Visual Exploration of Historical Semantic Data

by

Daniel Keenaghan B.A. B.A.I.

Dissertation

Presented to the

University of Dublin, Trinity College

in fulfillment

of the requirements

for the Degree of

Master of Science in Computer Science

University of Dublin, Trinity College

October 2012

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Daniel Keenaghan

October 1st 2012

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Daniel Keenaghan

October 1st 2012

Acknowledgements

I would like to thank my supervisor Owen Conlan for his advice and support throughout this project. I would also like to thank Mark Sweetnam for his assistance in designing the evaluation, and Eamon Darcy for his help during the evaluations, and I'm also very grateful to the members of the humanities department who took part in the evaluation. Finally I would like to thank Emma Lynch for her support, and her help in proofreading this dissertation.

Daniel Keenaghan

University of Dublin, Trinity College
October 2012

Visual Exploration of Historical Semantic Data

Daniel Keenaghan
University of Dublin, Trinity College, 2012

Supervisor: Dr. Owen Conlan

By presenting historical data visually, patterns and relationships can be more easily seen. There are challenges when it comes to designing an effective means to allow a user to visually explore historical data, which is often quite noisy, containing inconsistent, missing or malformed data.

This report will seek to produce a set of guidelines that should be followed to support users in the visual exploration of noisy historical data.

Through the creation of a basic visual exploration tool and by performing user evaluations on that tool a series of recommendations were arrived at. These recommendations include creating multiple different view of the data that the user has selected through the use of filters, and, the observation that while it is important to ensure that the user has access to a complete set of results for a give filter set, they do not all have to be plotted visually, the user will accept a certain level of missing or incomplete data.

Contents

List of Figures

Chapter 1

Introduction

Modern society either via social networks, research projects or online data driven collaborations (Such as Wikipedia or OpenStreetMap) produces a vast quantity of semantically rich data. This data is often a mixture of semantically annotated data and noisy raw data. The amount of data produced by many research or other projects is too much for a person to make any meaningful discoveries in a reasonable amount of time by browsing through the raw data alone. In order to find answers to research questions posed, a researcher must be able to find patterns within the data or be able to quickly navigate to a specific piece of data they are interested in. They must also be able to view related or similar data that will help them discover links within the data. A visual exploration tool also aids in the formulation of new research topics or questions by facilitating undirected exploration of the data.

The 1641 Depositions project is one such research project that produced a large semantically marked up data set. The 1641 Depositions are a series of oral accounts in which people who lived through the 1641 rebellion era discussed their experiences of that time. This project produced a database

of semantically rich data to be used by researchers who are interested in studying the 1641 Depositions. Due to the nature of the source of the data that was produced there are many difficulties that must be overcome in order to enable a meaningful exploration of the data set.

To produce a tool that is useful to a domain expert one must first understand how that person interacts with the data and what level of access to the raw data is required. A user may have extensive knowledge of the data that is being visualized, however they may not be proficient at using computers. Therefore they require a tool that is intuitive and simple to use while remaining a powerful tool for exploring the data at hand. This project will look at ways to enable such a user to explore the data and extract useful information from it.

A challenge faced in this project arises from the use of historical data. This is data that has been transcribed from historical manuscripts. The data is often much more messy than modern text, due to inconsistencies in spelling and punctuation. Often the same word is spelled in two different ways within the same sentence. Difficulties also arise when trying to map the various place names that occur within the text. Aside from the inconsistencies in spelling, many of the place names have changed since the data was recorded or they no longer exist, making them difficult or impossible to properly map. A similar issue arises for the names of people mentioned within the texts, it can be difficult to determine if two different spellings of the same name are two different people or just a lack of consistency in the spelling of their name.

The CULTURA project aims to build and exploit relationships in this data to aid in humanities research and thus generates a large volume of data with corresponding semantic models. Developing a tool that will utilize these models and data will be the primary focus of this project. The project

will develop a visual exploration tool to help researchers make sense of a large data set and explore underlying relationships within that data set. Part of this project will involve identifying the most suitable and intuitive way to present the content to a given user. There are a wide variety of visualizations available for this task, each of which has a number of strengths and weaknesses when it comes to displaying a particular type of data.

1.1 Motivation

In large data sets there are usually many complex hidden relationships that lie within the data. These relationships often remain hidden due to the complexity of the data in its raw form but can be more easily seen if presented visually in various different ways.

By enabling researchers to explore the data in a tailored and visual way, previously unseen relationships can be exposed and related content discovered. Adapting the way the content is presented based on usage of the tools will enable the user to work more efficiently with the underlying data.

The techniques and tools developed for this project could then be adapted for different research areas or more generally for use with any semantically marked up data. One can imagine exploring connections between relationships between users on Twitter and Facebook and geographic data to discover trends or deliver more targeted online advertisements.

A secondary motivation for this project was to produce a tool that would allow historical researchers who are working with the 1641 Depositions to explore the depositions more effectively and efficiently. Currently the researchers must search through the depositions by hand in order to find interesting collections of depositions and the information contained within them,

often spending many weeks on a task that could be accomplished within hours through use of the correct tools. This tool could then later be expanded to allow for manipulation of the 1641 Depositions database, allowing users to add more meta data, which would in turn allow researchers to better explore the depositions.

1.2 Research Question

This project seeks to answer the question: *What guidelines should be followed to support users in the visual exploration of noisy historical data?* In this project noisy historical data refers to data that is historical in nature and has been transcribed into some sort of electronic database. This data contains numerous omissions and inconsistencies, such as different spellings used for people or place names. It also contains information that is out of date, such as place names that have changed since the original data was produced. An exploration of data is taken to mean the iterative process used to refine filters and views of data in order to answer a specific question, or more generally browse a data set.

The approach taken in the pursuit of answers to this question involved first creating a basic visualisation tool, liaising with members of the humanities department for feedback and then conducting user evaluations in order to determine the effectiveness of the resultant tool. The basic visualisation tool focused on utilizing the 1641 Depositions database. Conclusions were then drawn from the experience acquired during the project and the results of the evaluation conducted on the tool produced. In order to create an effective tool for evaluation, an iterative process was followed, whereby, following each round of improvements made to the tool, feedback was sought from a

member of the humanities department who had experience with the source material.

To support the investigation into this research question a number of objectives needed to be completed. These objectives are listed here in the order in which they were to be performed.

1. Perform State of the Art review.
2. Design and implement visual exploration system.
3. Design and conduct evaluation with authentic end users.
4. Draw conclusions from evaluation data.

1.3 Dissertation Overview

Following this chapter this dissertation will give a review of the current state of the art in data visualisation and user interaction techniques.

In Chapter 3 the major design decisions that were made during the course of this project will be explained along with the reasoning behind those decisions. This will be followed in Chapter 4 by a detailed summary of the implementation and development of the visualisation tool that was developed as part of this project.

A detailed analysis of the project will then be presented in Chapter 5, considering the feedback that was gained through performing evaluations of the tool with domain experts.

Finally in Chapter 6 conclusions and observations that were made during the evaluations and during the project as a whole will then be given. In this chapter guidelines will be offered to answer the research question that this dissertation asks. This will be followed by an outline of the future work that could be carried out.

Chapter 2

State of the Art

There are a vast amount of research papers and tools available in the field of data visualisation. In this dissertation they are broken up into two main sections. Section 2.1 on visualisations will provide an overview of the different types of visualisations that are used and comment on their strengths and weaknesses. Section 2.2 looks at more general papers that deal with the methods used to allow user to visually explore a data set. Finally, Section 2.3 will look at the various tools that are used to produce these visualisations and also touch on the data manipulation that is required to rearrange data into a format from which a visualisation can be made.

2.1 Visualisations

An overarching aim of this project was to take raw data and visually present it to a user in a useful manner. Therefore various different types of visualisation were examined to determine their value for use within the visualisation tool that was produced. This section will detail the different types of visualisations used and comment on their appropriate usage. Some of these along

with other visualisations are examined by Heer, Bostock and Ogievetsky[16]. The visualisations that are commonly used range from the simple Bar Chart and Scatter Graph to more the exotic Cartograms and Parallel Co-ordinate plots, each of which is suited to different types of data.

Figure 2.1: Circle Graph [4]

2.1.1 Circle Graph

The Circle graph was described by Aumann et al. in 1999[15]. This type of graph is used where there are many relationships in the underlying dataset. A viewer can quickly gain an overview of all of the relationships present. Due to the circular arrangement of the nodes in the graph the resulting visualisation is much cleaner and more compact than many other layouts, a linear layout (such as Arc Diagrams. See Section 2.1.8) can become very large in size for datasets that contain a large number of nodes.

In an interactive environment enhancements can be made to the static image. By hovering over a specific node all connections to that particular node can be highlighted. An sample of the circle graph can be seen in Figure 2.1.

2.1.2 Tree Diagrams

The tree diagram can take many forms, they are commonly employed to represent hierarchical data such as a file structure. These diagrams are typically used in file management software and consist of a list of nodes, each of which can be expanded to reveal its sub-nodes, which are typically placed slightly indented from the parent. A circular variant is shown in Figure 2.2, this circular variant is more efficient in its use of space, however it is not as easy to read through the list of nodes as a vertical tree. The vertical tree would however require much more vertical space.

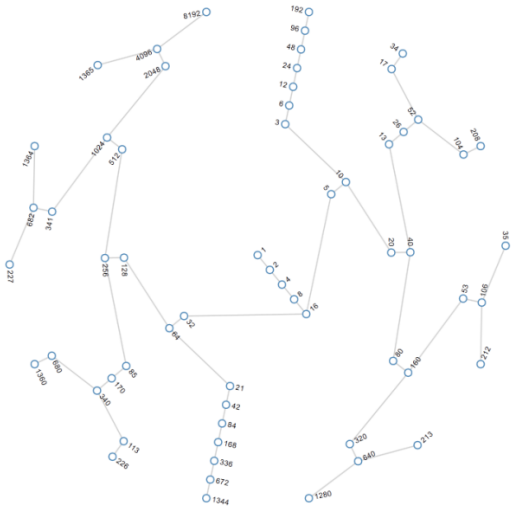

Figure 2.2: Collatz Tree Diagram[5]

2.1.3 Treemap

The treemap, as described by Schneiderman[22], is useful for representing hierarchical data. Each rectangle in the graph represents a node in the data set with its area proportional to the magnitude of the value that is being visualised. Each sub-node within the graph is contained by its parent node.

Shneiderman provides an algorithm for generating treemaps in his paper. By using the area to represent the value of a node, nodes can be quickly compared and contrasted aiding in discovery of relevant information. A common use of the tree graph is in hard disk usage analysis software. Such software produces a tree graph where each node represents the amount of space occupied by files on the hard disk.

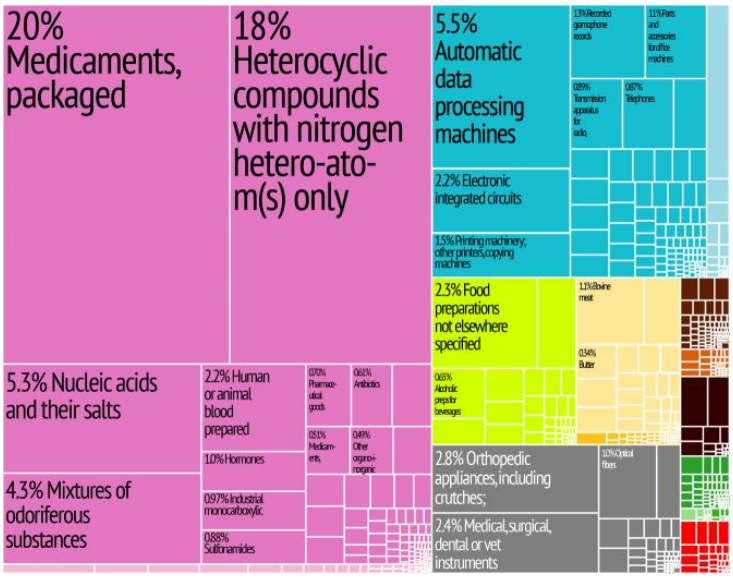

Figure 2.3: Treemap representing Irish exports; in various categories, denoted by colour. [10]

2.1.4 Adjacency Diagram

Adjacency Diagrams are a similar visualisation to the treemap in that they are used to represent hierarchical data using the area of a node to represent its value. They differ in that rather than sub nodes being wholly contained within their parent node they are placed adjacent to them. This allows for direct comparison to other nodes of the same depth irrespective of their parent nodes. Again they are used by hard disk usage analysis software, as

a more visually aesthetic visualisation.

Both treemaps and Adjacency Diagrams are suited to displaying hierarchical data where a specific data value of a node (represented by the area of a node) is just as important as the hierarchy itself.

2.1.5 Maps

There are a wide variety of different map visualisations, ranging from the basic geographic map to the more exotic such as the cartogram.

Choropleth Maps

This type of map displays a particular region where it's subregions are shaded in different colours to represent a variable. They are typically used to display data about particular geographic regions, such as population or election results. The tool produced as part of this project is expected to make use of a basic map, however it could easily be extended to show data using a Choropleth map.

Cartogram

Like a choropleth map these maps are used to represent information about particular geographic areas. Rather than simply using different colours to represent varying values the area is distorted in proportion with the variable that is being measured. Keim, Panse and North[20] look at different methods used to generate cartograms efficiently and suggest one method in particular that should be used.

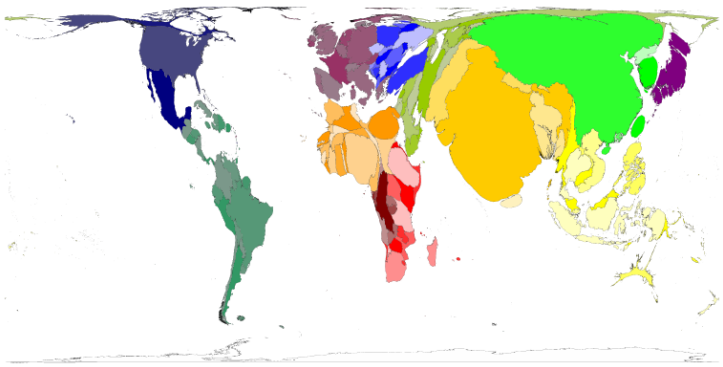

Figure 2.4: Cartogram showing countries with their area proportional to their population. [2]

2.1.6 Parallel Coordinates

This technique as outlined by Inselberg and Dimsdale[18] is used to display multi dimensional data. Each vertical line represents a dimension that is being measured, and the lines connecting points on these axes connect corresponding points. This particular visualisation can become rather noisy and hard to read if there is a large amount of data being represented, however this can be mitigated somewhat in an interactive environment by allowing the viewer to highlight only specific data points and their connections. An example of a parallel coordinates plot can be seen in Figure 2.5.

2.1.7 Voronoi Diagram

A Voronoi diagram is built using a collection of points, or seeds. The diagram is divided up into regions which are defined based upon the distances to these seeds. The region occupied by a seed will contain the area which is closest to that seed. This can be used to visualize the closest instance of a particular object to a specific region. For example a Voronoi diagram could be used to show areas which are closest to a particular airport.

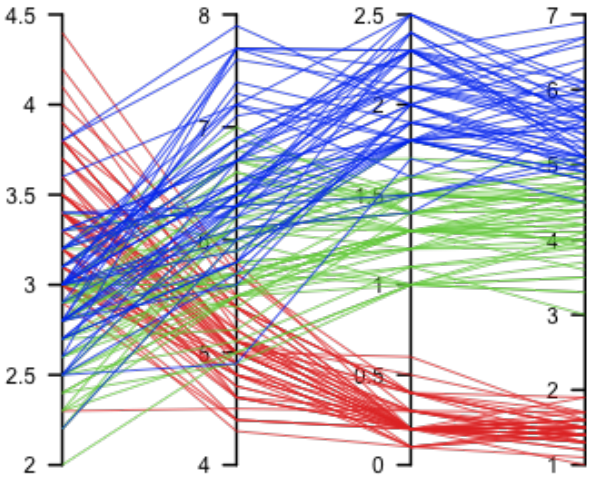

Figure 2.5: Parallel Coordinates example. [11]

2.1.8 Arc Diagram

Arc Diagrams[24] are generally used for displaying sequences in which there is some amount of repetition, one such example is visualizing co-occurrences of characters in a novel or play. Figure 2.6 shows links between books in the bible that reference each other. This example is particularly large and a very high resolution image is required to see all the detail. The example shows how visually appealing this particular visualisation can be. The Arc Diagram is also suited to smaller data sets.

2.2 Visualisation General

This section will look at a range of more general visualisation papers, and discuss how they are relevant to this project.

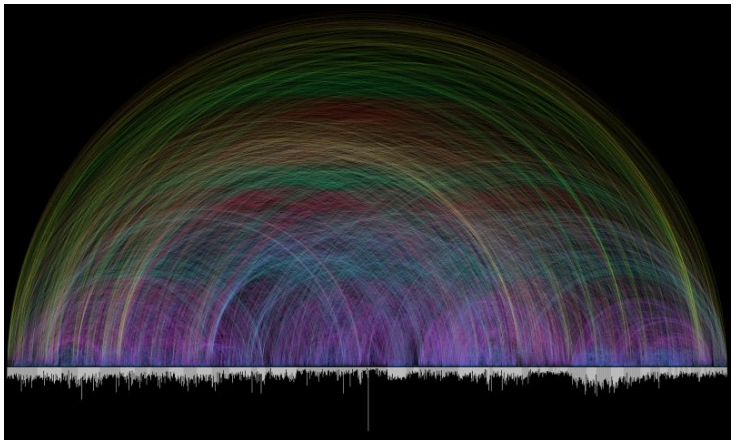

Figure 2.6: Arc Diagram [1]

2.2.1 Visualising Text

The 1641 depositions database contains a transcription of the original depositions. Visualising textual data is problematic as it is inherently non-visual data. Wise et al [25] discuss this problem. They present a way of visually exploring textual data by 'transforming it to a spatial representation that preserves informational characteristics from the documents' to be visualised.

2.2.2 The Eyes Have It

Ben Schneiderman discusses the concept of *Overview first, zoom and filter, then details-on-demand* in his paper 'The Eyes Have It: A Task by Data Type Taxonomy for Information Visualisations'[23]. This paper provides a list of seven different tasks that should be implemented in an application to support visual exploration and also suggests ways to visually display complex filter queries. This paper informs many of the design decisions made throughout the course of this project.

2.2.3 Evaluation Challenges

The paper 'The Challenge of Information Visualization Evaluation'[21] by Catherine Plaisant looks at the challenges facing researchers who are performing evaluations of information visualisation tools. The paper highlights ways in which the methods used to evaluate visualisation applications can be improved.

2.3 Visualisation Tools

In order to display data sets visually, they may require a certain amount of preprocessing. Raw data from historical datasets may need to be cleaned up or normalized, or the visualisations themselves may need to be processed further to deliver a better user experience. These visualisation tool will be discussed in this section.

2.3.1 Visualisation Libraries

Given that this project is intended to deliver a plug-in for an interactive online web based environment, a client side Javascript library is most suited to generate the visualisations required. This allows for a more responsive application compared to one where the visualisations were generated on the server and then transfered to the client. A large range of visualisation libraries are available, most of which have been recently developed and are being actively updated thanks to the growing popularity of the web as a platform for delivering interactive media and applications.

The Protovis[12] Javascript library was used by Heer, Bostock and Ogievetsky [16] to create the visualisations presented in their *"Tour through the Visualisation Zoo"*. This library has been superseded by the D3.js [6] library,

and has quickly become popular throughout the web for producing dynamic data driven graphics. The library generates standard HTML mark-up & SVG[13] graphics that can be then styled using CSS. This standards compliant approach allows the visualisation to be viewed by anyone using a browser that supports SVG[1].

Google Maps

To visually display the geospatial data a map was required. There are a number of options available, however Google Maps were chosen. It provides the best combination of features with active development support communities should a problem arise during its integration to the visual exploration tool. The Google Maps API provides a geocoding service that shall be utilized to convert the place names that exist in the 1641 Depositions database to co-ordinates. Also provided are a series of overlays, such as markers and info boxes, that can be used in the maps to display information. These overlays are all highly customizable and can be tailored to best represent the information at hand.

2.3.2 Polygon Amalgamation

When data is shown on a map specific regions of the map can be highlighted by polygonal shapes layered on top of the map. Where there are multiple regions adjacent to each other it can look visually more appealing to remove internal boundaries and show multiple polygons as if it were one larger polygon. Methods for combining several polygons into a single polygon are discussed in *"Efficient Polygon Amalgamation Methods for Spatial OLAP*

[1]Approximately 85%[14] of web users' browsers natively support SVG. Major browsers that do not support SVG are Internet Explorer 7 & 8, however a plug-in is available for them that provides SVG support[3], along with other improvements.

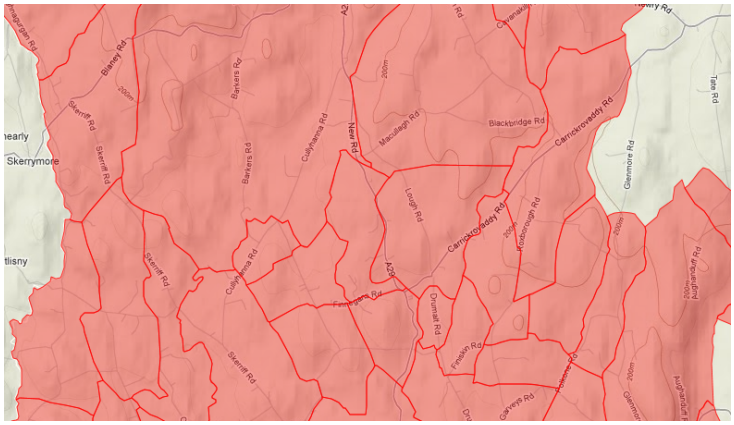

Figure 2.7: Polygonal regions overlaid on a map.

and Spatial Data Mining" [26]. The paper compares the processing speeds of the various algorithms and concluded that the occupancy-based method was faster. Figure 2.7 shows a map with several different regions highlighted. If the polygon amalgamation algorithm were used only the external boundary lines would be shown.

This polygon amalgamation technique was not used in the tool as implemented by this project. This is because markers were used to represent places on the *Map* rather than highlighting an entire region. This method would be useful in the future when improvements could be made to the tool to show areas on the *Map* rather than just points.

Chapter 3

Design

This chapter will present the major decisions that affected the design of the visual exploration tool. It will also look at some of the challenges faced throughout the project. The chapter will begin with the reasoning behind the choice of platform, followed by a closer look at many of the other design decisions made over the course of the project.

3.1 Platform Selection

One of the earliest decisions that needed to be made in the project was the choice of platform. This choice would then later inform system architecture choices. It was decided to use version 6 of the Drupal Content Management System[7] as a platform on which to build the visual exploration tool. The tool was to be implemented as a Drupal 6 module. The choice of the Drupal CMS was largely a result of ensuring integration with existing tools developed as part of the CULTURA project. The CULTURA project is based around Drupal, and thus by implementing the visual exploration tool as a module, it could later be used as part of the CULTURA system.

The choice of Drupal and therefore a client side application developed using HTML, CSS and Javascript technologies also offered a solid development base. There are a wide range of development tools and debugging software available, which greatly aid in the development of web applications. The use of these web technologies allowed for rapid development of suggested features, and easy integration with web services such as Google Maps. It also allowed for a flexible user interface implementation that could be changed around easily if a more optimal UI configuration was required.

3.2 Architecture Decisions

Due to the choice of implementing the visual exploration tool as a Drupal module the basic application architecture would be Client / Server. In order to keep the interaction between the client and the server to a minimum it was decided to implement the majority of the functionality of the tool on the client side. This decision was also influenced by the development ecosystem available. There are a wide range of tools available for building and debugging Javascript applications. The server side code would be kept to a minimum, only serving as an intermediary between the client and the database. It would also be used to save log files. The lack of proper debugging support for PHP played a role in this decision.

3.2.1 Filters

The choice of filters was important and needed to reflect how the users of the tool would wish to query the database. Six different filters were settled on. These filters were picked based on discussions with a humanities researcher. The six filters which were chosen are as follows:

County Often the user would be interested in depositions only from a certain county or collection of counties.

Nature of Deposition This allowed the user to filter the depositions by the type of content they contained, be it Killing, Robbery etc.

Occupation of Deponent This allowed the user to filter the deposition by the occupation of the person deposed.

Deposition Type The depositions fell into 5 main different types, with the rest being grouped under the title 'Other'. Often researchers would only be interested in depositions of a certain type.

Commissioner While not as used as the other filter types, being able to filter out depositions by Commissioner would be a useful feature to some in their research.

Date of Deposition By allowing the user to filter by date, they could see the progress made by the deposition commissioners as they traveled the country gathering their data.

The decision was made to create a general framework for the implementation of filters. The filters would be stored in a separate configuration file which specified a number of different parameters that would make up the filter. This allowed for the flexibility required to ensure that any type of data in the database could be used as a filter. Filters could then be easily added without having to edit the source code of the main application. This framework would also incorporate the normalization of certain pieces of data, such as commissioner names.

The UI to select the filters was placed at the top of the page as they are a crucial element of the visual exploration tool. The filters constitute

the 'zoom and filter' portion of Schneiderman's mantra[23]. Filters can be quickly added using the drop down menus, which include the ability to search for specific values.

3.3 Visual Components

The design of the interface aimed to follow the guidelines concerning visualisations as laid out by Schneiderman in the paper 'The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations'[23]. In this paper Schneiderman describes his 'Visual Information-Seeking Mantra' of '*Overview first, zoom and filter, then details-on-demand*'. The interface supported this approach, as all of the depositions can be shown either on the map, or represented in the statistics view so the user can get an overview of the data. Then the user can filter out depositions they are interested in, and select one they are interested in to view it's details.

There shall be three main sections in which the data is to be visualised. These are the *Map*, *Statistics View* and *Deposition View* each looking at the data from different perspectives and levels of details. Each of these sections can be viewed simultaneously to allow for comparisons between them. Each section allowed the used to explore and interact with the data in a different way, increasing the potential for discovery of interesting patterns or relationships.

The *Map* would provide a broad overview[23] of the geospatial data available within the database. This would allow users to see where there were concentrations of depositions that met a specific criteria, or when used in combination with the date filter, it could be used to investigate the methods used in collecting the depositions by tracking the paths of the commissioners

as they traveled around the country collecting depositions.

The *Statistics View* would provide a range of information about the depositions. It can be used to show co-occurrences of properties of depositions. It can also be used to rapidly get information such as the number of depositions that exist that meet specific criteria.

The *Deposition View* would give quick access to the underlying deposition transcription, from which all of the meta data has been previously extracted. This is important to researchers as they need access to the source material, this could be further improved by also providing links to images of scans of the actual documents from which the information was transcribed. This view would provide all of the details related to a deposition when demanded by the user, this is in line with the recommendations by Schneiderman[23].

To avoid the risk of confusing users, all data would be synchronized between the various different views. The exception to this being the *Deposition View*, which would retain the last loaded deposition until a different one was selected. This would allow the user to view the data from several different viewpoints at once, aiding them in determining patterns and relationships that exist within the data.

Numerous different types of charts were considered for use within the *Statistics View*. The different charts considered can be found in Section 2.1. However, in the end it was decided to only use a single chart type, the bar chart. This was sufficient to represent the data and was familiar to users. The implementation of the *Statistics View* was designed so that different chart types could be added in the future.

3.3.1 Sidebar

In order to allow a user to easily move between the different views a simple mechanism was implemented whereby they could select the view they were interested in from the side bar and the page would scroll to that view. This sidebar remained stationary relative to the screen, scrolling with the user as they scrolled up and down the page, meaning that it was always accessible.

3.4 Difficulties

3.4.1 Geocoding

Geocoding is the process of converting a place name. For example turning the place name*Dublin, Ireland* into the co-ordinates $53°20'52''N\ 6°15'35''W$. This is necessary in order to plot the positions of the depositions on the a map, such as the one used in this project. The difficulty arose where the place names that were in the 1641 depositions database were not valid. Often the name of the place now has a different spelling to the one given in the database, in other cases the place name had changed completely. This presented a challenge in that not all of the deposition would be able to be displayed on the map.

3.5 Alternatives Considered

Initially a different approach was taken to visualise the data in a more general way. Google created DSPL or Dataset Publishing Language[8] for use in it's online Google Public Data tool. DSPL allows one to create a description of a dataset which can then be used to produce visualisations of data within

the dataset. The hope was that by using this technology a generalized visualisation tool could be created that could visualise any given data provided there was a DSPL file to describe the dataset.

In the end however, it was decided that DSPL was not suitable for use with the 1641 depositions. The language is more suited to describing numerical and statistical data such as populations, GDP, birth rates etc. The 1641 depositions don't contain much of this type of data, containing instead mainly textual data.

Chapter 4

Implementation

In the previous chapter the reasoning behind each of the major design decisions that were made over the course of this project were explained. This chapter will outline the architecture of the visualisation tool that was implemented as a result of those design decisions. A screen shot showing part of the application is shown in Figure 4.1.

The basic framework of the application will be described first, followed by a closer look at each of the separate components that make up the tool. These are the *Filters*, *Map*, *Statistics View*, *Deposition View* and *Data View* components. Finally, the *Logging* component that was used during the evaluation to log the participant's interaction with the tool will be discussed.

The application consists of a number of different panels or views. The first panel is the *Filters* panel which allows one to filter the depositions to be view by various parameters. The rest of the panels offer a number of different views of the data. A floating sidebar, as shown in Figure 4.2, allows the user to quickly jump between each of the panels, and any of the panels can be collapsed or hidden by clicking the title of that panel.

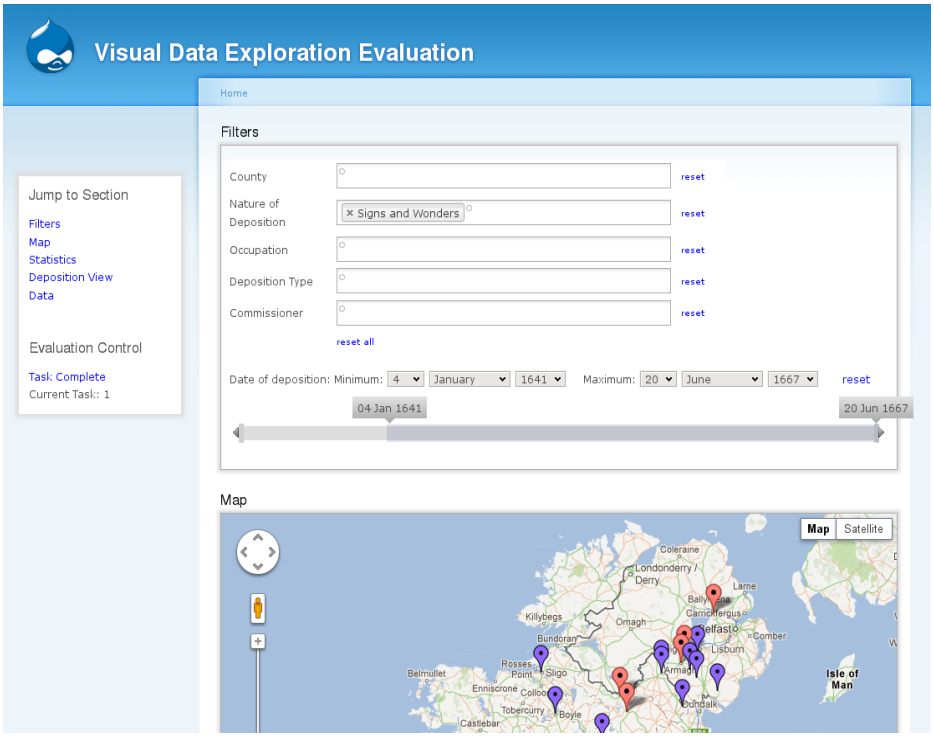

Figure 4.1: Overview of the visual exploration tool.

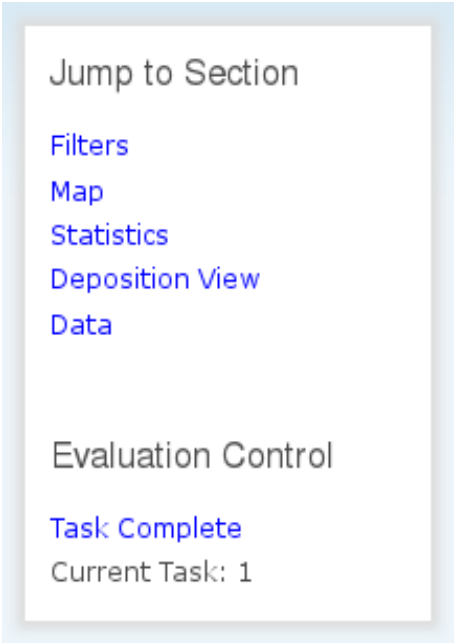

Figure 4.2: Hovering sidebar used to jump between sections of the visual exploration tool.

4.1 Basic Architecture

The tool is implemented as a module for the PHP based Drupal content management system. The reasoning behind the decision of creating the tool as a module for Drupal is that the CULTURA project is based around Drupal, therefore creating the tool as a Drupal module allows it to be integrated into the CULTURA environment relatively easily.

Due to the implementation of the tool as a web based application it has a Client \ Server architecture. The server is responsible for querying the MySQL database that contains the 1641 Depositions data as well as saving the log files generated as part of the evaluation. The server side code is minimal, consisting of a small number of PHP scripts tailored to return a different result (be it a list of filtered depositions, a specific transcription) from the database or update the database with new information. Everything else is handled by the client. The client uses Ajax requests at all times when requesting data from the server to avoid a page refresh. A server side PHP script serializes the response to JSON before sending it to the client.

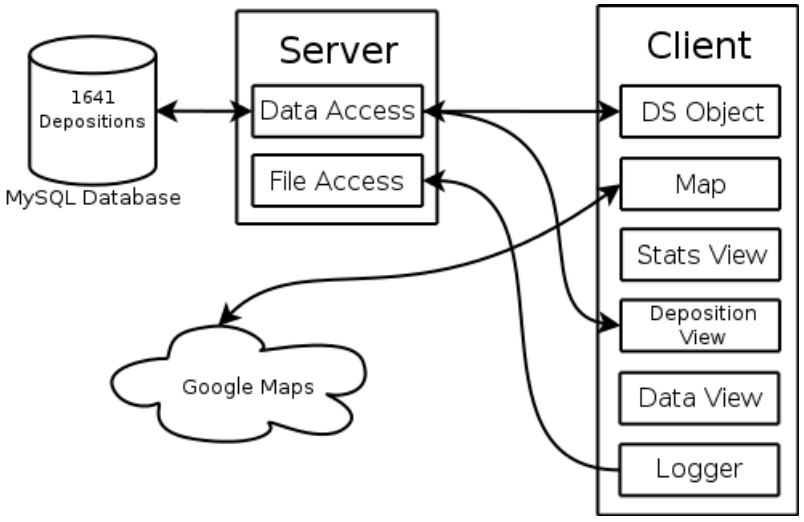

Figure 4.3: Basic architecture diagram of application.

On the client side, the application is broken up into a number of components, each corresponding to one of the different panels or views that made up the user interface. Each of the main components are represented as singleton *'Classes'*[1] which encapsulate all the functionality associated with that component. These main components will be discussed in more detail in further sections of this chapter. There is also a separate component that handles data retrieval, manipulation and filtering.

The tool makes use of several Javascript libraries such as jQuery, D3.js, moment.js and Underscore.js as well as some other minor libraries that provide extra visual controls such as the time line slider and filter selection boxes.

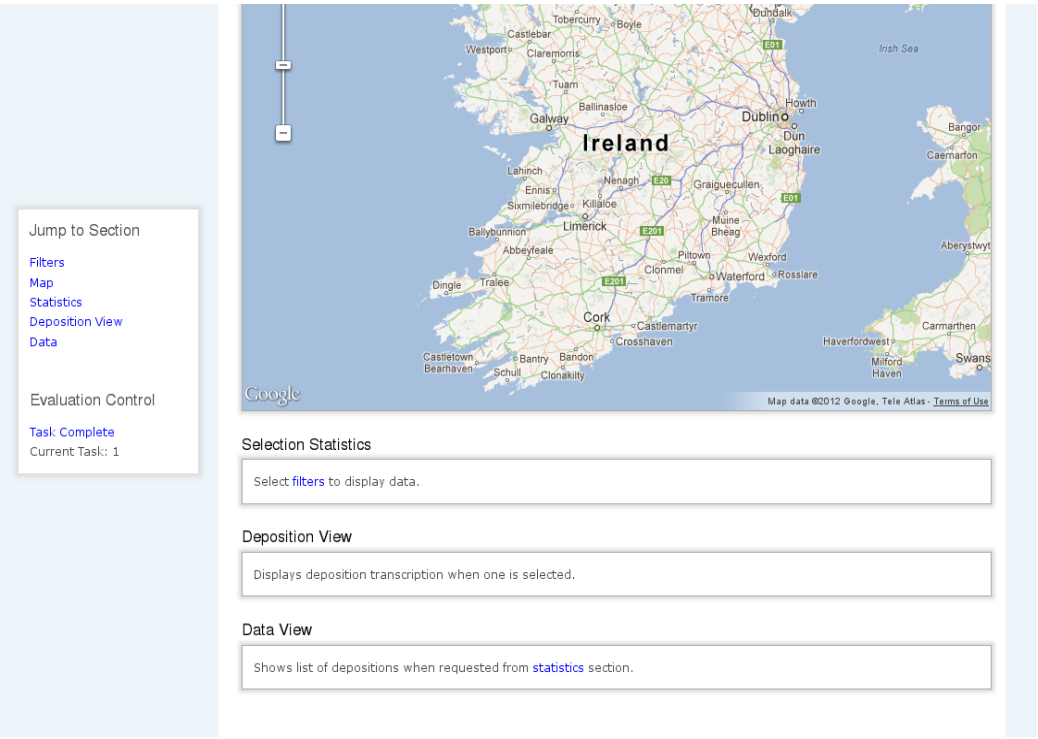

Figure 4.4: Overview of the bottom portion of the visual exploration tool.

[1]The Javascript language doesn't have actual classes, however they can be emulated.

4.2 Filters

Filters allow the user to view only the portion of the data set that they are interested in. For this implementation there are 6 different filter types that can be applied. These filter types are *County*, *Nature of Deposition*, *Occupation of Deponent*, *Deposition Type*, *Deposition Commissioner* and the *Date of Deposition*. When filters are added or removed the other views immediately reflect the new filters. If multiple filters within the same type are selected, i.e Wexford, Athlone and Meath from the county filter, then depositions from each of those counties are shown. This behaviour is a boolean *OR*. If there are multiple filters selected from different filter types then a boolean *AND* is applied. For example, if there are filters of "County: Wexford", "County: Wicklow" and "Commissioner: Adam Loftus" there the query will be "*(Wexford OR Wicklow) AND Adam Loftus*".

Most of the filtering is done at the server level. An SQL query is created from the various filters that have been selected, and the results returned to the client. This limits the amount of data that needs to be transmitted between the server and the client, thereby increasing response times. The exception to this is the time filter which is implemented on the client side. This was done because one can change the date filter by manipulating the slider much faster than adding any other type of filter, this would produce an unnecessary amount of data transfer back and forth between the client and server as the user changed the value of the date slider.

All of the filter configuration data required to show the filters in the application were stored in a separate file. The filter data was stored in a JSON format to allow for flexibility. This filter data contained information such as what columns from the database to filter. The filter system allowed for mappings. This means that a single entry in a list of possible filter values

could map to multiple values. This was used in the case of the *Deposition Type* filter, where, although there are many possible values for deposition type only the five most popular were shown. The rest were grouped under the value of 'Other'. This same system could also be used in filters for which normalization data was available. The canonical name would appear in the filter list, and when selected, be mapped to each of the alternate names, which was then used to search the database.

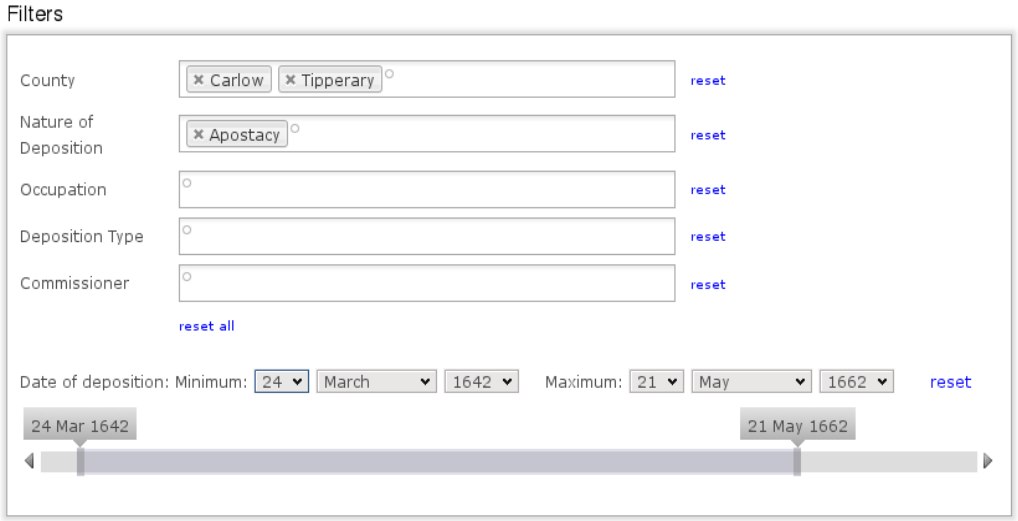

Figure 4.5: Filter section of the visual exploration tool.

An object called DS[2] stores and manages access to the information received from the server after filters have been changed. This object stores all of the depositions returned in an array, ensuring that no duplicate entries are added and grouping depositions that have the same place associated with them. The *DS* object also supplies statistical information about the depositions such as the number of depositions returned or the number of depositions that meet a certain condition. Also maintained by this object is a list of 'active depositions', these are depositions that have had a client side

[2]Abbreviation for Data Set

filter applied to them and have not been filtered out.

4.3 Map

The map plays a key role in the visual exploration tool as it allows one to visually view the locations of each of the depositions, providing an overview[23] of the data. When a set of filters are applied the map is immediately updated to reflect the new selection of data. The depositions are represented by markers placed onto the map. A deposition can then be viewed by clicking on one of these markers.

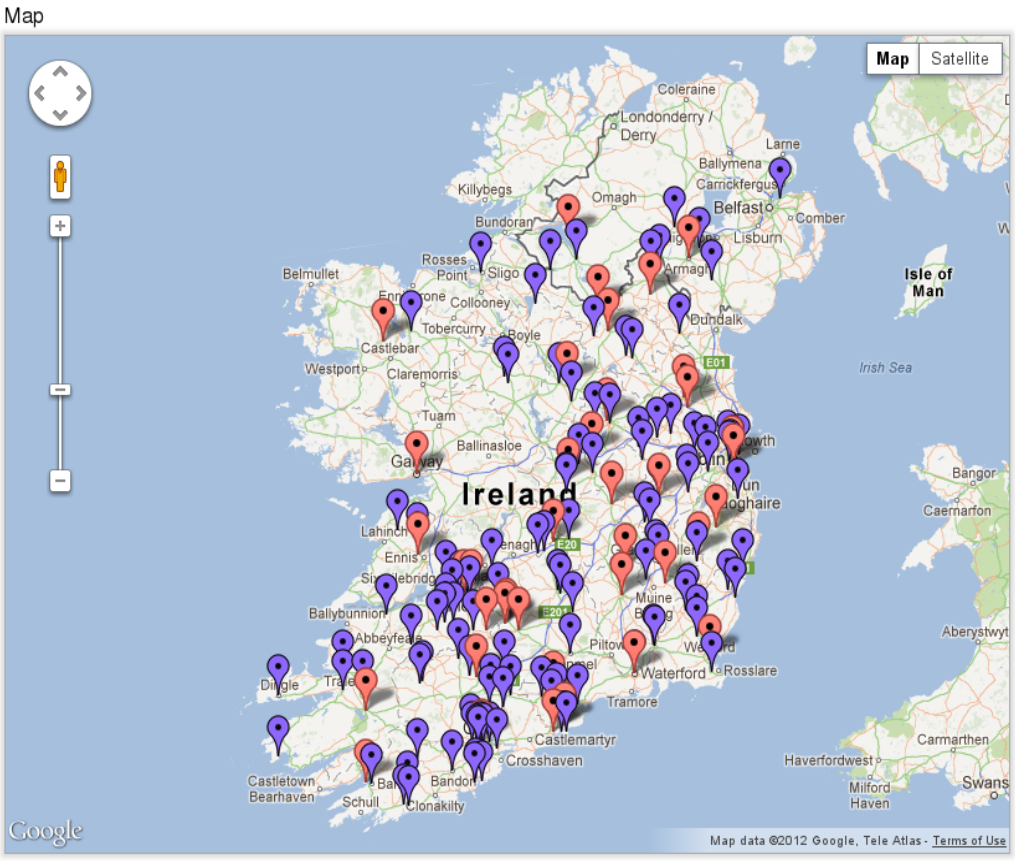

Figure 4.6: Map section of the visual exploration tool.

The map used is an embedded Google Maps map. A singleton *Map* object controls any functionality relevant to the map, such as initialization, adding markers, geocoding place names and the display of info boxes[3].

4.3.1 Geocoding

Geocoding is the process of converting a place name, such as *Dublin, Ireland* into co-ordinates, such as $53°20'52''N\ 6°15'35''W$. This is necessary in order to plot the positions of the depositions on the map, as the database that was used only contains the names of the locations and not their co-ordinates. To carry out this process the Google Maps geocoding service [9] was used. Extra columns were added to the existing 1641 Depositions database to hold the co-ordinates so that each location would only have to be geocoded once. A meta data column was also added to store information pertaining to the status of the geocoding process for a particular place, i.e whether or not the geocoding service was able to find a match for the place name or not.

The geocoding process was complicated by the historical nature of the data. Many of the place names in the database are now spelled a different way or have changed completely. In the former case the geocoding service was able to correctly map the locations to their proper locations, although some were still mapped incorrectly. In the latter case the place would either be mapped to the wrong location or there would be no results from the geocoding service at all.

Shown in Figure 4.7 are cases in which the geocoding process was particularly erroneous. The high rates of false positive matches for place names in the USA and Australia (not shown in the image) can be attributed to the

[3]The boxes that appear on the map when a marker is clicked, showing a list of depositions at that location. An example of an info box can be seen in Figure 4.8 on Page 33.

fact that many of the places names in these countries are similar to Irish places names.

Depositions for which locations were unable to be mapped are reported as such in the *Statistics View* (See Section 4.4, Page 34) and could be displayed as a list in the *Data View* (See Section 4.6, Page 38).

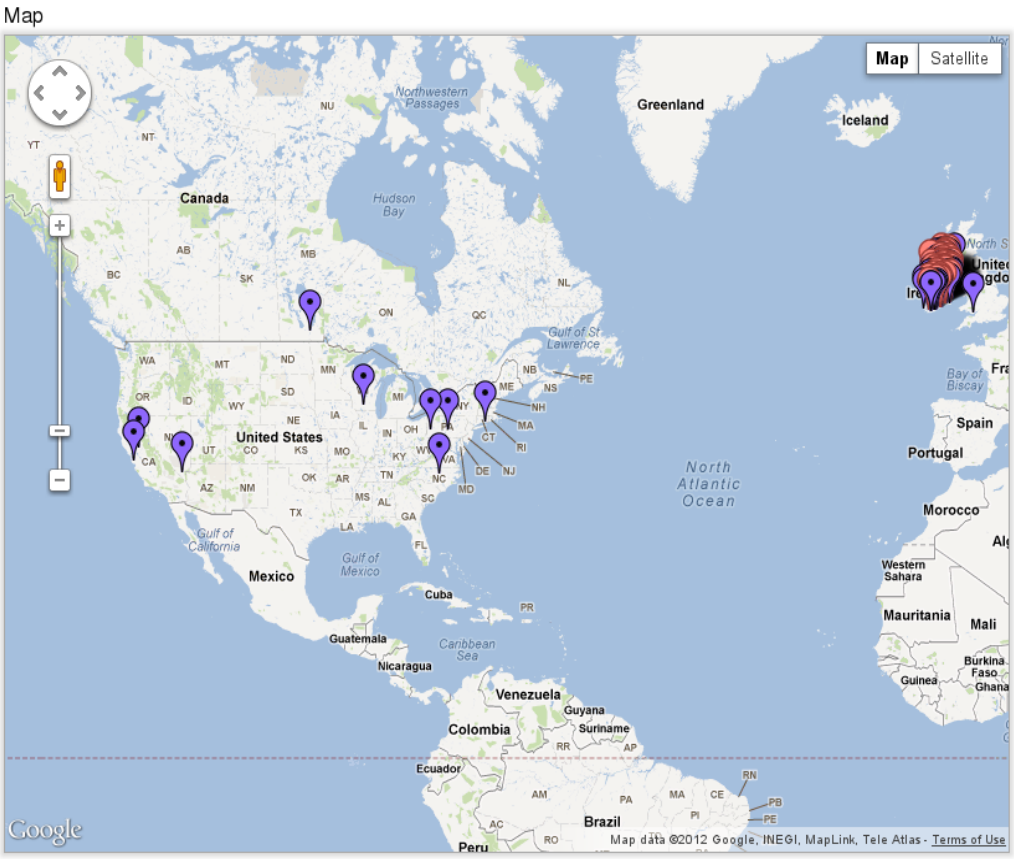

Figure 4.7: Map showing incorrectly geocoded locations.

4.3.2 Map Markers

The markers used on the *Map* were the default Google Maps teardrop marker. They can be seen in Figure 4.6. The blue markers indicate that there is a single deposition at that location and the red markers indicate that there are

multiple deposition at that location. To select a specific deposition from a group marked by a red marker one simply clicks the marker and is presented with a list of deposition at that location. This is shown in Figure 4.8.

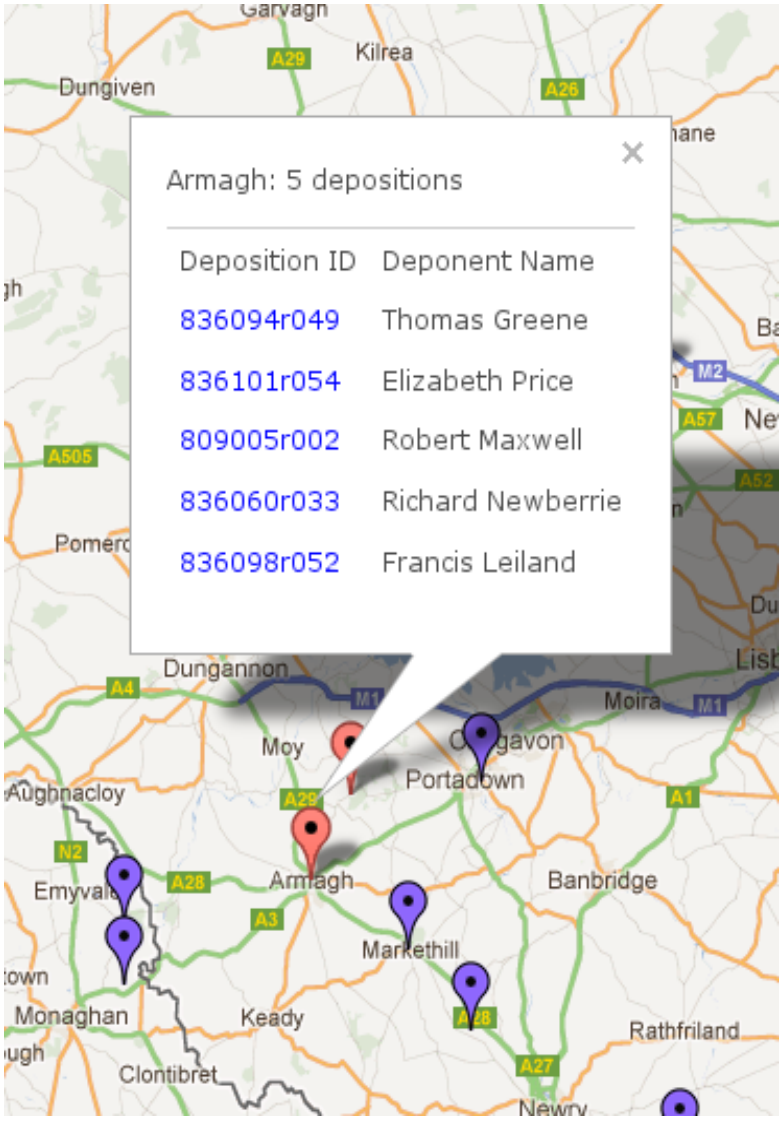

Figure 4.8: Information box showing multiple depositions in a single location.

Criticism of Marker Choice

The teardrop markers used in this tool turned out to be less than optimal, as they are very large which can lead to a very cluttered view on the map. This combined with the fact that the majority of the area of the marker is not positioned at the position it is marking would lead one to suggest that an alternative marker be used. A simple small circle positioned with it's centre point over the target point would be a better choice. This would allow one to gain a more accurate view of the locations of the depositions being visualised.

4.4 Statistics View

The *Statistics View* shows various statistics about the currently selected group of depositions. It will show statistics about the selection of depositions that were returned from the server. If a red marker (indicating multiple depositions in a single location) is selected on the *Map*, the *Statistics View* will show statistics about the depositions that are represented by that marker. Also shown are the total number of depositions that match the set filters and the number of depositions within that set which are not shown on the *Map*. There is a link beside each of these totals that when clicked will display the corresponding list of deposition in the *Data View* (See Section 4.6).

To draw the charts in the *Statistics View* the D3.js library is used. This library makes use of SVG to draw interactive charts, allowing for rich charts to be generated from data in the client. The only chart type currently implemented in the tool is a bar chart, however it is relatively trivial to modify the code to implement other types of charts to display data.

Currently there are 4 different sets of data that can be displayed on the chart. These are *Nature of Deposition*, *Deposition Type*, *County* and

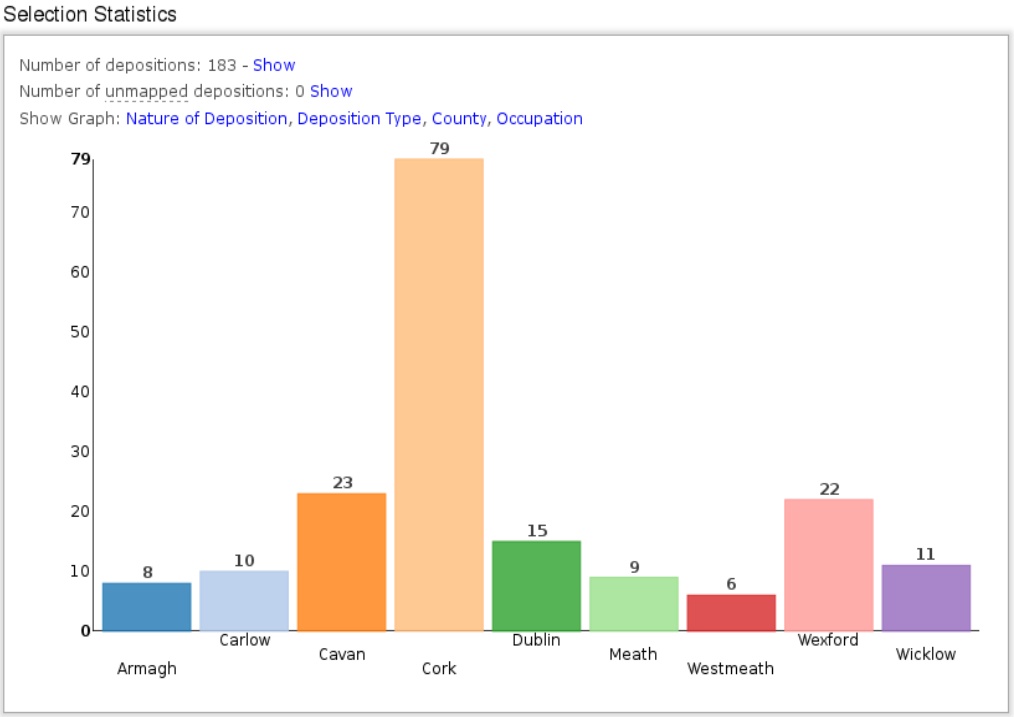

Figure 4.9: Selection statistics section of the visual exploration tool.

Occupation of Deponent. As with the addition of extra chart types, it would also be straightforward to add more data types that can be graphed. The user can switch between these by clicking in the links displayed immediately above the chart.

4.4.1 Criticism of Graph

The graph had a shortcoming when it came to plotting things where there were a high number of different values. This happened when a large number of depositions were on display and the 'Deponent Occupation' graph was chosen. Due to the high number of different occupations that would be contained withing the selection the widths of the bars in the bar chart were very thin, and the labels underneath them covered each other up making them unreadable. This was somewhat mitigated by the ability to hover over a bar, which would reveal what it represented and it's value. An alternative to the bar chart that would be able to display a larger number of labels is the Tree Map which was discussed in Section 2.1.3 on Page 8.

4.5 Deposition View

This view displays information about and the transcription of a single deposition. When a deposition is selected either from the *Map* or the *Data View* a request is sent to the server to retrieve all the information that exists about that particular deposition, including a transcription of the deposition. All of this information is formatted and displayed to the user. This view can been seen in Figure 4.10. Similar to the other views available in the visual exploration tool, the functionality of the *Deposition View* is encapsulated within it's own singleton object, in this case one called *Deposition View*. This ob-

ject is relatively simple and is only responsible for fetching and displaying a specific transcription when requested.

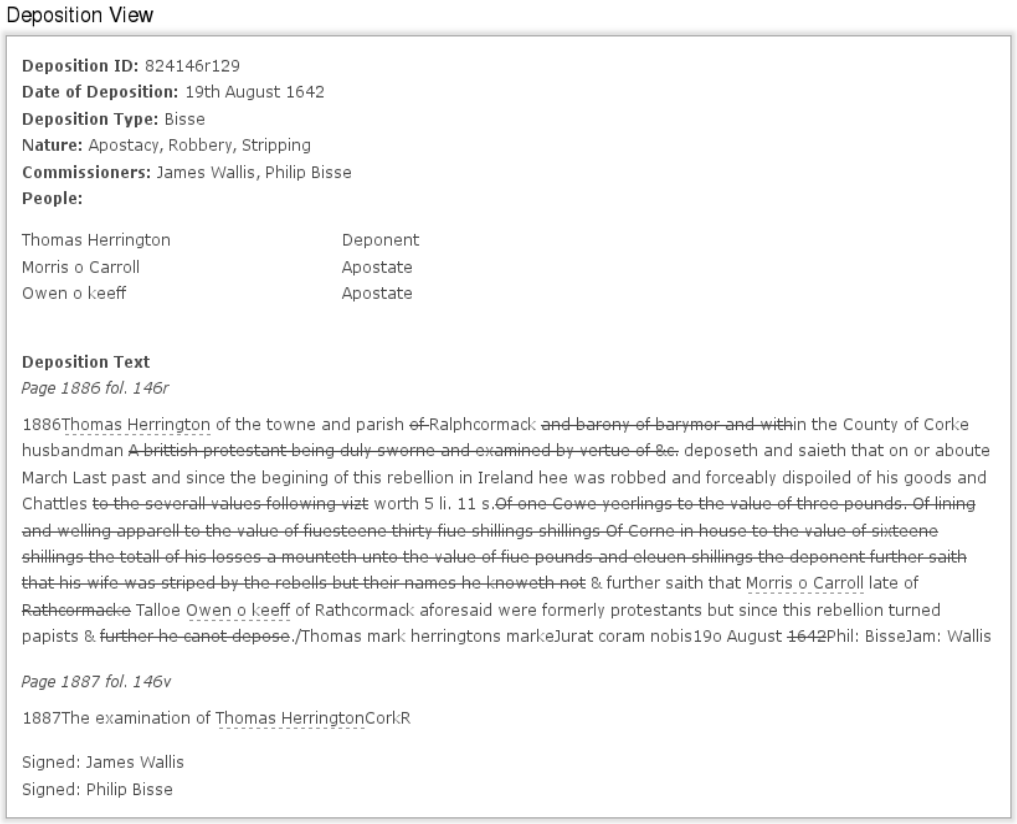

Figure 4.10: Deposition view section of the visual exploration tool.

4.5.1 Deposition Transcription

A limited amount of processing of the transcription takes place before it is displayed. There are a number of markup tags embedded within the transcription that contain meta data such as folio numbers and marginalia. This formatting of the text is accomplished by making use of the fact that the transcriptions have been marked up with XML tags, which allows the transcription to be formatted using CSS without any pre-processing requirement.

The transcription is searched for the occurrence of any person's name that is associated with the deposition. If an occurrence is found the user will be able to determine the relationship that person has with the deposition by hovering over the person's name in the transcription. The name will be underlined with a broken line indicating that more information is available.

4.6 Data View

Occasionally the need arises for the deposition to simply be displayed as a list. This view allows for unmappable depositions to be presented to the user. A screen shot of this view can be seen in Figure 4.11. This view is simply a list of depositions displayed in a table. When a row is clicked the corresponding deposition is loaded into the *Deposition View*.

This view also has the option to export the table shown to CSV format. This could be useful for viewing the data in a spreadsheet application. Also provided is the ability to generate a reading list of depositions. When clicked, this link will create a page that contains the transcription of every deposition on the list.

4.7 Logging

In order to obtain user interaction data from the evaluations a logging system was implemented. This was a fairly simple class that logged a number of specific interactions the user performed while carrying out the tasks that has been assigned to them. The *Logger* object would create a log entry that consisted of a string describing the type of interaction that had occurred along with the time of the interaction and an ID number that had been set

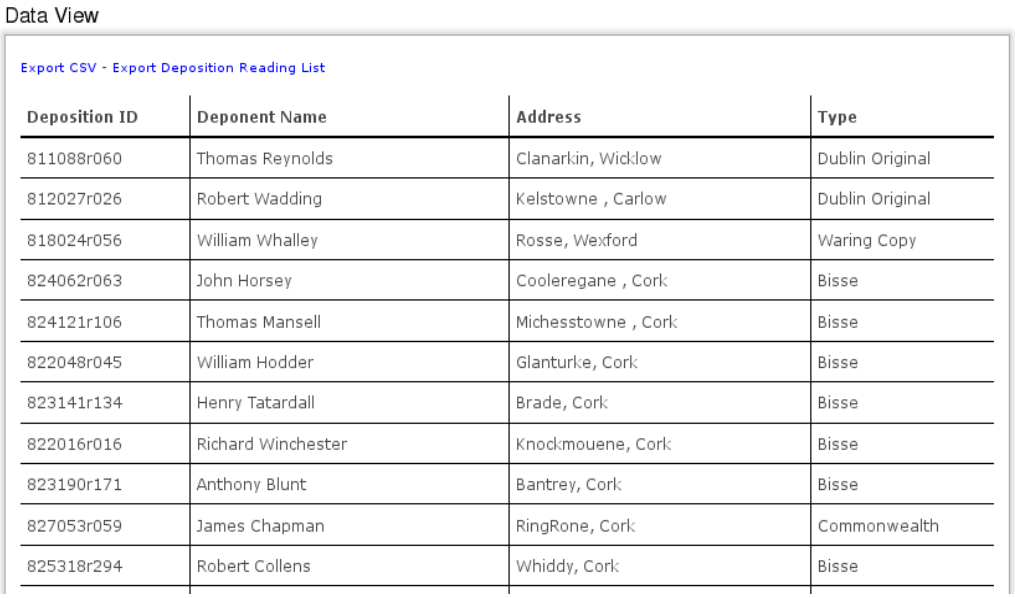

Figure 4.11: Data view section of the visual exploration tool.

at the beginning of the evaluation to differentiate the different evaluators. This log entry was then sent to the server where it was appended to a text file.

The floating sidebar was also used for logging. Each participant was asked to click the 'Task Complete' link in the sidebar after they had completed each task, allowing the individual tasks to be differentiated in the log files. This link can be seen near the bottom of Figure 4.2 on page 25. The sidebar also displayed the number of the current task.

The following user interactions were logged by the *Logger* module as part of the evaluation.

- Filter added or removed.
- Filters reset.
- Marker on map clicked.
- Item in map info box (Figure 4.8, Page 33) clicked.
- Jump to section via sidebar.

- Section hidden or shown.
- Statistics View graph type changed.
- Row clicked in Data View.
- Task Complete.

4.8 DSPL

A significant amount of time was invested in to implementing the aborted attempt at creating a generalized visual exploration tool based around DSPL. This section will provide a brief overview of that work.

The work for this was mainly done on the server side. DSPL defines a number of concepts, each of which were implemented as separate PHP classes. The PHP scripts were able to read the data from the DSPL file and then send that data as JSON to the client for display.

The DSPL specification itself was modified through the addition of a method for specifying additional types of data sources. In the original specification only CSV files are supported. This was expanded to allow SQL queries to be used as well. A DSPL file for the 1641 Depositions was in the process of being created when it was decided that it would not be a suitable method of enabling the visualisation of the Depositions, due to its focus on statistical data.

Chapter 5

Evaluation

This chapter outlines the user evaluation of the visual exploration tool. An overview of how data was collected and a summary of the results obtained are then discussed.

5.1 Methodology

Evaluations of the tool were performed to assess the usability, usefulness and appropriateness of the tool developed to allow for the visual exploration of semantic data. The data collected from these user studies could then be used to determine if an effective design had been arrived at, and what further improvements could be made in order to support the user in effectively visually navigating through the 1641 Depositions database. Furthermore the participants would be asked which sections of the tool they found most useful in accomplishing the tasks set out for them.

Both quantitative data via a questionnaire (which can be seen in the appendix on page 56) and qualitative data via a short interview with each of the participants at the end of the evaluation sessions were captured during the

evaluations. A number of tasks were set for the evaluator to perform. These tasks were designed to get the evaluator to use the tool to find specific information by using the various different views of the data that were available. These tasks can be found in the appendix on page 56 of this dissertation.

5.2 Ethics

Ethical approval was sought and received for the evaluations. All of the participants were aware that they could withdraw from the evaluation at any time without penalty. They were informed that all information gathered during the evaluations would be anonymised and could not be linked to the participant personally.

5.3 Participants

A number of people were approached and asked to participate in the evaluations. The Participants that were chosen to participate in this study were selected because of their familiarity with the 1641 Depositions and expertise in historical research. There were relatively few participants, with only four experts completing the study, however there were three other experts who were given access to the tool to use and they provided similar, albeit unrecorded and informal, feedback to the four who took part in the evaluation proper.

5.4 Evaluation Design

The evaluation was designed with the help of an expert from the Faculty of Arts, Humanities and Social Sciences in Trinity College, Dublin. This expert

has experience in research into the 1641 Depositions. The first section of the evaluation consisted of a number of tasks each participant was asked to perform using the tool.

Once the evaluator had completed these tasks they were then asked to complete a short questionnaire. This questionnaire can be found in the appendix. The questions were intended to ascertain how well the tool performed from a technical viewpoint and whether the tool would be useful to the participant in their investigations of the 1641 Depositions. This portion of the evaluation provided the quantitative data used to assess the success of the tool.

After the questionnaire was completed, the evaluator was given a chance to leave any comments they might have had about their experience using the visual exploration tool. Following this there was a short informal verbal interview. This interview was used to gain deeper insights into the users experience of the tool. This portion of evaluation process was helpful in gathering qualitative data and feedback that could be used to assess the success of the tool.

While the answers to the questions were not particularly important all of the evaluators gave similar answers. The exceptions to this being any of the tasks that required the evaluators to extract information from the depositions transcriptions themselves.

5.4.1 Tasks

The tasks set in the first part of the evaluation involved using the tool to search through the 1641 deposition database looking for answers to questions contained either within a specific deposition or about a group of depositions. These tasks were designed to get the participant to use as many of the features

of the tool as possible, with each task focusing on directing the user to use a specific view of the data. Some of the tasks built upon previous tasks, allowing the user to experience the process of further refining a filter set. Participants were given seven tasks in total.

Tasks 1 and 3 each took the shortest amount of time to complete for most participants. These tasks were fairly simple, asking the evaulator to apply a single filter and report the number of depositions returned.

Task 4 (See Appendix, Page 56) took the longest amount of time to complete for most participants, this task was a continuation of the previous task. The user was given enough information to apply filters such that only four depositions were shown on the *Map*. They could then find the specific deposition using the deponent's name. This task then involved reading through the text of the deposition to find the answer to the task. Despite the complexity of this task, all participants found the correct answer. In some cases however they did not make use of all the information given in the question, opting instead to apply a more general filter returning a large set of results, and then searching through the *Data View* to find the specific deposition mentioned.

For the most part all of the participants gave the same answers to the questions, and they all successfully completed all tasks successfully. Exceptions to this were the answers to tasks 5 and 6. In task 5 there were multiple possible correct answers that could be given, depending on what specific deposition the participant chose. Task 6 was a subjective question, asking the participant to comment on the distribution of a set of depositions, therefore each participant gave slightly different answers, although they all were broadly similar in nature.

5.4.2 Data Capture

During the initial section of the evaluation all of the evaluator's interactions with the tool were logged. The user was asked to click the 'Task Complete' link after they had completed each of the set tasks. This allowed the researcher to determine exactly what steps each evaluator took in order to complete each of the tasks. A complete list of the interactions that were captured and logged can be found in Section 4.7 on Page 38.

5.5 Evaluation Sessions

The evaluations took place in late September near the end of the project. Each participant was given access to a laptop computer on which they could use the tool. Each session took approximately 40 minutes, with each participant evaluating the visual exploration tool in separate sessions. Before the start of each evaluation, after they had been given instruction in using the tool, the evaluators were given a chance to familiarize themselves with the tool.

The laptop that was used in the evaluations was a Dell Inspiron 6400 from 2006 with 1.5GB RAM, a 1.83GHz Intel Core 2 Duo Processor and integrated Intel graphics running Kubuntu Linux. A separate mouse was provided as it was easier to use the application with it than the built in track pad.

5.6 Feedback

Overall the feedback from the evaluations was positive. A summary of the results of the questionnaire part of the evaluation are presented in the tables

below. Whilst there were 5 options available as responses to each statement, *1) Strongly Disagree, 2) Disagree, 3) Neutral, 4) Agree* and *5) Strongly Agree*, only the final two (*Agree* and *Strongly Agree*) were chosen by the participants. Therefore for the sake of brevity the other options are omitted as they all have a value of zero.

The questions in the Table 5.1 were designed to gain feedback about the overall experience the user had while using the tool.

Table 5.1: General statements.

Statement	**Agree**	**Strongly Agree**
I found the tool easy to use	2	2
The tool would be useful to me when working with the 1641 Depositions	1	3
The tool behaved as I indented when I performed an action	0	4
The information presented was easy to understand	1	3
The tool was responsive to my actions	1	3
The tool was intuitive to use	2	2
I enjoyed using this tool	2	2

The questions in the second Table 5.2 were designed to gather data on what particular sections of the tool the users found useful.

Table 5.2: Section specific statements.

Statement	**Agree**	**Strongly Agree**
I found the map section useful	2	2
I found the selection statistics section useful	1	3
I found the deposition view section useful	0	4
I found the data view section useful	1	3

It can be seen from these tables that the responses to these statements was slightly more biased towards the response of *Strongly Agree*. It is worth noting that not everyone strongly agreed with the statement that they found the map section useful. This is supported by observations made during the evaluations whereby some of the participants preferred to rely on only the *Filter*, *Statistics* and *Data View* sections of the tool when looking for specific depositions. The user interaction logs also corroborate this observation. It should also be noted however that other participants said that for them the map was the most useful visual element of the tool.

There were a number of evaluators that gave comments on how the tool could be improved in the future. One request was to provide a way to view the scans of the original source material from which the depositions were transcribed. A common request was the ability to save the state of the application for later use. This would allow a user to set up whatever filters and views they needed and then be able to come back at a later date or compare to another state by swapping between two saved states.

5.6.1 Negative Feedback

A common complaint of users was that the they would accidently change the zoom level of the map when they were scrolling up or down the page. This happens because on Google Maps one can use the scroll wheel to zoom in and out when hovering the mouse pointer over the map. When scrolling the cursor would move over the map and then cause the zoom level to change. The tool did not contain an easy way to reset the map to its original position and zoom level.

Chapter 6

Conclusion

In the previous chapter the method of evaluation was examined, and an overview of the results and some discussion of these results were given. This chapter will draw conclusions from these results and make recommendations for future works on how to best support users in visually exploring noisy historical data. Finally a number of criticisms of the project will be discussed followed by a brief look at future work that could be done.

6.1 Results

6.1.1 Guidelines

From the results of the evaluation a number of guidelines were deduced which answer the research question posed in Chapter 1.3 Introduction, that is: *What guidelines should be followed to support users in the visual exploration of noisy historical data?*. The guidelines that were arrived at are listed here.

Multiple Synchronized Views

Provide multiple views of the same data beside each other. The first guideline recommended by this dissertation is that one should provide multiple differing views of the same data alongside one another. Many of the participants in the evaluation commented on the fact that it was useful to them to see different aspects of the data simultaneously.

Incomplete Data

Users will except a small percentage of unvisualised data. This work showed that when plotting data users are happy to deal with a small percentage of unmapped points. While it is important to ensure that the user has access to a complete set of results for a given filter set, they do not all have to be plotted visually. A user will accept a certain level of missing or incomplete data without it adversely affecting the usability of the tool. This is evidenced by the interviews that were carried out as part of the evaluation. When asked if the missing data was an issue, all of the participants in the evaluation responded that it was not.

Simplicity

Simplicity is key. From the experience gained through conversing with humanities researchers, specifically in the area of history there seems to be a serious need for tools like the one developed over the course of this project. Many of the researchers do not make good use of technology in their work. This seems to stem from a lack of trust in technology to faithfully represent the source material. Many can also be put off by the complexity of existing applications that allow one to query databases. By keeping the application simple and providing an uncluttered user interface researchers can focus on

the task of investigating the source material and not spend time trying to figure out how to use the application. This guideline can be broken down into a number of smaller sub-guidelines as follows.

Integrate available meta data into textual data as simply as possible. Consider enhancing textual data with any available meta data. In this project the Deposition transcriptions were enhanced by highlighting any occurrences of people in the text. The user could then hover over the name of the person to find out more information.

A simple filter system will suffice. Powerful filter systems can be too complex or distracting. This is especially true when the user is not experienced with or even dislikes using computers. The use of simple boolean AND/OR relationship in filtering can be sufficient when combined with multiple views of the data.

6.1.2 Effectiveness of Tool

The overall feedback from the evaluators was positive. All of the evaluators expressed a desire to see the tool released online alongside the existing 1641 Depositions website. There were some who stated that they had recently conducted some research into the 1641 Depositions and that had they had use of the tool when they were doing the research, it would have saved them weeks of time they had spent manually searching through the depositions and collecting statistics.

One of the participants was particularly interested in researching depositions done by specific commissioners and had spend several weeks manually searching through a collection of text documents, extracting those that were done by a certain commissioner. This task can be achieved by the using visual exploration tool very quickly by simply setting the correct filter. It

would also have been possible to achieve the same by building an SQL query by hand and querying the database without the use of the tool. However the data would not be presented as simply, it would not be plotted on a map and the user would have had to structure the query in such as way as to include all of the alternative names of the commissioner. Normalized fields, such as commissioner, were handled transparently to the user, by the tool.

The tool also makes it easy to research an area which hasn't been looked into a lot, such as that of tracking commissioners date they went around the country taking depositions. This would be facilitated through the use of the commissioner filter in combination with the date range filter. This could be further enhanced by the introduction of a 'Play' button that would slowly increment the centre point of a set date range along the time line. Filtering by date is particularly difficult to achieve without use of the tool, through the use of a simple SQL query. The dates in the database are not in a standard format, being separated into separate columns for day, month and year. Querying this data would involve the creation of a complex SQL script, something which would be beyond the abilities of most[1] of the researchers that are interested in querying the 1641 depositions database. The tool however filters dates on the client side, converting the columns into a more easily queried form.

Throughout the evaluations users reported that the tool was intuitive to use, behaving as they expected in response to their interactions with it. All of the participants were able to complete all of the tasks given to them in a reasonable amount of time. Many of the tasks set were not trivial and required a high amount of searching of the data base, which would have taken a considerable amount of time, if the tool were not used. The total

[1]The same could also be said for a simple SQL query, as the users of this tool tend not to have knowledge of SQL

completion rate indicated that the tool was easy to use, even for a user that had only just been given instruction in it's operation.

6.2 Criticism

During the evaluations the participants would frequently clear all the filters to apply a different set of filters. When they did this there was a significant delay[2] as all of the depositions were downloaded and displayed on the map. During this time the tool would be unresponsive, only becoming responsive again when all of the depositions were displayed. A simple solution to this problem is to have the option to toggle whether or not depositions are loaded when all filters are cleared.

The map marker icons used on the map were not ideal, they were too large. Smaller icons would have made for a clearer visualisation, as the marker icons would not overlap as much when there was a high density of depositions on display.

There was no way to save a specific filter configuration in the tool. Nor was there a way to easily switch between two different sets of filters for comparative purposes. This limited the tool, as users were unable to easily compare two sets of filtered depositions.

Finally, the evaluation was quite limited, apart from the low number of participants. There should be more rigorous testing of different combinations of views and tool configuration in order to find an optimal solution.

[2]This delay was more noticeable during the evaluations due to the hardware that was used. The laptop used was a 2006 Dell Inspiron laptop with 1.5GB RAM and a 1.83GHz Intel Core 2 Duo Processor.

6.3 Future Work

This section will look at a range of improvements that could be made to the visual exploration tool itself and the evaluation process.

6.3.1 Deposition Transcriptions

When searching for the names of people within a transcription there are improvements that could be made. Firstly the search needs to be case insensitive. This is made more difficult by the fact that we are dealing with historical data. The conventions for capitalization were different at the time. A notable example is the case where in 17th century English the letters *ff* could be used as the upper case of the letter *f* rather than the modern *F*.

There is also more meta data available that could be integrated into the transcriptions. Place names and dates that are mentioned could be highlighted in the text. These could also be made into links that when clicked return a set of filtered data corresponding to the link that had been clicked. For example, clicking on a date could display all depositions that were taken within a certain span of time centred around that date.

The transcription text is currently broken up into paragraphs where each is originally from a different folio. The page number is also given. This page number, as pointed out by one of the evaluators, is not meaningful to researchers and should instead be replaced by the paragraph's manuscript number.

6.3.2 Manipulation of Database

The tool should be enhanced to allow for the manipulation of appropriate parts of the data in the database. As mentioned in Section 4.3.1 not all

depositions were placed in the correct location. The tool should allow for this to be corrected by an authorized user. The most intuitive way, in this authors opinion, would be to allow a user to simply drag and drop the existing marker to the new correct location, but only when the user is in an 'Edit' mode so that markers are not moved in error.

6.3.3 Further Evaluation

The study that was performed was rather limited in the number of participants. A larger study with a greater number of participants would help to reinforce or possibly refute the conclusions and recommendations that were drawn from the evaluation conducted as part of this dissertation.

6.3.4 Other Improvements

There are a number of improvements that could be made to the visual exploration tool. Many of which which were suggested by the evaluators during the evaluation sessions. A list of some of these improvements follows.

- Ability to save current filter state.
- Option to not load depositions when there are no filters applied.
- Ability to add an appropriate filter when the graph in the statistics view is clicked.
- Improvements to the map: heat maps, polygon overlays showing county, barony or townland boundaries.
- Addition of different graph types.
- Ability to sort the *Data View* table by column.
- Automatically scroll page to deposition view when deposition is loaded.
- Link to images of scans of original depositions.

- Ability to manually save individual depositions to a list. This list can then be viewed like any other group of depositions returned from the server.
- Ability to search for any arbitrary string contained within the depositions, allowing researcher to map depositions using data that hasn't been extracted as meta data.
- Integration with historical maps, making use of normalized place names.

Appendix A

Evaluation Questionnaire

Tasks

Using the visual data exploration tool please answer the following questions, you may write the answers to each below the question in the space provided. If you require extra space for answers please ask the investigator for more paper.

1. How many brewers are deposed, and in what counties?

2. What type of crime was most commonly reported in County Dublin and County Louth?

3. How many County Cork depositions report 'Words'?

4. Mathew Boulster of County Cork reported 'Words' when he was deposed on 16th March 1643. What were the 'Words' reported?

5. What are the total losses of the Keeper of the Park?

6. Compare the geographic distribution of depositions before and after the year 1650.

7. Compare the number of depositions before January 1644 that mention rape with the number after that date.

Evaluation

State how strongly you agree or disagree with the following statements.

I found the tool easy to use.				
☐ Strongly Disagree	☐ Disagree	☐ Neutral	☐ Agree	☐ Strongly Agree
The tool would be useful to me when working with the 1641 Depositions.				
☐ Strongly Disagree	☐ Disagree	☐ Neutral	☐ Agree	☐ Strongly Agree
The tool behaved as I intended when I performed an action.				
☐ Strongly Disagree	☐ Disagree	☐ Neutral	☐ Agree	☐ Strongly Agree
The information presented was easy to understand.				
☐ Strongly Disagree	☐ Disagree	☐ Neutral	☐ Agree	☐ Strongly Agree
The tool was responsive to my actions.				
☐ Strongly Disagree	☐ Disagree	☐ Neutral	☐ Agree	☐ Strongly Agree
The tool was intutive to use.				
☐ Strongly Disagree	☐ Disagree	☐ Neutral	☐ Agree	☐ Strongly Agree
I enjoyed using this tool.				
☐ Strongly Disagree	☐ Disagree	☐ Neutral	☐ Agree	☐ Strongly Agree

I found the map section useful.				
☐ Strongly Disagree	☐ Disagree	☐ Neutral	☐ Agree	☐ Strongly Agree
I found the selection statistics section useful.				
☐ Strongly Disagree	☐ Disagree	☐ Neutral	☐ Agree	☐ Strongly Agree
I found the deposition view section useful.				
☐ Strongly Disagree	☐ Disagree	☐ Neutral	☐ Agree	☐ Strongly Agree
I found the data view section useful.				
☐ Strongly Disagree	☐ Disagree	☐ Neutral	☐ Agree	☐ Strongly Agree

Comments

If you have any other comments you could like to make about the tool please provide them below.

Bibliography

[1] Arc Diagram of Bible Cross References Image. http://www.chrisharrison.net/index.php/Visualizations/BibleViz, September 2012.

[2] Cartogram Image. http://www.worldmapper.org/svg/map2/index.html, September 2012.

[3] Chrome Frame - Internet Explorer Plug-in. https://developers.google.com/chrome/chrome-frame/, September 2012.

[4] Circle Graph Image. http://mbostock.github.com/d3/talk/20111116/bundle.html, September 2012.

[5] Collatz Tree Diagram Image. http://www.jasondavies.com/collatz-graph/, September 2012.

[6] D3.js Library. http://d3js.org/, September 2012.

[7] Drupal Content Management System. http://drupal.org/, September 2012.

[8] Google, Dataset Publishing Language Overview. https://developers.google.com/public-data/overview, September 2012.

[9] Google Maps Geocoding API Documentation. https://developers.google.com/maps/documentation/geocoding/, September 2012.

[10] Ireland Export Treemap Image. http://atlas.media.mit.edu/explore/tree_map/export/irl/all/show/2009/, September 2012.

[11] Parallel Coordinates Image. http://upload.wikimedia.org/wikipedia/en/4/4a/ParCorFisherIris.png, September 2012.

[12] Protovis Library. http://mbostock.github.com/protovis/, May 2012.

[13] Scalable Vector Graphics (SVG) Version 1.1 Specification. http://www.w3.org/TR/SVG/, September 2012.

[14] Wikimedia Web Browser Usage Statistics. http://stats.wikimedia.org/archive/squid_reports/2012-08/SquidReportClients.htm, September 2012.

[15] Yonatan Aumann, Ronen Feldman, Yaron B. Yehuda, David Landau, Orly Liphstat, and Yonatan Schler. *Circle Graphs: New Visualization Tools for Text-Mining*, volume 1704, chapter 30, pages 277–282. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[16] Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky. A tour through the visualization zoo. *Commun. ACM*, 53(6):59–67, June 2010.

[17] Jeffrey Heer and Ben Shneiderman. Interactive Dynamics for Visual Analysis. *Queue*, 10(2), February 2012.

[18] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the First IEEE Confer-*

ence on Visualization: Visualization '90, pages 361–378. IEEE Comput. Soc. Press, 1990.

[19] D. A. Keim, S. C. North, C. Panse, and J. Schneidewind. Efficient cartogram generation: a comparison. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*, volume 0, pages 33–36, Los Alamitos, CA, USA, 2002. IEEE Comput. Soc.

[20] D. A. Keim, C. Panse, and S. C. North. Medial-Axis-Based Cartograms. *IEEE Computer Graphics and Applications*, 25(3):60–68, May 2005.

[21] Catherine Plaisant. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*, AVI '04, pages 109–116, New York, NY, USA, 2004. ACM.

[22] Ben Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.*, 11(1):92–99, January 1992.

[23] Ben Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336+, Washington, DC, USA, 1996. IEEE Computer Society.

[24] Martin Wattenberg. Arc Diagrams: Visualizing Structure in Strings. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*, INFOVIS '02, Washington, DC, USA, 2002. IEEE Computer Society.

[25] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Information Visualization,*

1995. Proceedings., volume 0, pages 51–58, Los Alamitos, CA, USA, October 1995. IEEE.

[26] Xiaofang Zhou, David Truffet, and Jiawei Han. Efficient Polygon Amalgamation Methods for Spatial OLAP and Spatial Data Mining. In *Proceedings of the 6th International Symposium on Advances in Spatial Databases*, SSD '99, pages 167–187, London, UK, UK, 1999. Springer-Verlag.