avoid server-side queries and propose a simple, client-side application. This limitation was deemed to be acceptable because the sites that host the collections already offer full-text search and the focus of this research project is to propose new features instead of reproducing existing ones.

The concept of crossreading and the design of the reading experience are based on the personal and subjective experience of working with the collections. To support further work on reading collections it will be necessary to know more about the reading process and to test and validate approaches such as crossreading.

Another limitation, especially for the case of Diggers Diaries, is one of data synchronisation, due to the fact that the collection of diaries hosted by the State Library of New South Wales grows as new donated diaries are digitised. To add them to the Diggers Diaries interface would require the data analysis to be rerun and an updated dataset to be rebuilt. At the moment, this is a process that one person could do in a matter of days. Depending on the volume of new data, topic analysis, clustering, and labelling would also need to be redone, which would be a more significant task.

There are two general preconditions that future applications of the methods defined in this research project should take into consideration: data access and licensing. In projects where text analysis is done at a collection-item level, full text access to every object of the collection is required. In those cases, accessing the data will require some technical expertise in order to use APIs or scrape the contents. In any case, the data need to be under a license that allows reuse, modification, and publication.

In relation to the nature of the collections, there is a question about the content constraints for the application of crossreading to a collection of texts. In this research project all the collections had some homogeneity. In the texts used in Crossreads I and II, each collection had a single author, although the texts were diverse in format and structure as they included interviews, opinion pieces, presentations, catalog notes, etc. These formats impact on the segmentation process that is part of the preparation of the collection for crossreading. In Diggers Diaries, the collection of diaries was highly homogeneous. The personal situation of the young men far from their families on the front line or at the opera in London, led the authors to share similar feelings, experiences, and situations that they expressed often in a similar style, in line with their common cultural background. The experiences with these collections, very different in both structure and content, is not enough to draw general conclusions. Certainly, more research is needed about the nature of texts suitable for crossreading and the way we read digital texts.

Deep interfaces could be applied to a range of other collections. Inspired by "Mining the Dispatch" (Nelson, 2010a), the deep interface concept could be applicable to crossread press news, which would make possible to jump to other articles with similar subjects. Another field that could be tested is poetry, jumping across authors, styles and topics, whilst reading the poems.

Still regarding interfaces to text collections, one more subject to work on in the near future is to revisit and expand the concept of digital reading of rich documents. Digital reading can be understood as reading, listening, and/or viewing. Text documents can be complemented with images, audio, and video from similarly accessible collections, so that these collections can not only be crossread, but also cross-listened to and cross-viewed.

The software produced in the project has been designed to be applied easily to new collections. From a technical point of view, this project is completely described in order to assure reproducibility and, even more importantly, repeatability with other collections of texts. All the techniques, source code, data, and other related documents are accessible in the respective repositories (Nualart, 2016). The software necessary to repeat the work is freely available and the libraries and computer languages used are generic.

As a possible real application of the methods, a start-up is currently considering using the techniques described in this project to build a deep interface to legal texts for experts and professionals. This collection would include all the laws and all the court sentences of a specific country. One of the challenges of this project is to work with a dynamic collection of texts, as is the case with court sentences. Probably the system would need to rebuild the model regularly to include the updated content.

This work shows the feasibility of creating deep interfaces to text collections, exposing the content of collections to improve their readability. A future challenge will be to implement such interfaces for other collections. Institutional collections could lead the way, but there are several factors that might slow this development. Challenges for institutional collections include limited resources, a reliance on proprietary (vendor provided) DL software, and limited in-house technical capacity. Alternatively, deep interfaces might be implemented outside the institutions, like some of the examples already discussed. This option has many advantages, but also introduces difficulties associated with data synchronisation, sustainability, and data access, as servers, databases, and APIs inevitably become obsolete.

At the moment of writing this paragraph, Google has announced the release as free software of their ”SyntaxNet” (Petrov, 2016), an open-source neural network framework that offers state-of-the-art syntactic analysis of natural language. This will empower everyone with access to a standard computer to analyse large amounts of text with a state-of-the-art tool. This can be seen as a further step in the democratisation of text analysis. Beside that, it is worth mentioning once more Mallet (McCallum, 2002), Machine Learning for Language Toolkit, which is free software that implements techniques for text classification, sequence tagging, and topic modelling.

Another avenue for future work relates to the advances expected in data analysis and in the so-called Natural Language Understanding. Deep learning techniques promise significant improvement in performance and in features (Unit, 2012). This will help the process of integrating data analysis and visualisation in education and supporting the development of new ways of using data, new interactions, and, as a result, new interfaces.

In short, the rationale of this research is as follows: we live in a sea of digital text —more than we can ever imagine. In fact, libraries, museums, and archives are digitising and preserving all kinds of materials. The value of all these processes of preservation depends on accessibility: what use is all this if it is not read? Therefore, there is an urgency for new forms of access and interaction with these digital texts. This project faces this urgency, demonstrating the feasibility of deep interfaces for text collections.

Bibliography

Mortimer J Adler and Charles Van Doren. *How to read a book: The classic guide to intelligent reading*. Simon and Schuster, 1972.

Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. Representing topics labels for exploring digital libraries. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 239–248. IEEE Press, 2014.

Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. Serendip: Topic model-driven visual exploration of text corpora. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 173–182. IEEE, 2014.

AAA American Anthropological Association. Cultural anthropology collections. `http://www.culanth.org/curated_collections`, 2016. (Visited on 01/31/2016).

Keith Andrews, Wolfgang Kienreich, Vedran Sabol, Jutta Becker, Georg Droschl, Frank Kappe, Michael Granitzer, Peter Auer, and Klaus Tochtermann. The infosky visual explorer: exploiting hierarchical structure and document similarities. *Information Visualization*, 1(3-4):166–181, 2002.

archive.org. Internet archive: Digital library of free books, movies, music & wayback machine. `https://archive.org/`. (Visited on 01/31/2016).

Jason Baldridge. The opennlp project. *URL: http://opennlp.apache.org/index.html ,(accessed 2 February 2012)*, 2005.

Laurent Baleydier. Wikipedia kartoo, 2001. URL `http://en.wikipedia.org/wiki/Kartoo`. [Online; accessed 12-November-2015].

et al. Bernhardt. Deutsche digitale bibliothek visualized. `http://infovis.fh-potsdam.de/ddb/`, 2015. (Visited on 02/01/2016).

Howard Besser. *The past, present, and future of digital libraries.* Wiley Online Library, 2004.

David Blei. Topic modeling and digital humanities. *Journal of Digital Humanities*, 2(1):8–11, 2012.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Cameron Blevins. Topic modeling martha ballard s diary. *Pers. Blog*, 2010. (Visited on 12/12/2015).

Bono. Data never sleeps 3.0. https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/, 2015. URL `https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/`. [Online; accessed 10-May-2016].

Christine L. Borgman. What are digital libraries? competing visions. *Inf. Process. Manage.*, 35(3):227–243, 1999.

Wray Buntine and Jaume Nualart. Icml 2013 simple topic model. `http://research.nualart.cat/visference/visference-topicmodels.html`, 2013. (Visited on 11/28/2015).

Vctnnevar Bush. As we mov think. *Perspectives on the computer revolution*, page 49, 1989.

L. Candy, S. Amitani, and Z. Bilda. Practice-led strategies for interactive art research. *CoDesign*, 2(4):209–223, December 2006. ISSN 1571-0882. doi: 10.1080/15710880601007994. URL `http://www.tandfonline.com/doi/abs/10.1080/15710880601007994`.

Linda Candy and Cognition Studios. Practice based research : A guide practice and research. 2006.

David Carter and Luiz Fernando Capretz. 3d user interface for a file management system. *IEEE Can. Rev*, 44:13–15, 2003.

Quim Castella and Charles Sutton. Word storms: Multiples of word clouds for visual comparison of documents. In *Proceedings of the 23rd international conference on World wide web*, pages 665–676. ACM, 2014.

Tim Causer and Valerie Wallace. Building a volunteer community: results and findings from transcribe bentham. *Digital Humanities Quarterly*, 6, 2012.

Matthew Chalmers and Paul Chitson. Bead: Explorations in information visualization. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 330–337. ACM, 1992.

Jason Chuang, Christopher D Manning, and Jeffrey Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77. ACM, 2012.

Cisco. The ciscoÂ® visual networking index, 2016. URL http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html. [Online; accessed 10-May-2016].

Jeff Clark. Spot, 2009. URL http://www.neoformix.com/2012/IntroducingSpot.html.

Jeff Clark. Grimm s fairy tale metrics, 2013. URL http://www.neoformix.com/2012/IntroducingSpot.html.

Collective. The eugenics archives. http://eugenicsarchive.ca/, 2016. (Visited on 08/04/2016).

Julio Cortázar. Hopscotch (rayuela). *New York: Pantheon*, 1966.

Fred D Davis, Richard P Bagozzi, and Paul R Warshaw. User acceptance of computer technology: a comparison of two theoretical models. *Management science*, 35(8):982–1003, 1989.

DCMI. Home: Dublin coreÂ® metadata initiative (dcmi). http://dublincore.org/, 2001. (Visited on 12/22/2015).

DDB. Deutsche digitale bibliothek - kultur und wissen online. https://www.deutsche-digitale-bibliothek.de/, 2009. (Accessed on 09/20/2016).

DDB. Startseite - deutsche digitale bibliothek. https://www.deutsche-digitale-bibliothek.de/, 2016. (Visited on 08/04/2016).

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

G. Deleuze and F. Guattari. Introduction: rhizome. *A thousand plateaus: Capitalism and schizophrenia*, pages 3–25, 1987.

Giorgio Maria Di Nunzio. Visualization and classification of documents: a new probabilistic model to automated text classification. *Bulletin of the IEEE Technical Committee on Digital Libraries (IEEE-TCDL)*, 2(2), 2006.

Marian. Carpendale Doerk, Sheelagh, and Carey Williamson. The information flaneur: a fresh look at information seeking. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1215–1224, 2011. URL http://dl.acm.org/citation.cfm?id=1979124.

Erik Duval, Wayne Hodgins, Stuart Sutton, and Stuart L Weibel. Metadata principles and practicalities. *D-lib Magazine*, 8(4):16, 2002.

Butler. Ennis and Mitchell Whitelaw. Australian prints + printmaking. http://printsandprintmaking.gov.au/, 2014. (Visited on 02/01/2016).

Europeana. Europeana - homepage. http://www.europeana.eu/portal/, 2008. (Visited on 12/07/2015).

Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

Gabriela Ferraro, Hanna Suominen, and Jaume Nualart. Segmentation of patent claims for improving their readability. *Proceedings of the 3rd Worshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL 2014*, pages 66–73, 2014.

Paula Findlen, Dan Edelstein, and Nicole Coleman. Mapping the republic of letters, 2011.

Bryant Foo. Navigating the green book | nypl labs. `http://publicdomain.nypl.org/greenbook-map/`, 2013. (Visited on 01/31/2016).

Trevor Fountain and Mirella Lapata. Taxonomy induction using hierarchical random graphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 466–476. Association for Computational Linguistics, 2012.

Edward A Fox. The digital libraries initiative: update and discussion. *Bulletin of the American Society for Information Science*, 26(1):7–11, 1999.

Edward A. Fox and Ohm Sornil. Digital libraries. 2003.

Michael Friendly and Daniel J Denis. Milestones in the history of thematic cartography, statistical graphics, and data visualization. *U RL http://www. datavis. ca/milestones*, 2001.

Edward B Fry. Readability. reading hall of fame book, 2006.

A. Goldst. dfr-browser. `https://github.com/agoldst/dfr-browser`, 2014.

Andrew Goldstone and Ted Underwood. What can topic models of pmla teach us about the history of literary scholarship. *Journal of Digital Humanities*, 2(1): 39–48, 2012.

Yihong Gong and Xin Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25. ACM, 2001.

Dale L Goodhue. Understanding user evaluations of information systems. *Management science*, 41(12):1827–1844, 1995.

Stephan Greene, Gary Marchionini, Catherine Plaisant, and Ben Shneiderman. Previews and overviews in digital libraries: Designing surrogates to support visual information seeking. *Journal of the American Society for Information Science*, 51(4):380–393, 2000.

Seth Grimes. Unstructured data and the 80 percent rule. *Carabridge Bridgepoints*, 2008.

Ralph Grishman. *Computational linguistics: an introduction*. Cambridge University Press, 1986.

Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.

Roland Hausser and R Hausser. *Foundations of computational linguistics*. Springer, 1999.

Jef Heer. A conversation with jeff heer, martin wattenberg, and fernanda viegas, 2010.

Alexander Hinneburg, Rico Preiss, and René Schröder. Topicexplorer: Exploring document collections with topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 838–841. Springer, 2012.

Peter Hirtle. A new generation of digital library research. *D-Lib Magazine*, 5:7–8, 1999.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

ICML. Proceedings of the 30th international conference on machine learning (icml-13). In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013.

Matthew L Jockers. *Macroanalysis: Digital methods and literary history.* University of Illinois Press, 2013.

Steve Jones, Stephen Lundy, and Gordon W Paynter. Interactive document summarisation using automatically extracted keyphrases. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*, pages 1160–1169. IEEE, 2002.

Mehmed Kantardzic. *Data mining: concepts, models, methods, and algorithms.* John Wiley & Sons, 2011.

Rob Kitchin and Nick Tate. *Conducting research in human geography: theory, methodology and practice.* Routledge, 2013.

Krista Lagus, Samuel Kaski, and Teuvo Kohonen. Mining massive document collections by the websom method. *Information Sciences*, 163(1):135–156, 2004.

Jey Han Lau, David Newman, Sarvnaz Karimi, and Timothy Baldwin. Best topic word selection for topic labelling. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 605–613. Association for Computational Linguistics, 2010.

Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1536–1545. Association for Computational Linguistics, 2011.

M Lee, Brandon Pincombe, and Matthew Welsh. An empirical evaluation of models of text document similarity. *Cognitive Science*, 2005.

Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5), 2013.

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive datasets.* Cambridge University Press, 2014.

library of congress. Library of congress. https://www.loc.gov/, 1800. (Visited on 01/31/2016).

Erika C Linke. Million book project. *Encyclopedia of Library and Information Science: Lib-Pub*, page 1889, 2003.

Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

MACBA. The listening eye. http://www.macba.cat/en/exhibition-eugeni-bonet, 2014. (Visited on 11/28/2015).

Pattie Maes et al. Agents that reduce work and information overload. *Communications of the ACM*, 37(7):30–40, 1994.

Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

Xian-Ling Mao, Zhao-Yan Ming, Zheng-Jun Zha, Tat-Seng Chua, Hongfei Yan, and Xiaoming Li. Automatic labeling hierarchical topics. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2383–2386. ACM, 2012.

Gary Marchionini, Catherine Plaisant, and Anita Komlodi. Interfaces and tools for the library of congress national digital library program. *Information Processing and Management*, 34(5):535 – 555, 1998. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/S0306-4573(98)00020-X. URL http://www.sciencedirect.com/science/article/pii/S030645739800020X.

Mario Perez-Montoro and Jaume Nualart. Visual articulation of navigation and search systems for digital libraries. *International Journal of Information Management*, 35(5):572 – 579, 2015. ISSN 0268-4012. doi: http://dx.doi.org/10.1016/j.ijinfomgt.2015.06.005. URL http://www.sciencedirect.com/science/article/pii/S0268401215000614.

D. Masad and S. Nayar. Sec document clustering - david masad and sanjay nayar, css 739. http://www.davidmasad.com/sandbox/FirmClusters.html, 1011. (Visited on 12/08/2015).

Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`, 2002. (Visited on 11/28/2015).

Meeks, E. Documents | digital humanities specialist. https://dhs.stanford.edu/comprehending-the-digital-humanities/documents/, 2011. URL `https://dhs.stanford.edu/comprehending-the-digital-humanities/documents/`.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499. ACM, 2007.

Ian Milligan. Illusionary order: Online databases, optical character recognition, and canadian history, 1997-2010. *Canadian Historical Review*, 94(4):540–569, 2013.

Jessica Milstead and Susan Feldman. Metadata: Cataloging by any other name... *ONLINE-WESTON THEN WILTON-*, 23:24–31, 1999.

Franco Moretti. *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005.

Tamara Munzner. Data types, 23, 2015.

Robert K Nelson. Mining the dispatch. `http://dsl.richmond.edu/dispatch/pages/home`, 2010a.

Robert K Nelson. Mining the dispatch. `http://dsl.richmond.edu/dispatch/Topics`, 2010b. (Visited on 29/04/2016).

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 215–224. ACM, 2010.

Jaume Nualart. Contributors to jaumet/visference Â· github. https://github.com/jaumet/visference/graphs/contributors, 2014. (Visited on 11/28/2015).

Jaume Nualart. Jaume nualart repos at github. https://github.com/jaumet, 2016. (Visited on 11/02/2016).

Jaume Nualart and Gabriela Ferraro. Towards a rhizomatic narrative. 2014.

NYPL. Collections. nypl digital collections. http://digitalcollections.nypl.org/collections?sort=recent#/?scroll=0, 2016a. (Visited on 07/04/2016).

NYPL. Nypl releases hi-res images, metadata for 180,000 public domain items in its digital collections. http://www.nypl.org/press/press-release/january-6-2016/nypl-releases-hi-res-images-metadata-180000-public-domain-items, 2016b.

NYPLlabs. Gutemberg authors. http://tools.nypl-labs.biz/gutenberg/, 2015. (Visited on 08/04/2016).

Kenton O'Hara and Abigail Sellen. A comparison of reading paper and on-line documents. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 335–342. ACM, 1997.

OpenDOAR. Opendoar - home page - directory of open access repositories. http://www.opendoar.org/index.html, 2005. (Visited on 12/07/2015).

OTA. [ota] the university of oxford text archive. http://ota.ox.ac.uk/, 1976. (Visited on 12/02/2015).

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

Stephanie Pappas. How big is the internet, really? http://www.livescience.com/54094-how-big-is-the-internet.html, 2016. [Online; accessed 10-May-2016].

F.V. Paulovich and R. Minghim. Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1229–1236, Nov 2008. ISSN 1077-2626. doi: 10.1109/TVCG.2008.138.

Slav Petrov. Announcing syntaxnet: The world s most accurate parser goes open source, 2016. URL `http://googleresearch.blogspot.be/2016/05/announcing-syntaxnet-worlds-most.html`. [Online; accessed 13-May-2016].

Simone Paolo Ponzetto and Michael Strube. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175(9-10):1737–1756, 2011.

D. Quaranta. *In Your Computer*. Lulu. com, 2011.

Daniel Ramage and Jason Chuang. Dissertation browser | information. `http://nlp.stanford.edu/projects/dissertations/`, 2012. (Visited on 01/27/2016).

Allen H Rehear. Text encoding. *bIGITAL IIUMA\ ITIES*, page 218, 2004.

Lisa Rhody. Topic modeling and figurative language. *Journal of Digital Humanities*, 2(1):19–35, 2012.

Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1986.

Roger C Schank. What we learn when we learn by doing. 1995.

Benjamin M Schmidt. Words alone: Disfmantling topic models in the humanities. *Journal of Digital Humanities*, 2(1):49–65, 2012.

Michael Seadle and William Y Arms. The 1990s: the formative years of digital libraries. *Library Hi Tech*, 30(4):579–591, 2012.

Michael I Shamos. Machines as readers: A solution to the copyright problem. *Journal of Zhejiang University Science A*, 6(11):1179–1187, 2005.

Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343, 1996.

Carson Sievert. Ldavis. `https://github.com/cpsievert/LDAvis`, 2015.

SLNSW. State library of nsw - transcripts. `http://transcripts.sl.nsw.gov.au/`, 2010. (Visited on 12/22/2015).

SLNSW. World war 1 diaries | transcripts. `http://transcripts.sl.nsw.gov.au/project/World%20War%201%20Diaries`, 2014. (Visited on 11/28/2015).

Justin Snyder, Rebecca Knowles, Mark Dredze, Matthew R Gormley, and Travis Wolfe. Topic models and metadata for visualizing text corpora. In *HLT-NAACL*, pages 5–9, 2013.

M Stefaner. X by y project data visualizations. *Moritz Stefaner Information Aesthetics and Ludwig Boltzmann Institute for media. art. research, nd [Retrieved on Dec. 17, 2010] from the Internet: http://moritz. stefa ner. eu/projects/x-by-y*, 2010.

G. Sullivan. *Art Practice as Research: Inquiry in the visual arts*. SAGE Publications, 2005.

Hanna Suominen, Tobias Schreck, Gondy Leroy, Harry Hochheiser, Lorraine Goeuriot, Liadh Kelly, Danielle L Mowery, Jaume Nualart, Gabriela Ferraro, and Daniel Keim. Task 1 of the clef ehealth evaluation lab 2014 visual-interactive search and exploration of ehealth data. In *Proceedings of CLEF 2014*, 2014.

Ah-Hwee Tan et al. Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, volume 8, page 65, 1999.

Ellen Taylor-Powell. Questionnaire design: Asking questions with a purpose. *University of Wisconsin Extension*, 1998.

Ted Nelson, et al. Project Xanadu, 1960. URL http://xanadu.com/. [Accessed: 2014-07-26. (Archived by WebCite at http://www.webcitation.org/6RLo0HzFo)].

Lucy A Tedd and J Andrew Large. *Digital libraries: principles and practice in a global environment*. Walter de Gruyter, 2004.

Miles A Tinker. *Legibility of print*, volume 1. Iowa State University Press Ames, 1963.

LLC TouchGraph. Touchgraph, 2001. URL http://www.touchgraph.com/seo. [Online; accessed 12-November-2015].

TranScriptorium. transcriptorium. http://transcriptorium.eu/, 2013. (Visited on 01/31/2016).

Trove. About trove. http://trove.nla.gov.au/general/about, 2009. (Visited on 01/03/2016).

Trove. Trove press collections. http://trove.nla.gov.au/newspaper/, 2010. (Visited on 01/03/2016).

Yuen-Hsien Tseng. Automatic thesaurus generation for chinese documents. *J. Am. Soc. Inf. Sci. Technol.*, 53(13):1130–1138, November 2002. ISSN 1532-2882. doi: 10.1002/asi.10146. URL http://dx.doi.org/10.1002/asi.10146.

Anne Tyle. The accidental tourist, 1985. URL https://en.wikipedia.org/wiki/The_Accidental_Tourist. [Online; accessed 09-May-2016].

UCL. Transcribe bentham. http://blogs.ucl.ac.uk/transcribe-bentham/, 2000. (Visited on 01/31/2016).

Economist Intelligence Unit. Rise of the machines. moving from hype to reality in the burgeoning market for machine-to-machine communication, 2012.

U.Washington. University of washington digital collections. http://digitalcollections.lib.washington.edu/. (Visited on 12/22/2015).

Marcos Weskamp. Newsmap. *Webdesigning Magazine, June*, page 86, 2004.

Mitchell Whitelaw. Discover the queenslander. `http://mtchl.net/discover-the-queenslander/`, 2014.

Mitchell Whitelaw. Generous interfaces for digital cultural collections. *Digital Humanities Quarterly*, 9(1), 2015a.

Mitchell Whitelaw. Representing digital collections. *Performing Digital: Multiple Perspectives on a Living Archive*, 2015b.

J. R. Williams. Guidelines for the use of multimedia in instruction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 42, pages 1447–1451. SAGE Publications, 1998.

James Wise, James J Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, Vern Crow, et al. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Information Visualization, 1995. Proceedings.*, pages 51–58. IEEE, 1995.

Wen Zhang, Taketoshi Yoshida, and Xijin Tang. A comparative study of tf-idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3): 2758–2765, 2011.

A. appendix

A.1 Visference: Results of the online questionnaire:

Number of participants: 43

- Gender:

Answer	Count	Percentage
Female	12	30.77%
Male	24	61.54%
Other	1	2.56%
No answer	2	5.13%

- How old are you?

Answer	Count	Percentage
less than 20	0	0.00%
20 to 30	4	10.26%
30 to 40	12	30.77%
40 to 50	16	41.03%
50 to 60	2	5.13%
more than 60	3	7.69%
No answer	2	5.13%

- How would you rate your technical knowledge of computers and the web?

Answer	Count	Percentage
None	0	0.00%
Some	0	0.00%
Good	13	33.33%
Very good	12	30.77%
Expert	12	30.77%
No answer	2	5.13%

- This section asks about your attitudes and opinions regarding web interfaces: How often do you use web browsers?

Answer	Count	Percentage
Once a month or less	0	0.00%
Once per week	0	0.00%
Several times a week	1	2.56%
Every day	5	12.82%
Several times a day	31	79.49%
No answer	2	5.13%

- Do you like to encounter new features in the pages you visit the most?

Answer	Count	Percentage
Not at all	0	0.00%
Rarely	2	5.13%
No opinion	7	17.95%
Sometimes	23	58.97%
Often	5	12.82%
No answer	2	5.13%

- Are you happy with the information tools and interfaces that you use?

Answer	Count	Percentage
Very happy	1	2.56%
Happy	19	48.72%
Indifferent	5	12.82%
Unhappy	11	28.21%
Very unhappy	1	2.56%
No answer	2	5.13%

- Have you ever published a paper in an academic conference?

Answer	Count	Percentage
Yes (Y)	19	55.88%
No (N)	15	44.12%
No answer	0	0.00%

This section will ask you to answer a set of questions using the Visference tool as well as the conventional presentation of the conference papers You'll need to visit the two interfaces: JMLR (volume 28) with existing interface: [`http://jmlr.org/`] JMLR (volume 28) with Visference interface: [`http://research.nualart.cat/visference`]

- Which tool makes it easier to answer each of the following three questions: [How many papers were accepted in the conference?]

Answer	Count	Percentage
JMLR existing interface	0	0.00%
JMLR Visference interface	31	91.18%
no difference	2	5.88%
No answer	1	2.94%

- Which tool makes it easier to answer each of the following three questions: [Which papers are related to machine learning theory?]

Answer	Count	Percentage
JMLR existing interface	2	5.88%
JMLR Visference interface	24	70.59%
no difference	7	20.59%
No answer	1	2.94%

- Which tool makes it easier to answer each of the following three questions: [How many papers talk about optimization?]

Answer	Count	Percentage
JMLR existing interface	0	0.00%
JMLR Visference interface	29	85.29%
no difference	4	11.76%
No answer	1	2.94%

- Which of the two interfaces would you prefer for each of these three tasks? [Understanding the topics and themes of the conference]

Answer	Count	Percentage
JMLR existing interface	0	0.00%
JMLR Visference interface	29	85.29%
no difference	4	11.76%
No answer	1	2.94%

- Which of the two interfaces would you prefer for each of these three tasks? [Finding papers related to your personal interests]

Answer	Count	Percentage
JMLR existing interface	1	2.94%
JMLR Visference interface	28	82.35%
no difference	4	11.76%
No answer	1	2.94%

- Which of the two interfaces would you prefer for each of these three tasks? [Exploring new topics and discovering new research in this field]

Answer	Count	Percentage
JMLR existing interface (A1)	0	0.00%
JMLR Visference interface (A2)	30	88.24%
no difference (A3)	3	8.82%
No answer	1	2.94%

A2. A.2 Diggers diaries II: List of the one hundred topics with terms and first labelling draft not

Note 1	Notes 1	ID	Terms
military life		62	0.20607 men time work great made troops number present days part fact large make place officers australian war general order military
		54	0.18175 good time lot back bit boys pretty big things don't chaps night morning round chap day put thing place bad
		10	0.16237 time told back found made man men thought asked officer place put called gave find knew make long decided looked
		6	0.09093 day sunday july friday morning monday today saturday thursday wednesday tuesday afternoon night june good august usual april september quiet
Personal?	thoughts of life	8	0.08711 man life men people war years english good world french young great country women australian work read soldier mind things
?		76	0.08259 time light great air lights long minutes guns smoke sight sound hear men noise round half hour side heard dark
Personal?	homesick	56	0.07774 day leave days time home good spent england year great months xmas dec australia week left weather france return christmas
Personal	Weather	3	0.07364 water day hot sand feet men night dust mud cold good wet heat sleep camp rain bad hard bath dry
?	lanscape	26	0.0708 miles side hill hills country river town view place road high top sea fine distance water long small great large
Personal	Weather	15	0.07068 cold night day morning rain snow weather wind heavy today wet fine raining yesterday pay frost warm ground blowing sunday
??		97	0.06925 time war days long back day france things don't home life months thing hope great feel good place hard ago
EDITOR NOTES		4	0.06825 transcribed page previous blank preceding transcript n/a error photograph html duplicate printed transcription reverse preceeding pages repeat certificate version multi-page
Personal	Letters Family	42	0.06768 letter letters dear received home mother time mail love hope write loving news send father good week writing wrote days
Personal	Morning / breakfast	93	0.06223 night sleep bed good room morning time tea breakfast slept mess till day put camp comfortable hut blankets men tent
Military life	Camping	90	0.06215 camp move orders ready arrived morning a.m left transport p.m horses men night pack march tents back day road moved
War	Front line	30	0.05909 line trench trenches front men night back party wire dug post work company yards firing battalion fatigue digging shell day
Personal	Food	60	0.05713 tea bread dinner biscuits tin food jam good day breakfast beef rations meat butter bully milk eggs fruit eat coffee
Personal / Tourist?	free time Entertainment	61	0.05636 good concert evening show night tea afternoon home dinner party hall music spent y.m.c.a band time room gave splendid enjoyed
Tourist	city life	9	0.05606 town people streets place city women soldiers fine native french large street natives shops round english houses places small buildings
Travelling	Railway	5	0.05508 train station arrived left camp p.m a.m journey leave night morning marched railway hours good trip miles caught reached town
Personal	Letters Family	20	0.0544 hope letter time dear good write don't i'm letters love hear things back mrs give news long glad i've writing
War	Combat	48	0.05434 wounded men killed shell dead man hit head poor shot died back leg left buried bullet badly blown wounds blood
War	Health	21	0.05377 hospital ward sick doctor bed patients bad feeling day cases medical leg put days wounded sister wound morning arm feel
Tourist	city tourist spots	55	0.05035 church building place built fine large stone walls room beautiful wall inside house high years cathedral side round tower small
Military Life	Parades & Marches	66	0.04831 parade march drill morning afternoon camp day route order inspection men marched guard full marching church evening company usual leave
War	Combat	34	0.04759 shells guns fritz gun shell night fire artillery heavy bombardment line shelling trenches firing enemy big close quiet fired day
War	Air force	95	0.0472 planes bombs dropped guns plane air machine night brought aeroplanes fritz enemy german aeroplane flying taube lines bomb flew raid
Travelling	Over the sea	45	0.04578 sea deck day ship passed boat weather night morning land calm board miles sight rough wind port today side hot
Military life	general	74	0.04528 general brigade officers major battalion division officer col capt staff lieut command colonel australian gen men corps birdwood army div
War	News	13	0.04474 news french war german british germans front great troops prisoners france today english fighting papers big australians army germany captured
Tourist	Countryside	32	0.04105 trees green country fields beautiful lovely pretty flowers grass crops tree village fruit garden place road small growing gardens leaves
War	Combat	46	0.04035 turks men enemy trenches wounded attack position left guns fire line killed machine heavy casualties night artillery fighting infantry captured
Personal	Personal care & clothes	80	0.03995 boots pair white socks black clothes wear red issued post hat shirt wearing picture card clothing cap clean trousers uniform
EDITOR NOTES		78	0.03836 diary letter written pages australian note page letters notes book copy war printed august records australia transcriber's paper april read
Tourist	London	73	0.03767 london train back tea met hotel home park place dinner afternoon walked house arrived bus left caught lunch city walk
War	Front line	88	0.03721 line front guns enemy night attack barrage artillery fritz road position forward stunt heavy left prisoners moved advance gun morning
Military life	General	38	0.0368 men officer officers man major court orderly sergeant charge corporal guard company told colonel mess military room put martial private
Military life	Turkey	14	0.0358 horses miles camp turks water night left camels turkish horse back moved camped regt wadi camel arish brigade desert sand
		39	0.03545 road shell wood village mud left shells roads german ground ypres bapaume place holes ruins town blown traffic fritz dead
Military life?		40	0.03498 miles marched march battalion camp night line village albert moved back left days day place move time billets morning big
Personal?	Killing time	99	0.03444 day home tea till bed afternoon evening morning dinner back fine read wrote letters breakfast good usual spent rain sunday
Military life?	General	57	0.03342 round good general men lunch afternoon mess found today evening morning rode tonight colonel returned officers h.q back day bde
Military life	navy	29	0.03286 ships ship troops submarine passed boat boats port british destroyers cruiser submarines board sea speed miles sunk transports night a.m
Personal	Colors & beauty	0	0.03279 sun blue sky white beautiful light bright green red colour black beauty grey night shining world clouds moon gold full
		98	0.03274 day morning afternoon tea jan wrote parade dinner night feb evening letters sunday march letter monday home good tuesday saturday
War?	Health	86	0.03262 wounded station dressing ambulance bearers day stretcher night field post amb cases fritz section back number busy patients line work
Travelling	French roads	27	0.03044 village town left miles road motor amiens billets marched omer french bailleul march arrived back somme place hazebrouck kilos lorry
??	Health	75	0.03021 capt sgt killed pte l.h wounded margin major indecipherable cpl batt jack sick met smith lieut col note bill left
Travelling	Over the sea	69	0.02931 boat ship boats ashore board harbour water wharf aboard ships shore men port alongside small deck beach side troops coal
War	Combat	58	0.02917 turks trenches beach fire night shells firing guns shrapnel turkish gun day morning hill heavy artillery quiet men rifle gully
		59	0.0289 work good day time to-day working hard job days hours things put week fair section to-night making usual duties deal
Military life	Sports	36	0.02886 played won football sports match game cricket afternoon team day playing good race held boxing games play cards officers parade
Military life	chaps & comrades	44	0.02742 band boys men great people passed crowd played gave flags playing soldiers king troops cheers marched past cheering french singing
		53	0.0274 head long side feet end sketch top small horse front piece black rope drawing left cut water wheel hand hair
Travelling	Over the sea	18	0.02731 port board sydney melbourne ship boat left wharf leave ashore arrived bay harbour p.m troops fremantle town sea cape colombo

Personal	sadness	71	0.02712	dear great son death sympathy god boys loss home heart feel soldier died mother sad life proud kind mrs men
Personal	Money	91	0.0271	pay money paid bank book dear office comforts leave london soldiers sydney fund mrs received send made funds australia case
		24	0.02664	morning oclock back today afternoon tonight night camp time good regt put horses duty received dinner troop arrived horse till
		23	0.02545	indecipherable today home crossed horses usual bed january monday cairo friday tuesday thursday april mail february letters wednesday tomorrow day
Personal	Family & memories	65	0.025	p.m sister sisters miss nice mrs duty i'm love home a.m time happy hope lovely day tea poor matron afternoon
Tourist	Drinking	41	0.02435	beer canteen bought money bottle wine drink french sold photos buy cost price francs bottles pay paid films glass coffee
Travelling	UK & France	28	0.02217	camp salisbury leave france weymouth left england london hut back draft arrived hill boat train miles training sutton plymouth days
Military life?	Egipt & Turquey	47	0.02133	hospital alexandria cairo left lemnos egypt arrived troops island camp april wounded heliopolis to-day aust gallipoli ship back anzac base
Personal	Religion	82	0.02129	church service sunday parade morning held chaplain padre sermon attended afternoon communion services y.m.c.a evening address rev holy present gave
Military life	General?	35	0.02073	sydney australia battalion served pte mrs embarked nsw private australian lieutenant gallipoli field hmat returned service enlisted france october father
		72	0.01979	generally make feel things long interesting find makes pass talk return thing takes mind times case word past trouble matters
Tourist	Egipt	11	0.01974	cairo nile pyramids egypt egyptian heliopolis desert mosque camp gardens mena native sphinx tram natives pyramid citadel hotel city donkeys
War	Context	50	0.01952	war german germany australia government military british terms minister sir peace conscription vote general president country britain forces governor states
		19	0.01947	canal camp cairo suez left col fuller desert arrived major kantara brigade march men train february officers sand port returned
Military life?	Camp & horses?	67	0.01819	html fine morning afternoon day cleaning harness round evening horses stables friday back paddock saturday night thursday warm tuesday december
Personal	Letters & family	49	0.01811	home mum dad wrote letter ellis write letters mrs george recd meet day indecipherable lovely play auntie walk bed night
Military life?	Camp & horses?	96	0.01789	horses bty battery horse lines wagon night line p.m fine a.m day men sgt today guns gun weather morning evening
Military life	Training & weapons	12	0.01778	school gun training rifle drill work practice range today bayonet instruction coy musketry shooting afternoon lecture class lewis machine parade
Tourist	France	22	0.01727	paris hotel french place rue dinner cafe visit opera indecipherable walked find club people english round city met tram girls
Travelling	Railway & roads	7	0.01653	train engine line railway large trucks arrived left back run station boys depot number engines made yard started road trains
War	Cementery	37	0.01641	graves dead war grave buried crosses battle cemetery land soldiers html cross death fought lie god broken peace bones soldier
		43	0.01488	house room back engine home html park load farm made french broken fire door fritz apl supplies left wood oil
		17	0.0146	envelope mrs active justice service reverse post south field walesaustralia censor card addressed australia n.s.w signed postcard image fergusonsupreme shows
		52	0.0137	good letter p.m letters girls day hope night morning nash hospital dear kitty a.m sydney australia joseph mrs colonel egypt
		83	0.01359	sydney german island rabaul ship emden men wireless board officers naval native germans islands captain guinea left suva sept fleet
War	Combat	92	0.01257	gas helmets helmet trenches alarm billets june bombardment masks shells attack tear heavy steel armentieres france issued trench box mask
		2	0.01193	aug wed sat mon tues sun fri jan letter nov mar sept dec june feb mother thur july thurs oct
		25	0.01138	turned p.m a.m breakfast till tea dinner fed rested fell camels cleaned camp returned guard day stand saddled hot watered
Military life	Communications	16	0.01123	stop division anzac landing turks telegraph position troops bay army ashmead attack positions suvla turkish august baba gallipoli report enemy's
War	Prisoners	94	0.01057	german prisoners camp today received food english germans officers dated germany parcels room cross red escape prisoner french tonight bridge
		84	0.01005	signal time date naval flag wireless message berrima ship brigadier station received herbertshohe despatch messages receiving beresford telephone transmitting troops
		89	0.00886	november mrs december rup met friday thursday october tuesday hotel london monday wednesday saturday lunch afternoon evening home miss club
Military life	navy	81	0.0088	sea ship ships fleet arrived anchored proceeded german p.m aug squadron admiral harbour convoy a.m anchor weather cruiser left firing
Personal	Poetry?? CHECK	85	0.00712	book tonight today books read mail night yesterday reading back morning full poems boche early day half indecipherable long frank
		64	0.00692	oct nov gen office left indecipherable afternoon col tidworth troops morning visited wednesday called sunday thursday tuesday pencil monday saturday
		70	0.00653	wher camp internees wich internee hawe pris military day soldiers comandant compound owing issued made australia guard police charge recieved
		87	0.00645	day night quiet fine letter rec sun wed tues mon posted html sat indecipherable writing letters fri amy heavy arr
		79	0.00634	south wales transcribed library state spelt misspelt transcriber's possibly judy gimbert john peter smith ditto mayo betty lynne adrian bicknell
War	Air force	33	0.00536	lieut enemy squadron bde machines air report pilot machine aerodrome flying capt observer reconnaissance reported a.f.c area pilots m.c wadi
		1	0.00508	hun food huns men parcels russian russians misery awful received whilst english prisoners control reliable promptly british caused bread french
Personal	Food	77	0.00439	men rations baked n.c.o's unit hospital n.c.os man temp leave bakery n.c.o.'s strength rejoined o/r flour ovens bakeries dough personnel
CHECK	songs?	63	0.00438	yer we're song puff voice we'll sing there's love don't tommy we've i'll it's you're blokes html snake you'll that's
Personal	In French	51	0.00402	les des pour bon vous nous c'est tres pas qui guerre est monsieur agrave par une dans france tout sont
		31	0.00312	miss mrs indecipherable hill rita coy r.n r.a.n c/o king grafton j.t batt john health nurse tindale baby trained testimonials
War	Context	68	0.00107	page kms miles france east south west north belgium called village resting harvey flanders german centre battle coast tel australian

.